2012
VIII Convegno Nazionale
dell'Associazione Italiana Scienze della Voce

LA VOCE NELLE APPLICAZIONI

a cura di Mauro Falcone e Andrea Paoloni
Fondazione Ugo Bordoni

Versione estesa

Bulzoni editore

2012
Atti dell'VIII Convegno Nazionale
dell'Associazione Italiana Scienze della Voce

LA VOCE NELLE APPLICAZIONI

a cura di Mauro Falcone e Andrea Paoloni
Fondazione Ugo Bordoni

Bulzoni Editore

LA VOCE NELLE APPLICAZIONI

Atti dell'VIII Convegno AISV 2012
Dipartimento di Linguistica Università Roma Tre - Roma
25-27 gennaio 2012

Grafica, impaginazione e realizzazione del CD-Rom a cura di
Consuelo Tuveri e Stefania Vinci

ISBN 978-88-7870-774-0

00185 Roma, via dei Liburni, 14
http://www.bulzoni.it
e-mail: bulzoni@bulzoni.it

INDICE

La lettura ad invito del Prof. Hirst "Annotating Intonation" e i contenuti delle tavole rotonde sulla Qualità della voce e sulla Verbalizzazione automatica sono disponibili nei formati forniti dagli autori nelle rispettive cartelle nel CD allegato

ORGANIZZAZIONE DEL CONVEGNO

COMITATO SCIENTIFICO

Cinzia Avesani	CNR-ISTC, Padova	Italia
Carla Bazzanella	Università di Torino	Italia
Sergio Canazza	Università di Padova	Italia
Carlo Cecchetto	Università di Milano “Bicocca”	Italia
Ugo Cesari	Università di Napoli	Italia
Claudia Ciancaglini	Università di Roma “La Sapienza”	Italia
Gianpaolo Coro	CNR-ISTI, Pisa	Italia
Piero Cosi	CNR-ISTC, Padova	Italia
Giovanni Costantini	Università di Roma “Tor Vergata”	Italia
Francesco Cutugno	Università di Napoli “Federico II”	Italia
Amedeo De Dominicis	Università della Tuscia, Viterbo	Italia
Renato De Mori	Università di Avignone	Francia
Mauro Falcone	Fondazione Ugo Bordoni	Italia
Antonella Giannini	Università di Napoli “L’Orientale”	Italia
Barbara Gili-Fivela	Università del Salento, Lecce	Italia
Mirko Grimaldi	Università del Salento, Lecce	Italia
Daniel Hirst	Università della Provenza	Francia
Michele Loporcaro	Università di Zurigo	Svizzera
Giovanna Marotta	Università di Pisa	Italia
Marco Matassoni	Fondazione Bruno Kessler	Italia
Pietro Maturi	Università di Napoli, “Federico II”	Italia
Franca Orletti	Università di Roma, “Roma 3”	Italia
Andrea Paoloni	Fondazione Ugo Bordoni	Italia
Roberto Pieraccini	ICSI, Berkeley	USA
Antonio Romano	Università di Torino	Italia
Renata Savy	Università di Salerno	Italia
Oskar Schindler	Università di Torino	Italia
Stephan Schmid	Università di Zurigo	Svizzera
Rosanna Sornicola	Università di Napoli “Federico II”	Italia
Mario Vayra	Università di Bologna	Italia
Claudio Zmarich	CNR-ISTC, Padova	Italia

ORGANIZZATORI DEL CONVEGNO

Associazione Italiana di Scienze della Voce, Fondazione Ugo Bordoni, Università degli Studi di Napoli Federico II, Università degli Studi “Roma Tre”.

SEGRETERIA DEL CONVEGNO

Annalisa Filardo, Rosita Lio, Pamela Mattana, Gianluca Mignini, Fabio Poroli, Massimiliano Todisco.

COMITATO ORGANIZZATORE

Piero Cosi	CNR-ISTC, Padova	Italia
Francesco Cutugno	Università degli Studi di Napoli “Federico II”	Italia
Amedeo De Dominicis	Università degli Studi di Viterbo	Italia
Mauro Falcone	Fondazione Ugo Bordoni	Italia
Franca Orletti	Università degli Studi “Roma Tre”	Italia
Pamela Mattana	Università degli Studi di Roma “La Sapienza”	Italia
Gianluca Mignini	Università degli Studi di Napoli “Federico II”	Italia
Andrea Paoloni	Fondazione Ugo Bordoni	Italia
Antonio Romano	Università degli Studi di Torino	Italia

PREMESSA

L' VIII edizione del convegno AISV si è svolta a Roma presso il Dipartimento di Linguistica dell'Università Roma Tre nei giorni 25-27 gennaio 2012. Tema del convegno gli aspetti applicativi degli studi sulla voce ed in particolare quindi la didattica e l'insegnamento delle lingue, la logopedia e la qualità della voce, la trascrizione e la verbalizzazione automatica del parlante e poi "le macchine parlanti", ovvero le interfacce vocali oggi sempre più diffuse che riproducono la voce e che riconoscono il parlato e il parlatore.

Per la prima volta congiuntamente al convegno AISV si è svolta la terza edizione del convegno EVALITA, una iniziativa avente lo scopo di valutare i sistemi di linguaggio naturale e le tecnologie vocali per l'italiano.

Il convegno AISV è stato aperto da una lettura ad invito di Federico Albano Leoni dal titolo: "Il fonema: realtà o illusione?" lettura che non ha certamente lasciato indifferenti coloro che non avessero ancora conosciuto l'ardita l'ipotesi di Albano Leoni secondo la quale il fonema non è una realtà fisica, ma solo un modello interpretativo mutuato dalla scrittura.

Nelle prime tre sessioni si sono affrontati i temi della didattica, della logopedia inclusi gli aspetti alla cura e delle diagnostiche per la salute della voce.

Nella sessione poster i contributi erano relativi alle interfacce vocali evolute, alla valutazione dell'intelligibilità, e alla realizzazione e conservazione di archivi di materiale vocale.

Al centro del convegno una seconda lettura ad invito: Daniel Hirst ha svolto il tema "The state of the art in the field of the various intonation annotation systems" suscitando anche in questo caso una vivace discussione con i partecipanti.

A seguire una ulteriore sessione che ha visto interessanti contributi nell'ambito della ricerca sugli aspetti fonetici del parlato.

Hanno costituito parte integrante del convegno due tavole rotonde.

La prima, coordinata da Antonio Romano, è stata dedicata alla "Qualità della Voce" con particolare riferimento agli aspetti della logopedia e della medicina, anche con l'obiettivo di aumentare la collaborazione tra i vari settori disciplinari interessati a quest'oggetto di studio.

La seconda, coordinata da Piero Laface, è stata dedicata ad un aspetto particolarmente significativo dal punto di vista applicativo: la verbalizzazione automatica. Si è discusso sulla concreta possibilità di trascrivere automaticamente un'udienza processuale, con considerevoli risparmi di denaro ma soprattutto di tempo; al dibattito hanno partecipato, insieme ad esperti di resocontazione, anche rappresentanti degli organi della pubblica amministrazione.

EVALITA 2011

Il convegno è stato preceduto dal workshop di EVALITA (Evaluation of NLP and Speech Tools for Italian), un'iniziativa dedicata alla valutazione degli strumenti di Natural Language Processing e degli strumenti di elaborazione per l'italiano. L'obiettivo generale di Evalita è di promuovere lo sviluppo delle tecnologie del linguaggio e della parola per la lingua italiana, fornendo un quadro di riferimento condiviso in cui diversi sistemi ed approcci possono essere valutati in modo coerente.

La diffusione dei compiti comuni e pratiche di valutazione condivisi è un passo fondamentale verso lo sviluppo di risorse e strumenti per lo NLP e le scienze del linguaggio. La buona risposta ottenuta nei precedenti workshop EVALITA, sia nel numero di partecipanti sia nella qualità dei risultati, hanno dimostrato che vale la pena di perseguire tali obiettivi per la lingua italiana.

PREMIO FRANCO FERRERO

Come ogni anno è stato assegnato il Premio Franco Ferrero al dottorando che ha contribuito con il lavoro di maggior interesse. È stato premiato l'Ing. Danilo Comminiello che ha presentato un lavoro dal titolo "Un'interfaccia acustica intelligente per comunicazioni immersive in ambienti non stazionari".

RINGRAZIAMENTI

Un ringraziamento particolare va alla professoressa Franca Orletti che, mettendo a disposizione la sua aula, ha contribuito a risolvere un problema organizzativo. Ringraziamo inoltre tutti i membri del comitato scientifico e tutti coloro che ci hanno aiutato con il lavoro di revisione degli atti a realizzare questo volume.

Ringraziamo di cuore tutti i membri della segreteria che ci hanno aiutato nella organizzazione convegno.

Mauro Falcone

Andrea Paoloni

IL FONEMA: REALTÀ O ILLUSIONE?

Federico Albano Leoni
Università di Roma La Sapienza
federico.albanoleoni@uniroma1.it

Abstract. The present paper, starting from the critique to the segmental model of language and the concept of phoneme, analyzes the phonological ideas that K. B. expresses in his *Sprachtheorie* (1934) and other minor writings. The investigation shows how Bühler, while appreciating the idea of phoneme as proposed by the Prague school and Trubeckoj as an example of the general principle of *abstraktive Relevanz,* integrates the idea with a representation model of the signifier based on the gestalt principle of the *Klanggesicht*. Therefore, the semiotic and linguistic principle of pertinence and of distinctive capability does not display itself always and only in the traditional phonologic oppositions (typically represented by minimal pairs), but can also be distributed along the entire string, analogously to what happens in the detection of faces, where the erasure of one or more features considered distinctive does not prevent the correct recognition. Ultimately, the principle of gestalt recognition, seemingly anti-economical from the viewpoint of human memory ability, finds its support in the integration of the symbolic and deictic fields. Thus the recognition of phonic faces, even when phonetically underspecified, is granted from context, namely the world as lived and thought, shared by the participants in a linguistic exchange. Bühler hence proposed a powerful and versatile phonological model, capable of taking into account the variability, instability and fragility of the vocal signal, opposed to that of the double articulation proposed by Martinet, shared today by all phonologies, and becoming a cliché.

1. PREMESSA.

Da qualche anno ho avviato una riflessione critica (presentata sistematicamente in Albano Leoni 2009) sulla prospettiva segmentale in linguistica, cioè la prospettiva secondo la quale le lingue sarebbero sistemi di elementi discreti (fonemi, morfemi ecc.) che, nella realizzazione, si disporrebbero in sequenze lineari, variando secondo regole predicibili.

Presento in maniera schematica sette argomenti che mi hanno portato a ritenere che il fonema sia un'entità certamente utile ai linguisti, ma che non riflette l'organizzazione funzionale delle lingue e non ne è un primitivo:

a) unità sprovviste di senso, quali sarebbero i fonemi, non possono essere vere unità linguistiche, né essere oggetto di attività cognitiva, ma possono essere al massimo l'oggetto della psicoacustica o della fisiologia della fonazione ; è un principio saussuriano al quale non so rinunciare (Albano Leoni, 2007);

b) la fonetica ha mostrato, almeno a partire dai lavori di Sweet (1878) e di Rousselot (1902), e di molti altri che la materia fonica è continua : è dunque da dimostrare che la rappresentazione discreta e segmentale imposta dalle fonologie sia la sola possibile;

c) il segnale fonico è fragile e drammaticamente variabile, al punto che nessuna concezione di spazio strutturale e funzionale e nessuna regola trasformazionale è in grado di dominare questa variabilità;

d) Miller & Johnson Laird (1976: 38) ci ricordano che : «Not having an explicit perceptual theory means that we cannot be sure what the primitives of system are»; ma nella co-

struzione delle fonologie non si tiene alcun conto della percezione, nonostante che in importanti filoni di studi sui modelli del riconoscimento del parlato si sappia molto bene che la percezione del parlato non muove dal basso in alto, ma dall'alto in basso e che in condizioni normali si ferma al livello di una unità significativa, comunque la si voglia chiamare, percepita come un tutto;

e) la pertinenza distintiva che si pretende di trovare nelle coppie minime (tipo /pere/ - /bere/) è una semplificazione cognitiva perché i percorsi ermeneutici che portano al riconoscimento di una parola sono molto più complessi e la percezione di un dettaglio fisico, come la presenza o l'assenza di sonorità, non è necessariamente determinante (come mostrano le innumerevoli sonorizzazioni e desonorizzazioni del parlato); inoltre paradossalmente, almeno nel caso di opposizioni di occlusive come in /para/ - /kara/, la logica delle coppie minime non tiene conto delle transizioni vocaliche che garantiscono il riconoscimento anche quando la consonante è soppressa o commutata; la commutazione dovrebbe quindi eventualmente essere tra sillabe;

f) porzioni di materia fonica significativa, perfettamente comprensibili nel contesto in cui sono inserite, divengono spesso incomprensibili e irriconoscibili quando vengono sottratte dal loro contesto; è questo il risultato di numerosi esperimenti, confermato di recente da un esperimento condotto sul francese;

g) la persistenza del modello fonologico segmentale conduce ad una asimmetria irriducibile tra le rappresentazioni delle due facce del segno, il significante e il significato : da un lato si ammette, ormai quasi universalmente, che i processi della significazione e le entità del significato e i sensi sono incalcolabili, vaghi, indeterminati, deformabili ; dall'altro lato, i processi e le entità del piano dell'espressione sono considerati discreti, calcolabili e combinabili secondo regole. Ma nessun punto di vista, né quello della psicologia cognitiva, né quello della semiotica, né quello delle teorie della percezione e della conoscenza è in grado di giustificare questa asimmetria. [1]

Aggiungerei che, dopo la pubblicazione di *Sound Pattern of English* (Chomsky & Halle : 1968), sono stati proposti una ventina di modelli fonologici (Durand & Laks, 2002 : 27): mi sembra un indizio chiaro del malessere teorico perché, per quanto eleganti e a volte geniali siano questi modelli, assomigliano alle scale delle stampe di Escher**,** che ci affascinano ma non portano da nessuna parte.

Sono dunque del parere che il fonema sia non un'entità primitiva delle lingue ma piuttosto uno strumento metalinguistico i cui fondamenti sono in una plurimillenaria tradizione occidentale di scrittura alfabetica, cioè nell'invenzione di uno strumento potente per la registrazione delle lingue. Non sarà infatti un caso se la fonologia segmentale si è sviluppata solo in seno a culture linguistiche e filosofiche di tradizione alfabetica. Naturalmente si può concepire la rappresentazione segmentale della lingua, cioè l'alfabeto, come un utile intermediario fra noi e la materia fonica (Mulligan, 2004, p. 8), o più in generale, come uno

[1] L'idea di Hjelmslev (ripresa da Prieto, Coseriu e Greimas) di un isomorfismo del segno dipende dall'ipotesi secondo la quale il piano del contenuto sarebbe analizzabile in costituenti, secondo il modello dei tratti dell'analisi fonologica. Questa idea è stata criticata da diversi punti di vista (p. es. da Martinet, 1970, 27-30). De Palo (2003) l'ha criticata discutendo del fallimento delle semantiche strutturali e componenziali. Come si vedrà più avanti, io sostengo la tesi, già formulata in Albano Leoni (2009), secondo la quale la simmetria tra le due facce del segno sarebbe piuttosto da vedere nella natura indeterminata di ambedue.

schema mediatore tra la sensazione e la ragione, o tra la sensazione e la categoria (Rousseau, 2004, 12-13), purché si ricordi che questo schema è determinato storicamente, è uno fra altri possibili, bel lontano dall'essere trascendentale. Anche la geometria euclidea è uno schema intermediario tra la percezione e la categoria, ma non è l'unica possibile e i suoi assiomi si sono rivelati non universali.

L'idea che la scrittura alfabetica sia per noi la chiave per la rappresentazione lineare e ordinata delle lingue è profondamente radicata nel pensiero occidentale. Cito due autori molto diversi ma che concordano su questo punto.

Leopardi (in Gensini 1998, 48-54) vedeva nell'alfabeto il principio fondante delle unità foniche altrimenti inconoscibili o addirittura inconcepibili:

> L'alfabeto è la lingua col cui mezzo noi concepiamo e determiniamo presso noi medesimi l'idea di ciascuno di detti suoni. Quegli che non conosce l'alfabeto, parla, ma non ha veruna idea degli elementi che compongono le voci da lui profferite. Egli ha ben l'idea della favella, ma non ha per niun conto le idee degli elementi che la compongono […] Ma per determinare gli elementi della voce umana articolata, l'unica lingua, come ho detto, è l'alfabeto.

Sullo smarrimento conseguente all'abbandono della rappresentazione alfabetica, ancora Saussure (1922:44[55]) si esprime con parole che ricordano Leopardi:

> Quando mentalmente si sopprime la scrittura, chi è privato di questa immagine sensibile rischia di non percepire più niente altro che una massa informe di cui non sa che fare. È come se si levasse il salvagente a chi sta imparando a nuotare.

Come contrappunto a quanto sto dicendo, nasce una domanda: dato che il linguaggio e le lingue sono fenomeni naturali, e che la produzione e la percezione della materia fonica linguistica avvengono per mezzo degli stessi strumenti psicofisici dei quali ci serviamo per respirare e masticare, o per vedere, annusare, toccare, e per elaborare le nostre sensazioni, non si potrebbe allora pensare a un modo di produrre e di percepire e rappresentare il parlato diverso da quello implicito nel modello segmentale? Un modo per il quale la elaborazione del percetto uditivo linguistico non sia così drasticamente diversa dalle altre modalità umane di percezione sensoriale? Un modo per cui si possa dire anche della percezione linguistica che**:**

> […] la percezione è un flusso e la descrizione che William James […] dava del flusso della coscienza vale anche per il flusso della percezione. I percetti discreti, come le idee discrete, sono "cose di fantasia come il Fante di Picche" (Gibson, 1986:364)

Se si mette in discussione un modello lineare, discreto, categoriale, la risposta a queste domande può venire da un modello olistico, gestaltico, fisiognomico.

2. COSA CI INSEGNA LA CONOSCENZA FISIOGNOMICA?

Il mio punto di partenza è la constatazione banale e intuitiva che ciascuno di noi riconosce volti (e paesaggi, strade, abitazioni ecc.) ed è in grado di distinguerli l'uno dall'altro. Se ci chiediamo come ciò avvenga, ci rispondiamo che di un volto riconosciamo il colore dei capelli o degli occhi, la forma del naso o delle labbra, la curva del mento e delle guance e così via, e più o meno gli stessi indizi ci consentono di distinguere un volto dall'altro. Apparentemente abbiamo applicato un principio generale di molte forme di conoscenza umana, che è quello della pertinenza distintiva, cardine, tra l'altro, delle fonologie e, più in generale, di molti livelli di analisi linguistica.

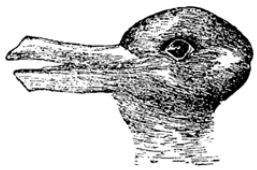
Ma se ci chiediamo cosa avvenga quando il colore dei capelli cambi, quello degli occhi sia velato dagli occhiali da sole, la linea delle labbra sia modificata chirurgicamente, o quando il volto ingrassi o dimagrisca, o invecchi, o impallidisca o si abbronzi, o quando, più semplicemente, pianga o rida e dunque si alteri, o anche quando due o più di questi cambiamenti succedano insieme, dobbiamo dirci che non accade un gran che, perché il volto rimane perfettamente riconoscibile e distinguibile da tutti gli altri. Dunque, riconosciamo un volto anche se si modificano quei tratti che abbiamo considerato salienti e che magari lo sono oggettivamente.

Già questa semplice riflessione su una banale esperienza quotidiana pone un problema teorico che riguarda la definizione e la localizzazione della pertinenza distintiva. Perché è ovvio che, se identifichiamo un oggetto o lo distinguiamo da un altro, ciò avviene perché in qualche modo abbiamo identificato costanti e differenze, ma la percezione, nel caso dei volti, è evidentemente olistica, gestaltica, e la costanza o la differenza sono nell'insieme e non nelle sue parti, e dunque la pertinenza è diffusa e non puntuale (come sarebbe ad esempio nella identificazione di una sequenza numerica, o nella distinzione tra due sequenze numeriche diverse).

La figura che segue, ideata dallo psicologo Joseph Jastrow a fine Ottocento, poi largamente ripresa e divenuta celebre, illustra bene due proprietà fondamentali della percezione (visiva). La figura, detta bistabile, può infatti essere vista o come una lepre che guarda verso destra, o come un'anatra che guarda verso sinistra e mostra: a) che il mero percetto e le parti fisiche che lo compongono (tratteggio, ombreggiature, linee ecc.) non sono niente al di fuori dell'insieme di cui sono parte; b) il percetto è fisicamente costante, ma esso può essere elaborato in base a schemi mentali diversi e dare quindi luogo a due immagini mentali diverse.

La percezione è dunque olistica e mentale.

Lepre e anatra



3. FISIOGNOMICA E FONOLOGIA.

E' possibile applicare questo modello ai suoni linguistici? Penso di sì.

In effetti, esistono rappresentazioni della voce che possiamo considerare fisiognomiche, nel senso che le proprietà che vi percepiamo, o immaginiamo di percepirvi, non sono localizzate in un punto della materia fonica, ma sono diffuse, appunto come nei volti: così, se ci domandiamo in cosa consista una voce *sensuale*, o *pastosa*, o *torbida*, o *sbigottita e*

deboletta, o *domenicale*, o *cattedratica* (Albano Leoni, 2002)[2], o in cosa consista un *tono imperioso* o di *sarcasmo quasi amaro* (Albano Leoni, 2003), dobbiamo riconoscere che queste impressioni risultano da una configurazione di elementi simultanei, ciascuno variante lungo una scala continua, nessuno dei quali è di per sé determinante per il riconoscimento. Dunque, anche in questo caso, il giudizio ingenuo ma non infondato, che ci porta a distinguere tra voci e inflessioni diverse, si basa su una pertinenza diffusa e non puntuale, come si è detto dei volti.

Penso che questa prospettiva "ingenua" potrebbe essere fruttuosa anche sul piano delle analisi fonologiche.[3]

Come si sa, l'assunto fondamentale delle fonologie, quasi un assioma, è la natura discreta e segmentale delle lingue. Il flusso sonoro, materialmente continuo, viene segmentato in unità minime, i fonemi,[4] dei quali si studiano le proprietà, i tratti, e si osserva quali siano distintivi.

E' opinione generalmente condivisa che in ciò si manifesti un principio regolatore delle lingue, un principio di economia, perché attraverso un numero limitato di tratti, dodici nella versione classica di Jakobson e Halle (1956) e delle loro combinazioni, è possibile costruire una matrice per ciascuno dei fonemi di tutte le lingue note, e ogni lingua, attraverso la combinazione dei fonemi di cui dispone (in genere poche decine) genera un numero illimitato di parole. Su questo stesso principio si basa la teoria martinettiana della doppia articolazione, sulla quale tornerò più avanti.

Ma, come ho detto all'inizio, l'assioma della natura segmentale e discreta dei significanti fonici, o quanto meno della sua applicazione tirannica è scarsamente fondato.

Un modello diverso di rappresentazione del significante linguistico era stato proposto da Karl Bühler, psicologo tedesco con forti interessi filosofici e linguistici, attivo a Vienna negli anni Trenta del Novecento.[5]

Proverò a sintetizzare la sua complessa posizione a proposito della fisiognomica delle parole (Bühler, 1934[1983]).

---

[2] Diverso il caso di *acuto* o *grave*, che sono metafore originariamente sinestetiche, ormai lessicalizzate, e riferite a proprietà misurabili di un suono.

[3] Menziono qui di sfuggita il caso delle analisi prosodiche (più dettagli sono in Albano Leoni 2009, 41-75). Queste oggi si ripartiscono tra un indirizzo che in qualche modo può essere considerato gestaltico, (Cruttenden 1986; Bolinger 1986, 1989; Halliday 1967, 1970; Rossi 1999), perché la prosodia vi è osservata nel suo contorno complessivo, e un indirizzo di tipo binarista (Pierrehumbert, 1980; Ladd, 1996), nel quale la prosodia è vista come il risultato di una successione discreta di toni alti e bassi.

[4] Ma non diversa era la *littera* della tradizione antica, definita come *pars minima vocis articulatae*. Sul rapporto tra *littera* e *fonema* v. Albano Leoni (2009, 83-85, 141-148).

[5] Su Bühler e il suo pensiero linguistico, sui suoi legami filosofici con Kant, Cassirer e Husserl, sulle sue complesse relazioni con le diverse scuole di psicologia della *Gestalt*, si vedano i saggi, la bibliografia e i documenti raccolti in Friedrich & Samain (2004), la monografia di Persyn-Vialard (2005), i testi che accompagnano la traduzione francese della *Sprachtheorie* (Bühler 1934 [2009]) e infine i saggi contenuti in «Verbum. Revue de linguistique», XXXI, 1-2, 2009, numero monografico curato da Janette Friedrich. Su Bühler da un punto di vista linguistico cfr. Albano Leoni (2011) e per i suoi rapporti con la scuola fonologica praghese cfr. Albano leoni 2009 [2011].

Innanzi tutto egli sottolineò l'importanza e la capacità esplicativa della psicologia della forma:

La psicologia però ha riconsiderato nel corso delle sue ricerche sul pensiero e della discussione sulla nozione di *Gestalt*, il problema del rapporto forma-materia: si tratta di mettere a frutto tale progresso nella teoria linguistica (Bühler 1934 [1983], p. 203).

In secondo luogo suggerì che la parola fosse caratterizzata innanzi tutto dall'avere un senso e poi da proprietà fisiognomiche peculiari:

Non è la fonologia, ma la grammatica o, più precisamente, la teoria lessicale quella che è legittimata a qualificare certe parti del flusso sonoro del discorso come parole e costituenti di parole […] Inoltre la psicologia moderna sottolinea vigorosamente la presenza, oltre che delle marche sonore = fonemi, di certe qualità gestaltiche nell'impronta sonora di queste forme […] In altre parole, ciascun termine presenta un aspetto sonoro che non è esclusivamente determinato dall'espressione, ma che indica in parte pure il valore simbolico e la valenza sintattica del termine (Bühler 1934 [1983], 228-229).

Questo punto di vista è ribadito con grande chiarezza a più riprese:

Si tratta del semplice fatto che nessun essere umano è in grado di distinguere migliaia di forme, caratterizzate, analogamente alle uova del nostro esempio, solo da combinazioni di notae, in un modo praticamente così agevole, rapido e sicuro qual è quello – basato sulle immagini sonore delle parole – di qualsiasi interlocutore normalmente esercitato di una comunità. È un'affermazione che invero non ho provato in modo sperimentale ma che ricavo da un'analisi dei meccanismi di riconoscimento nella lettura e da molti altri dati. Si tratta di un fatto che […] rinvia all'ampia efficacia esercitata dall'*aspetto* acustico delle immagini sonore con la loro funzione diacritica. *L'attuale fonologia ottempera al compito di una teoria diacritica sistematicamente costruita solo in un primo stadio d'avanzamento, mentre nel secondo dovrà ricevere lezioni dalla psicologia della* Gestalt (Bühler 1934 [1983], 339, ultimo corsivo mio).

Inoltre, come si vedrà nella prossima citazione, Bühler confutava con energia l'idea che la lingua fosse costruita montando insieme componenti che esistono al di fuori o prima dell'atto linguistico:

[…] l'opera linguistica in fieri non si sviluppa nel sistema psicofisico di un parlante in due tempi *successivi*, come avviene nella costruzione di una casa di mattoni, dove si ha prima la cottura dei mattoni e poi la costruzione dei muri. […] sta di fatto che a partire da Sweet la fonetica ha reso impossibile in linea di principio qualsiasi teoria di questo tipo (Bühler 1934 [1983], 316).

Infine, egli mostrava di essere ben consapevole del fatto che nel parlato avvengono naturalmente fenomeni di alterazione della materia fonica, e che ciò ha evidentemente conseguenze sul meccanismo della diacrisi, o sia della distinzione.

Questo fatto però diventa teoricamente fecondo se possiamo indicare con sufficiente precisione quali aspetti e costituenti della forma fonica vanno primariamente e massimamente soggetti, nelle circostanze menzionate, a indebolimento, logorio e distorsione. Dal punto di vista acustico sono i rumori, da quello fonetico sono i suoni esplosivi ad alterarsi prima di ogni altro suono […] risultano più resistenti i suoni vocalici, e, collegati ad essi, certi caratteri globali ben caratterizzati (qualità gestaltiche), come per es. la melodia […], inoltre, la struttura ritmica (forte – debole, breve – lunga) e infine le onde di acutezza e di saturazione della vocalità. *È un fatto che tali caratteri globali, nel loro insieme, sono sufficienti a soddisfare le esigenze diacritiche ridotte. Le immagini verbali saranno allora individuate preminentemente sulla base del loro aspetto acustico e in nessun caso in forza della sola segnalazione* (Bühler 1934 [1983], 341, corsivo mio).

Oggi sappiamo che queste manifestazioni di indeterminatezza del segnale linguistico sono molto più massicce e pervasive di quanto non pensasse lo stesso Bühler. Inoltre nel passo appena citato, con la menzione dei *caratteri globali*, è introdotto di fatto il principio di una pertinenza distintiva non affidata a un determinato segmento (che infatti potrebbe non essere fisicamente presente), come vorrebbero le fonologie segmentali, ma diffusa sull'intera stringa, in un modo analogo a quello che abbiamo visto per il riconoscimento visivo di volti. Il principio della distinzione, che, a ben guardare, corrisponde a quello saussuriano secondo il quale nella lingua non ci sono che differenze, e che certo non può essere messo in discussione, è dunque confermato, ma la sua manifestazione è qui configurata in modo originale e diverso da quello corrente.[6]

Ma, come si è detto, la teoria di Bühler è molto complessa e la prospettiva fisiognomica che qui si viene presentando va integrata con un altro strumento importante: il concetto di campo e la teoria dei due campi.

4. LA NOZIONE DI CAMPO.

Introdurre in linguistica la nozione di "campo", che Bühler ricava dalla psicologia della percezione, è certamente cosa nuova che implica una prospettiva dinamica non solo delle relazioni tra le unità, ma anche delle unità stesse. Basta pensare alla differenza tra l'idea che una percezione sia definibile a seconda delle condizioni al contorno, continuamente variabili, e il concetto corrente di "struttura", nel quale è vero che ogni entità è definita dai suoi rapporti con le altre e dalla sua alterità, ma queste entità, una volta definita la struttura, sono statiche e lo rimangono fino a che qualche evento non determini una riorganizzazione che porta a una nuova struttura. Nella citazione che segue è annunciata la teoria dei due campi:

> Il concetto di campo (Feldbegriff) qui adottato è stato introdotto dalla psicologia moderna […]. Noi […] ricaveremo in maniera puramente logica il campo d'indicazione e il campo simbolico del linguaggio dai più ampi ambiti delle condizioni che contribuiscono dovunque a determinare il senso linguistico. Che nel linguaggio non esista un unico campo, bensì due, è una teoria nuova. […] Ciò che Cassirer […] descrive come i due stadi di sviluppo del linguaggio umano, è una duplicità di momenti ineliminabilmente inerente a ogni fenomeno linguistico e che fa parte oggi come ieri del tutto linguistico. […] sosteniamo, in base alla *teoria dei due campi (Zweifelderlehre)*, che l'indicazione visiva e la presentazione in molteplici modi rientrano precisamente nell'essenza del linguaggio naturale, a cui non sono più estranee dell'astrazione e della comprensione concettuale del mondo. Questa è la quintessenza della teoria linguistica qui affrontata (Bühler, 1934 [1983], 44-45).

In sintesi, per Bühler la facoltà umana del linguaggio si estrinseca attraverso l'intreccio indissolubile (e non dunque la mera giustapposizione) tra un campo simbolico (le lingue in senso tecnico, come dispositivi arbitrari e, appunto, simbolici) e un campo deittico, o indicale. Quest'ultimo tuttavia non è da intendere solo nel senso della deissi in senso proprio, che sarebbe la *deixis ad oculos*, cioè l'indicazione materiale, gestuale di un oggetto materialmente presente e visibile per chi parla e per chi ascolta, ma anche nel senso della deissi

[6] Bühler non rifiuta i fonemi, ma li vede subordinati alla fisionomia, efficaci nei casi di difficoltà. Su questo v. Albano Leoni (2009). Vale la pena di ricordare che, mentre nessuna delle fonologie correnti sembra aver raccolto ed elaborato queste indicazioni, un punto di vista simile emerge, probabilmente in maniera indipendente da Bühler, in lavori di psicolinguistica dedicati al riconoscimento delle parole, come p. es. in Coleman (1998, 2002). Una rassegna di questi studi è in Nguyen (2005).

fantasmatica, che sarebbe, in senso lato, il rinvio al mondo conosciuto (e alla esperienza condivisa che ne hanno i parlanti) e all'universo di discorso (p. es. attraverso procedimenti anaforici e cataforici).

5. UN ESPERIMENTO.

Nel presentare le considerazioni che ho riportato, Bühler (1934 [1983], 339) si lamentava di non disporre ancora di risultati sperimentali a sostegno di quanto veniva dicendo. Oggi invece i dati sulla indeterminatezza fonica del parlato sono largamente disponibili e confermano pienamente il discorso di Bühler. Ora, se si ritorna con la mente a esperimenti del tipo di quello descritto in Albano Leoni &Maturi (1992 e poi ripetuto molte volte) lo si può facilmente rileggere alla luce delle variazioni nell'interazione fra i due campi.

L'esperimento mostra al di là di ogni ragionevole dubbio, che i presunti pezzi autonomi di lingua (singoli fonemi, sillabe, e spesso anche le parole), pur conservando esattamente la stessa materia fonica in tutte le tappe dell'esperimento, non hanno in realtà di per sé alcuna consistenza linguistica e la acquistano invece a mano a mano che intorno a loro si ricostruisce il tutto di cui sono parte e cioè li si reinserisce in una fisionomia complessiva. In effetti, si vede che solo un tutto semioticamente ben formato, risultante dalla sinergia dei due campi, è pienamente riconosciuto, come mostrano i diversi gradi di riconoscimento osservabili nelle diverse tappe dell'esperimento, in analogia a quanto avviene guardando il papero/coniglio.

6. FISIOGNOMICA E ECONOMIA.

Secondo l'ipotesi che qui sostengo, il riconoscimento di unità foniche significative della lingua parlata, avverrebbe in base a un procedimento di tipo gestaltico e non in base a una combinatoria di unità minime autonome. Detto in altre parole, ciò significa che ogni unità significativa della lingua, diciamo, per comodità, un monema, o una parola, avrebbe il suo volto.

Si potrebbe obiettare che poiché il numero dei monemi di una lingua è teoricamente illimitato e ne nascono continuamente di nuovi, ne consegue che noi ci troveremmo di fronte alla necessità di conoscere e ricordare un numero illimitato e in continuo aumento di volti fonici, e ciò sarebbe in apparente contrasto con le capacità della memoria umana. Il modello fisiognomico sarebbe dunque suggestivo, ma ingenuo e irrealistico. E' questa appunto la posizione di una delle voci più autorevoli della linguistica del Novecento. Scrive infatti Martinet:

> In vista della grande varietà e ricchezza del linguaggio umano, la doppia articolazione doveva per forza diventare un tratto del linguaggio umano : proviamo ad immaginare che cosa ci capiterebbe se dovessimo distinguere, sia quando parliamo che quando ascoltiamo, fra le migliaia di grugniti omogenei che ci occorrerebbero per ognuno dei nostri monemi se non esistesse la seconda articolazione. […]. In definitiva, sia la forma che il significato sarebbero in perenne stato di oscillazione e questo impedirebbe lo stabilirsi di unità discrete significative, cosa che i monemi delle nostre lingue realmente sono, grazie alle loro forme stabili e ben definite (Martinet, 1962, 45-46).

A prima vista il ragionamento è ineccepibile, ma, guardando meglio, si vede che esso riposa in un certo senso su un assioma che ha attraversato molta linguistica del Novecento, e che Chomsky (p. es. 1980) ha reso celebre: l'assioma della povertà dello stimolo (criticato da Lombardi Vallauri, 2004). In Martinet esso si manifesta nell'assunzione implicita del fatto che il parlante/ascoltatore debba ricavare ogni conoscenza linguistica del mondo dal

segnale fonico, e poiché questo segnale è fatalmente insufficiente, esso dovrà strutturarsi secondo un principio di economia combinatoria, di modo che i grugniti si trasformino in sequenze lineari di fonemi.

Prescindo qui dalla considerazione che le «forme stabili e ben definite» dei monemi sono un illusione, se per *forma* si intende la forma fenomenica che essi assumono nel parlato. Prescindo anche dalla considerazione che è illusorio pensare che l'aver identificato una successione di unità di seconda articolazione esaurisca il compito ermeneutico dell'ascoltatore (basterebbe ricordare gli innumerevoli casi di omofonia/omonimia).[7]

Mi soffermo invece su un punto che è di cruciale importanza. L'assioma è falso perché esso, come ho appena ricordato, riposa sull'assunzione implicita che le lingue siano solo dispositivi simbolici. Il che non è.

L'aspetto forse più importante del pensiero linguistico di Bühler, e che legittima l'idea di una dimensione gestaltica delle unità della lingua, è, come ho ricordato, la formulazione di una teoria dei due campi. Vale la pena di riportare per esteso ancora un passi di grande importanza.

> Il linguaggio umano quale sistema di rappresentazione, […] è il risultato di un certo sviluppo che rivela una sorta di progressivo affrancamento dall'indicazione e di progressivo distacco dalla raffigurazione imitativa. […]Forse *sopravalutiamo* l'affrancamento dal campo d'indicazione, forse *sottovalutiamo* il fatto dell'essenziale apertura, nonché l'esigenza, da parte di ogni rappresentazione linguistica di uno stato di cose, di integrare quest'ultimo sul piano conoscitivo. O, in altre parole, esiste forse una componente integrativa di tutto il sapere costituito linguisticamente che scaturisce da una fonte che non si riversa nei canali di un sistema simbolico linguistico e tuttavia genera un vero sapere (Bühler, 1934 [1983], 309).

In questo passo, e in altri simili, si vedono *in nuce*, anzi più che *in nuce*, molti degli sviluppi che, nei decenni successivi, e per vie in parte autonome e indipendenti (almeno stando alle citazioni esplicite), daranno luogo alla teoria degli atti di linguistici (Austin 1962; Searle 1969), alla pragmalinguistica (Schlieben Lange 1975), alla sociologia della conversazione (Goffman 1981), all'analisi della conversazione (Sacks, Schegloff e Jefferson 1974) e del discorso (Brown e Yule 1983), alla cosiddetta teoria H&H (Lindblom 1986, 1990, 1992), e che produrrà sorprendenti analogie perfino con la biologia della cognizione (Maturana e Varela 1980), almeno per quanto riguarda il ruolo centrale del contesto. Ma in Bühler, più che in ogni altro studioso di linguistica a me noto, fatta salva l'importante eccezione di Benveniste (De Palo, 2009, 110-111), l'integrazione tra i due campi attiene non alle condizioni d'uso della lingua ma alla natura profonda della lingua stessa fino nelle sue manifestazioni foniche.

In altre parole, la conoscenza del mondo e delle situazioni, condivisa dai parlanti, è parte integrante del funzionamento delle lingue, ed è appunto questa la cornice dentro la quale si situa la prospettiva fisiognomica, che viene così sgravata dell'eccessivo carico mnemonico che indebitamente le si rimprovera.

---

[7] Una volta identificate fonicamente le sequenze it. /para/ e /bara/, il lavoro ermeneutico è lontano dall'essere finito : infatti, che significa *porta* ? che significa *para* ? che significa *bara* ? e in che modo si risolve il problema di it. /melo'dia/? Sarà *melodia* o *me lo dia*? E così in innumerevoli altri casi in ogni lingua.

Una prima conseguenza è che la distintività non è una proprietà categoriale assoluta, ma è un valore variabile: infatti «le richieste diacritiche si riducono allorché la forma acustica di un termine è costruita *empraticamente*» (Bühler (1934 [1983], p. 341).

La seconda conseguenza investe il complesso dei processi cognitivi insiti nell'agire linguistico.

> Confrontiamo ancora una volta con il linguaggio il sistema a una classe di comunicazione simbolica del tipo dei segnali con bandierine […] Con l'ausilio d'un patrimonio di opportuni fonemi diacritici si poteva scambiare un numero praticamente sufficiente di «segnali» (manteniamo il termine) ; il procedimento è comodo e verosimilmente conciso. E presenta poi, certo, ancora altri vantaggi non riscontrabili nei sistemi di campo. Ma uno gli è sempre negato, mentre lo si ottiene subito con un sistema di campo: la possibilità, cioè, di descrivere in modo sufficientemente differenziato ed esatto l'illimitata molteplicità con un patrimonio circoscritto di convenzioni e corrispondentemente di forme linguistiche […] Le lingue umane, che oggi conosciamo, avanzano tutte la pretesa di essere dei sistemi simbolici «*produttivi*» e per ciò stesso «*universali*». […] Ci limitiamo a stabilire che per principio tale pretesa può essere accampata con buone prospettive soltanto da un sistema di campo. Un codice di simboli globali, scritto o non scritto, dev'essere delimitato […] a motivo semplicemente della limitata capacità della memoria umana. […] Infatti, se siamo tutti capaci di rappresentare continuamente, attraverso il linguaggio, cose nuove in modo comprensibile intersoggettivamente e praticamente all'infinito, ciò non è perché noi e gli altri siamo acrobati della mnemotecnica, ma è perché queste prestazioni non sono affatto richieste con un sistema di campo del tipo della lingua (Bühler (1934 [1983], 127-128).

Come si vede il concetto di economia proposto da Bühler non solo è diametralmente opposto a quello della doppia articolazione di Martinet,[8] ma è ben più potente perché riesce a conciliare l'indeterminatezza e variabilità costitutive del significante fonico delle lingue con la loro straordinaria efficacia comunicativa e rappresentativa.[9]

Vorrei richiamare l'attenzione sulla posizione di Lindblom. Le idee di questo grande fonologo in merito al ruolo del contesto, sono per certi versi simili a quelle di B., ma c'è una differenza che commenterò brevemente e che riguarda il concetto di 'riduzione'. Per farlo mi servirò di una analogia con il concetto di 'ellissi'.

L'ellissi è un dispositivo che tenta di spiegare le manifestazioni linguistiche che sembrano derogare da un presunto principio universale per cui ogni enunciato presupporrebbe un soggetto e un predicato (verbale) o una copula e un predicato nominale. Le sue manifestazioni sono tematizzate in forme che vanno dall'artifizio del "verbo (o soggetto) sottinteso", di scolastica memoria, all'artifizio delle regole di cancellazione della linguistica generativa nella fase trasformazionale, e in ciò appare la continuità di un punto di vista che si conserva intatto dalla retorica antica fino alla più popolare teoria linguistica del Novecento. Questo punto di vista presuppone che ogni sapere necessario per lo scambio linguistico tra umani debba passare attraverso una stringa logico-predicativa esplicita. E infatti ciò che le definizioni di 'ellissi' hanno in comune, come si legge in opere di riferimento anche molto diverse, è appunto il concetto di 'mancanza' (p. es. Beccaria, 1996, s. vv. *ellissi* e *cancella-*

[8] Un concetto di economia altrettanto angusto affiora in ambienti linguistici anche molto lontani da Martinet, come p. es. in Pierrehumbert 2001.

[9] Anche qui, volendo, si potrebbe ricorrere a qualche analogia con la conoscenza dei volti : potrei non riconoscere (o riconoscere a fatica e confusamente) il volto dell'impiegato del mio ufficio postale, se inopinatamente lo incontro al mare, ma non ho dubbi quando lo vedo al suo posto di lavoro. Anche questi sono effetti del campo.

zione; Crystal, 1993, *passim* e p.119 ; Lewandowski, 1984, s.v. *Ellipse*, più sfumato), a volte espresso tramite la locuzione equipollente «sostituente zero» (Simone, 1995, 228-230).

Ma esistono anche punto di vista che ribaltano completamente il problema e vedono nelle forme cosiddette 'ellittiche' una possibile forma normale, vista come il naturale risultato dell'integrazione tra il sapere contenuto nella stringa linguistica e il sapere contenuto nella nostra conoscenza del mondo condiviso nel quale agiamo. Leggiamo ancora un passo di Bühler:

> L'uomo adulto è certo un essere parlante, ma non un homo loquax nella misura in cui il fautore dell'ellissi sembra implicitamente supporre. Perché dunque parlare, quando senza la parola le cose, nella vita pratica, vanno altrettanto bene, se non meglio? Quando un segno linguistico diacritico viene inserito in un'azione, spesso non è necessario aggiungervi una serie di altri segni linguistici. […] Il campo periferico attivo in cui il segno si trova è in questo caso una prassi (Bühler, 1934, trad. it., p. 210).

Ancora più esplicito è Wittgenstein. Immaginando un muratore al lavoro che dice al suo assistente "lastra!", Wittgenstein risponde a chi vede qui l'abbreviazione della frase "portami una lastra":

> Ma perché non dovrei dire, viceversa, che la proposizione "portami una lastra!" è un prolungamento della proposizione "lastra!"? (Wittgenstein, 1953, § 19):

Il mio punto di vista sulle rappresentazioni fonologiche è molto simile a questo: anziché dire che le forme del parlato sono 'ridotte' 'underspecified' ecc., direi che le forme di citazione sono 'allungate", per dirla con Wittgenstein, o 'iperdeterminate' o 'overspecified'. Perché le forme cosiddette 'ridotte' sono la manifestazione normale del parlato e le forme cosiddette 'normali' o 'di citazione' sono quelle che si realizzano nelle condizioni artificiali della dimensione metalinguistica del laboratorio o di certe manifestazioni, pure artificiose, del parlato da parte di persone addestrate alla rappresentazione scritta delle lingue.

6. CONCLUSIONI.

Vorrei concludere sottolineando un punto teorico generale implicito nel modello fisiognomico che Bühler ha proposto : mentre per le fonologie la variabilità e l'indeterminatezza del significante fonico sono, se non una malattia, certamente un accidente (nel senso greco del termine) di superficie, che non tocca l'essenza invariabile delle entità, qui gli stessi fenomeni sono aspetti della fisiologia delle lingue e caratteri immanenti della loro natura. Infatti, a ben guardare, la variabilità e la fragilità dei suoni delle lingue fanno parte delle cose del mondo percepito e dunque sono cose da interpretare: la questione sarà dunque non quella di sapere se il segnale linguistico sia variabile o costante, integro o danneggiato, vago o determinato, ma di sapere che cosa gli serva di volta in volta per essere compreso.

Mi sembra che questo sarebbe un obiettivo interessante per una fonologia che si voglia linguistica.

BIBLIOGRAFIA

Albano Leoni, F. (2002), Sulla voce, in La voce come bene culturale (De Dominicis A., a cura di), Carocci, Roma, 41-65.

Id. (2003), I correlati spettroacustici di una "voce leggermente rauca, con un tono di sarcasmo quasi amaro". Fonetica e linguistica della parole, in Voce, canto, parlato. Studi in onore di Franco Ferrero (Magno Caldognetto E. & Cosi P. a cura di), Padova, Unipress, 31-36.

Id. (2007), Saussure, la sillaba e il fonema, in La lezione di Saussure. Saggi di epistemologia saussuriana (Elia A. & De Palo M. a cura di), Roma, Carocci, 56-85.

Id. (2009), Dei suoni e dei sensi. Il volto fonico delle parole, Bologna, il Mulino.

Id. (2009 [2011]), Karl Bühler et le Cercle Linguistique de Prague, Verbum, XXXI, 1-2, 89-114.

Id (2011), Attualità di Bühler, Paradigmi. Rivista di critica filosofica , XXIX, 3, 121-134.

Id. & Maturi P. (1992), Per una verifica pragmatica dei modelli fonologici, in La linguistica pragmatica, (Gobber G. a cura di), Bulzoni, Roma, 39-49.

Austin, J. L. (1962), How to Do Things with Words, Oxford, Oxford University Press.

Beccaria, G. L., a cura di (1996), Dizionario di linguistica e di filologia, metrica e retorica, Torino, Einaudi.

Bolinger, D. (1986), Intonation and Its Parts: Melody in Spoken English, Stanford, Stanford University Press.

Id. (1989), Intonation and Its Uses. Melody in Grammar and Discourse, London, Arnold.

Brown, G. & Yule G. (1983), Discourse Analysis, Cambridge, Cambridge University Press.

Bühler, K. (1934 [1983]), Sprachtheorie. Die Darstellungsfunktion der Sprache, Jena, Fischer (trad. it. Teoria del linguaggio. La funzione rappresentativa del linguaggio, Armando, Roma, 1983).

Chomsky, N. (1980), Rules and representations, Oxford,Basil Blackwell.

Chomsky, N. & Halle, M. (1968), The Sound Pattern of English, New York & London, Harper & Row.

Coleman, J. (1998), Phonological Representations. Their Names, Forms and Power, Cambridge, Cambridge University Press.

Id. (2002), Phonetic Representations in the Mental Lexicon, in Phonetics, Phonology and Cognition (Durand, J. & Laks B. a cura di), Oxford, Oxford University Press, 96-130.

Cruttenden, A. (1986), Intonation, Cambridge, Cambridge University Press.

Crystal, D. (1993), Enciclopedia Cambridge delle scienze del linguaggio. Edizione italiana a cura di P. M. Bertinetto, Bologna, Zanichelli.

De Palo, M. (2003), L'asymétrie du signe chez Saussure», in Ferdinand de Saussure (Bouquet, S. éd.), Paris, Editions de l'Herne, 246-259.

Ead. (2009), L'io, i sensi e il linguaggio. Dall'antipsicologismo alla semantica della persona, in L'immagine e i sensi (P. De Luca & F. Fimiani, a cura di), Milano-Udine, Mimesis, 95-111.

Durand, J. & Laks, B. eds (2002), Phonetics, Phonology and Cognition, Oxford, OUP.

Friedrich, J. & Samain D. éds (2004), Karl Bühler. Science du langage et mémoire européenne, Dossiers d'HEL n° 2 (supplément électronique à la revue «Histoire Epistémologie Langage»), Paris, SHESL, n.2 (http://htl.linguist.jussieu.fr/dosHEL.htm).

Gensini, S. a c. di (1998), La varietà delle lingue, Scandicci, La Nuova Italia.

Gibson, J. J. (1986), The Ecological Approach to Visual Perception, Hillsdale (N.J) – London, Elbaum (trad. it. Un approccio ecologico alla percezione visiva, il Mulino, Bologna, 1999).

Goffman, E. (1981), Forms of Talk, Philadelphia, University of Pennsylvania Press (trad. it. Forme del parlare, Bologna, il Mulino, 1987).

Halliday, M. A. K. (1967), Intonation and Grammar in British English, The Hague – Paris, Mouton.

Id. (1970) Intonation and meaning, in A Course in Spoken English. I. Intonation, Oxford University Press, London (poi in Id., System and Function in Language. Selected Papers, ed. by G. R. Kress, London, Oxford University Press, 214-234).

Jakobson, R. & Halle, M. (1956), Fundamentals of Language, The Hague, Mouton.

Ladd, D. R. (1996), Intonational Phonology, Cambridge University Press, Cambridge.

Lewandowski, T. (1984), Linguistisches Wörterbuch, 3 voll, Heidelberg, Quelle & Meyer.

Lindblom, B. (1986), On the Origin of Discreteness and Invariance in Sound Patterns, in Invariance and Variability in Speech Processes (Perkell J. S. &Klatt D. H. eds), Hillsdale & London, Lawrence Erlbaum Associates, 493-523.

Id. (1990), Explaining Phonetic Variations. A Sketch of the H&H Theory, in Speech Production and Speech Modelling (Hardcastle, W.J. & Marchal, A. eds), Kluwer, Dordrecht, 403-439.

Id. (1992), Phonological units as adaptive emergents of lexical development, in Phonological development: models, research, implications (Ferguson C., Menn L. & Stoel – Gammon C. eds), Timonium, York Press, 131-163.

Lombardi Vallauri, E. (2004), The relation between mind and language: The Innateness Hypothesis and the Poverty of the Stimulus, The Linguistic Review, 21, 345-387.

Martinet, A. (1969[4] [1960]), Eléments de linguistique générale, Paris, Colin (trad. it. Elementi di linguistica generale, Bari, Laterza, 1971 [1966]).

Id. (1962), A Functional View of Language, Oxford, At the Clarendon Press, (trad. it. La considerazione funzionale del linguaggio, Bologna, il Mulino, Bologna,1965).

Id. (1970), La linguistique synchronique. Études et recherches, Paris, Presses Universitaires de France.

Maturana, H. R. & Varela, F. J. (1980), Autopoiesis and Cognition. The Realization of Living, Dordrecht, Reidel (trad. it. Autopoiesi e cognizione. La realizzazione del vivente, Venezia, Marsilio, 1985 [2001]).

Miller, G. A. & Johnson-Laird P. N., 1976, Language and Perception, Cambridge (Ma), The Belknap Press of Harvard University Press.

Mulligan, K., 2004, «L'essence du langage, les maçons de Wittgenstein et les briques de Bühler», in Friedrich & Samain.

Nguyen N. (2005), Perception de la parole, in Phonologie et phonétique. Forme et substance (Id. *et alii* éds), Paris, Lavoisier – Hermes Science, 425-447.

Persyn-Vialard, S. (2005), La linguistique de Karl Bühler. Examen critique de la Sprachtheorie et de sa filiation, Rennes, Presses Universitaires de Rennes.

Pierrehumbert, J. (1980), The Phonology and Phonetics of English Intonation, PhD Dissertation (rist. Indiana University, Bloomington, 1987).

Ead. (2001), Exemplar dynamics: word frequency, lenition and contrast, in Bybee Joan, Hopper Paul (a cura di), Frequency and the emergency of linguistic structures, Benjamins, Amsterdam – Philadelphia, 137-157.

Prieto Luis (1969), La découverte du phonème. Interprétation épistémologique, «La Pensée», 148, 35-53 (trad it. in Lineamenti di semiologia. Messaggi e segnali, Laterza, Bari, 1971, 169-194).

Rossi Mario (1999), L'intonation. Le système du français : description et modélisation, Ophrys, Paris.

Rousseau, A., 2004, «L'éclectisme intellectuel et linguistique de Karl Bühler : de l'axiomatique aux schèmes cognitifs», in Samain & Friedrich.

Rousselot, l'Abbé, 1902, Principes de phonétique expérimentale, Paris-Leipzig, H. Welter.

Sacks, Harvey, Schegloff, Emmanuel A. & Jefferson, Gail (1974), A Simplest Systematics for the Organization of Turn-Taking for Conversation, «Language», 50, 696-735.

Saussure, F. de, 1922, Cours de linguistique générale, Paris, Payot (trad. it. a c. di Tullio De Mauro, Bari, Laterza, 1967. da cui cito).

Schlieben Lange, B. (1975), Linguistische Pragmatik, Kohlhammer, Stuttgart, (trad. it., Linguistica pragmatica, il Mulino, Bologna, 1980).

Searle, J., (1969), Speech Acts. An Essay in the Philosophy of Language, Cambridge University Press, Cambridge (trad. it. Atti linguistici. Saggio di filosofia del linguaggio, Boringhieri, Torino, Boringhieri).

Simone R. (1995), Fondamenti di linguistica, Bari-Roma, Laterza.

Sweet, H., Handbook of Phonetics, Oxford, Clarendon Press.

Trubeckoj, N. S. (1939), Grundzüge der Phonologie, Prague, «Travaux du Cercle Linguistique de Prague», VII (trad. it., Fondamenti di fonologia, Torino, Einaudi, 1971).

Wittgenstein, L. (1953), *Philosophische Untersuchungen*. Oxford, Blackwell (trad. it. *Ricerche filosofiche*. Torino: Einaudi, 1967).

FONETICA I

IL PROGETTO DI RICERCA LONGITUDINALE "INDICI FONETICI PREDITTIVI DI BALBUZIE CRONICA IN ETÁ PRESCOLARE": PRIMI RISULTATI

*Lenoci G., #Allegri S., °Bernardini S., #Chiari F., °Crivelli N., #Dadamo V., #de Biase M., ^Galatà V., °Pisciotta C., #Polesel L., #Stanchina S., °Stocco D., §Vayra M., ^Zmarich C.
*Scuola Normale Superiore, Pisa
#Università di Padova, Padova
°Centro Medico di Foniatria, Padova
^CNR-ISTC, Padova
§Università di Bologna, Bologna
*giovanna.lenoci@sns.it, °{sbernardini, cpisciotta, dstocco}@centrofoniatria.it,
^claudio.zmarich@pd.istc.cnr.it

1. SOMMARIO

La balbuzie è un disordine della fluenza della parola che è presente nell'1% della popolazione mondiale e secondo ricerche recenti compare all'età media di 33 mesi (Yairi & Ambrose, 2005). Gli studi epidemiologici registrano un alto tasso di guarigione spontanea: benché colpisca il 5% dei bambini prescolari, l'80% di essi guarisce spontaneamente entro il quinto anno dalla comparsa. Il restante 20% che, probabilmente a causa di una forte predisposizione genetica (Yairi & Ambrose, 2005), è destinato a cronicizzare, deve essere individuato e sottoposto a trattamento terapeutico il prima possibile, in modo che la guarigione possa essere rapida, completa e duratura (Starkweather, 1993).

Il presente lavoro illustra i primi risultati di un progetto di ricerca (CNR RSTL 995) volto all'individuazione degli indici clinici predittivi, di tipo percettivo e acustico, che dovrebbero permettere, già dalle prime fasi di comparsa del disturbo, di discriminare i soggetti candidati a cronicizzare rispetto a coloro che, invece, guariranno spontaneamente. Dei circa 40 soggetti finora reclutati, tutti con familiarità al disturbo, quelli studiati a tutt'oggi sono tre (Anna, Giuseppe, Alessandro). Di loro sono state selezionate e analizzate le registrazioni relative alle fasi significative di evoluzione del disturbo: il mese precedente la comparsa della balbuzie (per 2 dei 3 soggetti), il periodo dell'insorgenza, il periodo a distanza di 12 mesi e quello intorno ai 15 mesi (fine della fase di osservazione ed inizio dell'eventuale trattamento). Qui presentiamo i risultati relativi agli indici sperimentali *Profilo delle disfluenze* (Yairi & Ambrose, 2005) e all'analisi acustica del grado di coarticolazione intrasillabica, secondo il metodo delle 'Equazioni del Locus' (Sussman et alii, 1999; Zmarich e Marchiori, 2005).

I risultati evidenziano che per due dei tre soggetti il valore prognostico del *Profilo delle disfluenze* (remissione *vs.* cronicità) nel secondo semestre è simile a quello dei soggetti destinati a cronicizzare (Yairi & Ambrose, 2005). Il terzo soggetto (Anna), invece, a distanza di 12 mesi dall'insorgenza sarebbe, secondo il *Profilo delle disfluenze*, candidato a remissione. Con riferimento ai dati relativi all'analisi della coarticolazione intrasillabica, quest'ultimo soggetto e uno degli altri due (Alessandro) sarebbero destinatati a guarire spontaneamente dal disturbo poiché, a distanza di 12 mesi dall'esordio, il grado di coarticolazione nelle sillabe percettivamente fluenti si riduce rispetto alle tappe precedenti (come per i soggetti balbuzienti di Subramanian et alii 2003, che in seguito avrebbero recuperato

spontaneamente), mentre il terzo soggetto sarebbe destinato a cronicizzare poiché il grado di coarticolazione intrasillabica si mantiene su livelli alti per tutte le tappe d'età studiate.

2. INTRODUZIONE

L'ICD (*International Classification of Diseases*) definisce la balbuzie come un "parlato caratterizzato da frequenti ripetizioni o prolungamenti di suoni o sillabe, o da frequenti esitazioni o pause che ne interrompono il flusso ritmico. Dovrebbe essere classificato come disturbo solo nel caso in cui la sua gravità è tale da ostacolare marcatamente la fluenza del parlato" (ICD-10, F98.5, W.H.O, 2007)[1]. Le ripetizioni di sillaba, le ripetizioni e i prolungamenti udibili o silenti di fono sono considerati sintomi primari, e fanno parte delle cosiddette "disfluenze" (costituiscono, più precisamente, le "disfluenze da balbuzie", *cfr*. *Stuttering Like-Dysfluencies* o *SLD,* Yairi & Ambrose, 2005).

I più importanti risultati in ambito epidemiologico nello studio della balbuzie si devono ai contributi dell'attività trentennale di E. Yairi, riassunti in Yairi & Ambrose (2005). Secondo questi autori essa si manifesta essenzialmente in età prescolare: il 95% dei soggetti, infatti, inizia a balbettare entro il quinto anno d'età, nella finestra temporale che va dai 16 ai 66 mesi (media e mediana attorno ai 33 mesi), che coincide con un periodo altamente critico per il bambino perché caratterizzato dal contemporaneo sviluppo delle strutture anatomofisiologiche e delle capacità cognitive e motorie. Il tasso di prevalenza del disturbo ammonta all'1% nella popolazione mondiale, mentre il tasso di incidenza è del 5%. La differenza fra il tasso di incidenza (5%) e il tasso di prevalenza (1%) suggerisce che la maggior parte dei soggetti che è stata balbuziente in un dato momento della propria vita è poi guarita spontaneamente. Il recupero spontaneo, che può avvenire entro il quinto anno dalla comparsa (ed entro i 3 anni per il 75% dei soggetti) interessa, infatti, 4 bambini che hanno incominciato a balbettare su 5.

Un bambino che inizia a balbettare ha quindi il 20% di possibilità di cronicizzare il disturbo, probabilmente a causa di una forte predisposizione genetica (Kang et alii, 2010; Kraft & Yairi, 2012), e questa aumenta quanto più la remissione spontanea si allontana temporalmente dall'inizio. Il tempo trascorso dalla comparsa del disturbo rappresenta, dunque, un indice prognostico fondamentale della cronicità. In considerazione di questo fatto, e dell'altro fatto che una volta che la balbuzie è diventata persistente il trattamento terapeutico diventa lungo, difficile e quasi mai risolutivo, è pertanto essenziale individuare e sottoporre il prima possibile a trattamento terapeutico in modo elettivo i soggetti candidati a persistere, senza attendere la chiusura della finestra temporale dei 5 anni: solo così sarà possibile eliminare o ridurre i sintomi del disturbo e le relative conseguenze psicologiche e attitudinali che esso comporta. Per queste ragioni, e traendo ispirazione dalla strategia di ricerca illustrata da Yairi & Ambrose (2005), l'obiettivo del progetto di ricerca intitolato "Indici predittivi della balbuzie cronica in età prescolare" è quello di individuare nel comportamento dei soggetti che hanno iniziato a balbettare da poco (meno di 12 mesi), gli indici predittivi di tipo clinico (cioè validi e relativamente facili ed affidabili da rilevare) che permettano di discriminare quelli che risultassero positivi per riservare loro il trattamento terapeutico precoce, mantenendo semplicemente sotto osservazione quei soggetti candidati, invece, a guarire spontaneamente. Questi indici predittivi di natura comportamentale sono stati proposti da Yairi & Ambrose (2005) ed insieme a indici di natura biologica (essere di sesso

[1] Traduzione a cura di C. Zmarich.

maschile, avere precedenti familiari per la balbuzie) fanno parte dei cosiddetti fattori di rischio. Questi indici, sui quali si basa, attualmente, la decisione clinica di intraprendere il trattamento terapeutico sono, per il momento, solo probabilistici. È essenziale, pertanto, condurre ulteriori ricerche per convalidarne il valore prognostico.

Il progetto di ricerca succitato nasce grazie ad un finanziamento *una tantum* del CNR nel 2008 (cod. CNR RSTL 995) al Dr. Claudio Zmarich, ricercatore dell'ISTC-CNR di Padova, che aveva partecipato ad un bando del 2006 per il finanziamento delle 'Attività di Ricerca a Tema Libero'. Dopo che il finanziamento si è esaurito in seguito al pagamento di due annualità di borsa di studio, il progetto attualmente sopravvive solo grazie alla collaborazione del Centro Medico di Foniatria (CMF) di Padova (clinica specializzata nelle patologie del linguaggio, convenzionata col SSN, partner scientifico sin dall'inizio del progetto) e dell'A.I.BA.COM[2] (dal 2010), oltre che per il lavoro di tesisti e tirocinanti del corso di laurea in Logopedia (Università di Padova, Università di Ferrara), di Psicologia (Università di Padova) e di Lettere (Università di Bologna).

Per ragioni di tempo e di risorse economiche a disposizione (ambedue scarsi), il progetto si è finora concentrato sul reclutamento di soggetti balbuzienti con familiarità al disturbo (figli o fratelli di balbuzienti), residenti nel Veneto. Le probabilità di incominciare a balbettare per soggetti con storia familiare di balbuzie (che hanno quindi congiunti di primo grado balbuzienti cronici) è, infatti, sei volte maggiore rispetto a quei soggetti nati in famiglie di non balbuzienti (Kloth et alii, 1999). L'attuale *database* del progetto consta di 40 famiglie con genitori o fratelli balbuzienti nelle quali, all'epoca del reclutamento, c'erano bambini di età compresa fra i 12 e i 24 mesi. In base al protocollo della ricerca, al raggiungimento del 24° mese questi soggetti vengono sottoposti a video e audio registrazioni del loro eloquio spontaneo presso il loro domicilio e i genitori sono istruiti a mettersi in contatto con il Dr. Zmarich se il bambino inizierà in seguito a mostrare i primi sintomi della balbuzie. Solo in questo caso, infatti, il bambino verrà indirizzato al CMF per la diagnosi formale e per essere sottoposto ad una serie di accertamenti clinici, tra cui quello ORL, neurologico, logopedico, psicologico, da ripetersi alla fine del periodo di osservazione dopo 15 mesi, e contestualmente incomincerà ad essere registrato a intervalli regolari presso il suo domicilio da operatori inviati dall'ISTC-CNR. Questa parte del progetto infatti prevede che vengano effettuate delle registrazioni audio-video domiciliari ogni tre mesi (6 in tutto), che hanno lo scopo di raccogliere dati sullo sviluppo fonetico (tramite un test fonetico, Bonifacio e Zmarich, in preparazione, prima dei 40 mesi; Bortolini, 1995, dopo i 40 mesi), sullo sviluppo lessicale (tramite il PVB, Caselli, Pasqualetti e Stefanini, 2007), e sull'evoluzione della gravità della balbuzie (SSI-3, Riley, 1994). Per quanto riguarda le variabili sperimentali, il parlato connesso viene analizzato per calcolare il numero delle disfluenze da balbuzie su 100 sillabe, mentre l'elicitazione di produzioni ripetute di sillabe CV comincianti per occlusiva, insieme con quelle prodotte nel parlato connesso, permette la misurazione del V.O.T. (*Voice Onset Time,* o tempo di inizio della sonorità, *cfr*. Lisker & Abramson, 1964) e del grado di coarticolazione intrasillabico C-V (col metodo delle 'Equazioni del Locus', Sussman et alii, 1999; Sussman et alii, 2010).

Dopo 15 mesi dall'esordio, se il bambino sta ancora balbettando e se i genitori lo richiedono, viene iniziato il trattamento clinico. L'efficacia del trattamento, che viene eseguito presso il CMF e prevede un ciclo standard di 15 sedute, è valutata con test standard alla

[2] Associazione Italiana per la Balbuzie e la Comunicazione.

sua fine (dopo circa due mesi) e nel *follow-up* dopo altri 12 mesi dalla fine del trattamento. È previsto anche un gruppo che non riceverà trattamento (per i figli di quei genitori che non lo desiderano, o che non riescono a programmarlo dopo il periodo di osservazione).

Al momento in cui stiamo scrivendo, 39 bambini sono già stati registrati a 24 mesi e di questi 13 hanno cominciato a balbettare. Tra questi vi sono 3 soggetti (due maschi e una femmina), la cui balbuzie all'inizio è stata classificata di severità medio-grave, su cui è già stata compiuta un'analisi delle variabili sperimentali. Il livello di gravità della balbuzie viene stimato nel progetto attraverso l'uso di due scale: lo *Stuttering Severity Instrument* (Riley, 1994) e la *Illinois Stuttering Severity Scale* (Yairi & Ambrose, 2005). Questi strumenti puntano a rapportare i dati emersi dalla valutazione del bambino a quelli della popolazione balbuziente: il loro principale vantaggio consiste nel poter estrarre dai diversi parametri, alla fine della valutazione, un giudizio complessivo della gravità del disturbo su una scala qualitativa di livello. La *SSI-3* di Riley (1994) è uno strumento statisticamente valido ed affidabile: su due campioni non consecutivi costituiti da almeno 200 sillabe ciascuno, la *SSI-3* valuta tre diversi parametri quali la frequenza delle disfluenze (percentuale di sillabe balbettate), la loro durata (media in secondi dei tre più lunghi eventi di balbuzie) e le caratteristiche fisiche concomitanti (*cfr*. *physical concomitants*, Riley, 1994: 11). I punteggi rilevati per i diversi parametri vengono trasformati, attraverso delle *Tabelle di conversione,* in punteggi equivalenti a specifici livelli di gravità che possono essere espressi o in percentili o in giudizi qualitativi ('lieve, molto-lieve, moderata, grave, molto-grave'). *L'Illinois Stuttering Severity Scale* di Yairi & Ambrose (2005) non è un test standardizzato e commercializzato come l' *SSI-3*. È piuttosto uno strumento clinico interno al gruppo di ricerca di Yairi e collaboratori. Esso consiste nella classificazione di quattro parametri su due campioni non consecutivi di 100 sillabe ciascuno: frequenza (numero di *SLD* per 100 sillabe), durata (media delle cinque disfluenze più lunghe), tensione delle disfluenze e caratteristiche accessorie. Ai primi tre parametri viene affidato un punteggio che va da "0" (fluenza normale) a "6" (balbuzie grave). Le caratteristiche accessorie, invece, possono ricevere un punteggio che va da "0" a "1" e questo punteggio va aggiunto alla media dei primi tre parametri.

La valutazione per la prognosi di cronicità viene condotta attraverso l'analisi del cosiddetto *Profilo delle disfluenze* che si basa sulla classificazione delle disfluenze proposta da Yairi & Ambrose (2005) e secondo i quali le disfluenze individuabili nel parlato di bambini balbuzienti possono essere discriminate qualitativamente in due diverse categorie: le *Stuttering-Like Disfluencies* (*SLD*) e le *Other Disfluencies* (*OD*). Le *SLD* sono le vere e proprie disfluenze da balbuzie e comprendono tre principali tipologie: ripetizioni di parti di parola e di parole monosillabiche, prolungamenti e blocchi. Le *OD*, rilevabili anche nel parlato di persone normofluenti, includono le interiezioni, le ripetizioni di parole polisillabiche e di frasi, le revisioni e le frasi interrotte. Il *Profilo delle disfluenze* viene calcolato contando il numero di *SLD* in un campione rappresentativo di linguaggio connesso di circa 400 sillabe *fluenti* (vale a dire le sillabe dei target lessicali) ed esprimendolo in percentuale.

L'analisi dell'evoluzione temporale della percentuale di *SLD* costituisce un importante fattore prognostico: secondo Yairi & Ambrose (2005), infatti, un numero relativamente alto di *SLD*, ma non di *OD*, distingue i bambini balbuzienti dai non balbuzienti; inoltre, se questo numero si riduce a partire dal secondo semestre dall'insorgenza del disturbo, i soggetti interessati avrebbero più possibilità di guarire spontaneamente rispetto a coloro per i quali la percentuale di *SLD* rimane alta durante tutto il corso del primo anno. Questi ultimi quindi sarebbero i soggetti destinati a cronicizzare (vedi Tabella 1).

Gruppo soggetti		0-6 mesi	7-12 mesi	13-18 mesi	19-24 mesi	25-36 mesi	37-48 mesi	49-60 mesi
Cronici	**SLD**	11.31 (6.12)	9.76 (6.32)	7.82 (5.31)	7.34 (6.75)	7.93 (6.40)	5.85 (8.37)	3.61 (4.42)
	OD	4.64 (2.05)	5.41 (2.09)	5.42 (2.48)	5.75 (3.57)	7.49 (4.20)	5.54 (1.67)	6.38 (3.01)
Guariti	**SLD**	11.03 (6.74)	5.38 (4.37)	3.01 (2.65)	1.99 (1.51)	1.62 (1.56)	1.18 (0.81)	0.91 (0.64)
	OD	5.85 (3.00)	5.21 (2.34)	5.13 (2.92)	4.80 (2.25)	4.93 (2.98)	5.07 (2.06)	5.75 (2.51)
Controlli	**SLD**	1.42 (1.01)		1.11 (0.79)		1.08 (0.97)	0.93 (0.89)	
	OD	4.42 (2.27)		4.39 (1.60)		4.67 (2.17)	5.42 (2.02)	

Tabella 1: Valori medi e deviazione standard (tra parentesi) della percentuale del numero di *Stuttering-Like Disfluencies* (*SLD*) e di *Other Disfluencies* (*OD*) su 100 sillabe, in bambini con balbuzie persistente, con recupero spontaneo e nel gruppo di controllo (modificata da Yairi & Ambrose, 2005 p. 173).

Questi dati hanno portato Yairi & Ambrose (2005) a proporre, con il *Profilo delle disfluenze,* un possibile indice prognostico basato sulla variazione della percentuale delle *SLD* a partire dal settimo mese *post-onset*. Se questo indicatore venisse confermato da ulteriori studi, costituirebbe un’importante risorsa per differenziare la balbuzie cronica dalla balbuzie remissiva perché economico e di facile utilizzo nel *setting* clinico così da permettere, dunque, la rapida individuazione dei soggetti con predisposizione verso la balbuzie cronica a cui offrire immediatamente trattamento terapeutico.

In alternativa agli indici percettivi, che forniscono una valutazione qualitativa delle disfluenze, si è fatta strada l’idea che esse debbano essere considerate anche dal punto di vista fonetico (motorio), essendo la balbuzie un disturbo che può essere definito neuromotorio in quanto determinato da disfunzioni nei processi del controllo motorio del sistema pneumo-fono-articolatorio e, più precisamente, da un *deficit* nella coordinazione spazio-temporale dei movimenti articolatori coinvolti nella produzione del parlato fluente. Tutti gli studi riconducibili a questa impostazione, hanno infatti rilevato differenze significative tra il comportamento motorio (verbale) dei balbuzienti e quello di soggetti normofluenti, non solo durante il parlato disfluente, ma anche in quello fluente dal punto di vista percettivo (van Lieshout et alii, 2004; Namasivayam & van Lieshout, 2011).

Anche non considerando le ricerche di tipo cinematico, sono molte le ricerche di tipo acustico dedicate alla balbuzie (per una rassegna vedi Zmarich & Marchiori, 2005). Alcune tra le più significative sono state quelle relative allo studio della coarticolazione intrasillabica, attraverso l’analisi dell’andamento della seconda formante (F2), il cui valore è indice del luogo di occlusione nel cavo orale, lungo la direzione antero-posteriore (Fant, 1970, Recasens, 1999). Diversi studi hanno, infatti, registrato la presenza di forti anomalie nelle transizioni formantiche dei soggetti balbuzienti adulti (riassunti in Pisciotta et alii, 2010a).

Questi comportamenti articolatori anomali sono il frutto della difficoltà ad eseguire rapidamente compiti complessi o nel mantenere una corretta organizzazione spaziale degli organi articolatori, in quanto le variazioni rilevate della struttura formantica lungo i domini temporali e di frequenza riflettono proprio le dinamiche articolatorie (e nello specifico della lingua nel cavo orale) necessarie alla produzione linguistica. Lo studio di Robb & Blomgren (1997), il cui obiettivo era l'analisi ed il confronto della coarticolazione anticipatoria della vocale sulla consonante nelle sillabe CV prodotte fluentemente da soggetti balbuzienti e normofluenti adulti, ha evidenziato che le transizioni della F2 per i soggetti balbuzienti mostravano pendenze maggiori, ad indicare che la lingua, nel passare dall'articolazione del *target* consonantico a quello vocalico, si muove molto velocemente nel cavo orale. I due gesti articolatori vengono realizzati, dunque, secondo i due autori, con una coordinazione spazio-temporale ridotta (basso grado di coarticolazione). Sembrerebbe, dunque, che i soggetti adulti balbuzienti rispetto ai normoparlanti, per riuscire ad essere fluenti, eseguano movimenti articolatori più ampi e più veloci nel passaggio dal *target* consonantico a quello vocalico (van Lieshout et alii, 2004). La presenza di queste anomalie nelle transizioni formantiche ha portato gli studiosi a interrogarsi se esse possano essere una manifestazione diretta del disturbo (da intendersi come limitazione/restrizione del sistema articolatorio dei balbuzienti) o una strategia di adattamento ad esso. Il balbuziente che da anni, infatti, convive col proprio disturbo, potrebbe aver imparato a sviluppare delle strategie articolatorie 'alternative' che gli permettano di parlare fluentemente. Per queste ragioni, per riuscire a dare una corretta interpretazione ai fenomeni studiati, si è deciso di potenziare le ricerche sui soggetti balbuzienti prescolari: queste in teoria permetterebbero di intercettare il disturbo sul nascere prima che le strategie compensative siano diventate parte integrante del modo di parlare di questi soggetti. Gli studi compiuti sui soggetti balbuzienti in età prescolare sono relativamente scarsi ma quelli svolti finora hanno tutti evidenziato sottili differenze di tipo motorio nelle dinamiche articolatorie dei soggetti balbuzienti rispetto ai loro coetanei normofluenti (*cfr*. Pisciotta et alii, 2010a). Il potenziale valore prognostico di questo indice acustico è stato confermato dallo studio di Subramanian, Yairi & Amir (2003) nel quale è stata effettuata l'analisi del parlato percettivamente fluente di bambini balbuzienti registrati tra i 6 e i 12 mesi dopo la comparsa della balbuzie e poi controllati con un monitoraggio di diversi anni. Dal confronto dei dati ottenuti nella prima registrazione dopo l'insorgenza del disturbo con quelli di un gruppo di controllo, è emerso che un *deficit* della transizione formantica è significativo e forse centrale in presenza di balbuzie cronica. I parametri analizzati sono stati la durata della transizione (l'intervallo che intercorre tra il rilascio dell'occlusione consonantica (F2*onset*) e la fine della transizione (F2*vowel*) e l'estensione della variazione di frequenza nel passaggio dall'articolazione della consonante a quello della vocale. Se per il primo parametro non sono state evidenziate differenze significative fra i due gruppi, si è osservato, invece, che i soggetti che in seguito avrebbero cronicizzato, mostravano cambiamenti di frequenza significativamente minori (mediati sui valori assoluti delle estensioni in Hz, attraverso i diversi luoghi consonantici e contesti vocalici), probabilmente perché la lingua si muove in un *range* spaziale meno ampio (vedi Tabella 2).

Misura	Balbuzie persistente	Balbuzie a remissione spontanea	Controlli
Durate (ms)	58.7 (22.7)	59.0 (19.6)	66.1 (37.1)
Variazione in frequenza (Hz)	**395.7 (196.7)**	583.99 (229.7)	502.3 (232.1)

Tabella 2: Media e deviazione standard (tra parentesi) dei valori di durata (ms) ed estensione (Hz) delle transizioni di F2 dal locus consonantico all’inizio dello stato stazionario della vocale per i tre gruppi di dieci soggetti di Subramanian et alii (2003).

I risultati sperimentali di Subramanian et alii (2003) forniscono i presupposti per affermare che la ridotta coarticolazione rifletta una restrizione del sistema motorio dei balbuzienti piuttosto che un adattamento con cui costoro reagiscono al loro disturbo, in base alla considerazione che i soggetti studiati erano vicini all’insorgenza della balbuzie, di modo che non avessero ancora potuto sviluppare comportamenti di tipo reattivo. Questo dato possiede una grande potenzialità prognostica che, se confermata, potrebbe costituire un indice clinico.

Qui è opportuno precisare che la metodologia d’analisi impiegata da Subramanian et alii (2003) per lo studio della coarticolazione è diversa da quella delle ‘Equazioni del Locus’ usata da Sussman et alii (2010) e anche da noi, nel presente studio: nel metodo tradizionale, infatti, viene misurato il valore di F2 all’inizio e alla fine della transizione, definiti rispettivamente come la prima pulsazione glottidale dopo il rilascio dell’occlusione e il primo ciclo riconducibile allo stato stazionario della vocale. E’così possibile misurare oltre che l’estensione, anche la durata e la velocità di cambiamento della transizione. Le ‘Equazioni del Locus’, invece, poiché misurano i valori di frequenza di F2 all’inizio della transizione e a metà della vocale, non possono misurare la durata della transizione.

Nell’ottica di una valutazione globale e multifattoriale del disturbo è importante non sottovalutare gli aspetti psicologico - attitudinali che accompagnano il disturbo e che possono, secondo alcuni autori, contribuire al suo persistere (Yairi & Ambrose, 2005). È stato dimostrato, infatti, che a partire dai tre anni i bambini balbuzienti mostrano consapevolezza delle difficoltà di fluenza e attitudini più negative rispetto ai coetanei normofluenti. Un test di facile somministrazione, per valutare l’attitudine comunicativa dei soggetti prescolari è il *Kiddy-Cat* (Vanryckegem & Brutten, 2006; Bernardini et alii, 2009). Questo test prevede che venga chiesto al bambino prescolare che balbetta di rispondere a 12 domande dirette e poste verbalmente sulle sue idee ed emozioni riguardo al suo modo di parlare. Questo indice di tipo psicologico potrebbe contribuire a un quadro prognostico più dettagliato del disturbo dei singoli soggetti.

Nel protocollo del progetto, gli indici sperimentali del Kiddy-Cat, del Profilo delle disfluenze, del V.O.T. e della coarticolazione intrasillabica, sono valutati nel loro potere prognostico confrontando i valori ottenuti nella tappa dei 9 mesi (nella quarta seduta di registrazione domiciliare) con i valori ottenuti alla tappa dei 15 mesi (nella sesta ed ultima seduta di registrazione) e con i punteggi di gravità del SSI-3 (Riley, 1994) a 15 mesi.

3. PROCEDURA SPERIMENTALE

3.1. Soggetti

I soggetti sperimentali del lavoro che qui presentiamo sono tre: Anna, Giuseppe e Alessandro che al momento della stesura dell'articolo hanno rispettivamente 6 anni e 11 mesi, 7 anni e 8 mesi, 5 anni.

Tutti e tre i soggetti hanno una storia familiare di balbuzie. Anna, figlia di padre balbuziente, ha iniziato a balbettare all'età di 30 mesi. Giuseppe ha entrambi i genitori balbuzienti ed ha iniziato a balbettare all'età di 40 mesi. Questi due soggetti facevano parte di un progetto pilota precedente al progetto RSTL-CNR, per il quale si era iniziato a registrarli, mensilmente, a partire dal sesto mese dalla nascita. L'ultimo soggetto è Alessandro, con padre e fratello maggiore balbuzienti, che ha iniziato a balbettare all'età di 31 mesi e che è stato registrato due mesi dopo secondo il protocollo del progetto RSTL-CNR.

3.2. Stimoli e procedure

Per ciascuno dei tre soggetti sono state analizzate dalle 3 alle 4 registrazioni relative alle tappe significative del disturbo: di Anna abbiamo analizzato la registrazione a 29 mesi (tappa -1, relativa al mese precedente la comparsa della balbuzie), a 30 mesi (tappa 0), a 31 mesi (tappa +1) e a 43 mesi (tappa +12, a distanza di un anno dall'insorgenza). Di Giuseppe sono state analizzate la registrazione a 39 mesi (tappa -1), a 40 mesi (tappa 0), a 52 mesi (tappa +12) e a 60 mesi (tappa +20). Per Alessandro, sono state analizzate le seguenti registrazioni: a 33-35 mesi (tappa +2), a 41 mesi (tappa +10) e a 44 mesi (tappa +13). Le audio-video registrazioni, ciascuna della durata di circa 75 minuti, sono state effettuate a casa dei soggetti, tramite un registratore digitale Edirol R-09, collocato in un marsupio indossato dai bambini e con un microfono esterno Sony ECM DS70P tipo *Lavalier* attaccato al colletto della vestina del bambino e a una distanza di circa 15 cm dalla sua bocca. Oltre all'interazione spontanea madre-figlio, l'eloquio dei bambini è stato sollecitato tramite l'uso di un test fonetico per stabilire l'inventario fonetico dei soggetti, attraverso la denominazione di immagini che, a seconda dell'età del bambino, poteva essere il TFPI (*cfr*. Zmarich et alii in questo volume) o il PFLI (Bortolini, 1995).

3.3. Trascrizione fonetica

La prima fase dell'analisi è consistita nella trascrizione fonetica in IPA del segnale acustico acquisito a 44.1 kHz e 16 bit (eccetto i suoni di tipo vegetativo o di difficile identificazione). Successivamente, sulla base della trascrizione fonetica, sono stati individuati e salvati gli enunciati contenenti sillabe CV o CVC, trascritte con consonante occlusiva bilabiale (p/b), dentale (t/d) e velare (k/g) seguita da vocale, necessarie per l'analisi della coarticolazione. Per Alessandro, il soggetto studiato più di recente, è stato creato un *Textgrid* in PRAAT, composto da un *Tier* per la trascrizione ortografica delle sillabe potenzialmente utili all'analisi e da un secondo *Tier* per la trascrizione fonetica, in alfabeto SAMPA, dei foni effettivamente prodotti. A causa della relativa rumorosità del segnale la segmentazione tra foni e tra sillabe è stata svolta manualmente, apponendo dei confini all'interno dei *Textgrid*, sulla base dell'andamento delle formanti, in particolare di F2 (Salza, 1990).

3.4. Analisi acustica

L’analisi acustica della coarticolazione intrasillabica CV (con C= [p/b; t/d; k/g] e V= qualsiasi vocale) è stata effettuata solo sulle sillabe percettivamente fluenti (cioè quelle non affette da disfluenze nella valutazione degli ascoltatori) usando come indice l’andamento della transizione di F2 da C a V, con il metodo delle ‘Equazioni del Locus’ (Sussman et al., 1999). Le ‘Equazioni del Locus’ sono equazioni differenziate per locus consonantico che descrivono la retta interpolante i valori in frequenza (Hz) della seconda formante (F2) misurati all’inizio della transizione (F2*onset*) sul primo ciclo utile della vocale conseguente al rilascio dell’occlusione e al centro della vocale (F2*vowel*). Queste equazioni, secondo Sussman e collaboratori, permettono di quantificare il grado di coarticolazione anticipatoria. Ogni transizione di F2 fornisce, infatti, una coppia di valori frequenziali (F2*onset*, F2*vowel*) che disposti nel piano cartesiano, con il primo valore in ordinata ed il secondo in ascissa, permettono di individuare un punto in maniera univoca. L’insieme dei punti, relativi a una singola categoria di luogo (bilabiale, dentale e velare) si addensa lungo una distribuzione di valori che è interpolata dalla retta di regressione lineare descritta dalla formula: *F2onset=k*F2vowel+c* (Lindblom, 1963), con *k* e *c* costanti reali che rappresentano rispettivamente la pendenza della retta di interpolazione e l’intercetta con l’asse delle ordinate (Sussman et al., 1999), e dove il valore di *k* esprime il grado di coarticolazione.

Per Anna e Giuseppe l’analisi acustica, basata sullo spettrogramma delle sillabe CV, è stata effettuata con *PRAAT* applicando la tecnica FFT (*Fast Fourier Transform*) per ricavare gli effettivi valori in *Hz* di F2 dagli inviluppi spettrali estratti negli istanti temporali selezionati. Per ogni sillaba sono stati generati due spettrogrammi, uno a banda larga e l’altro a banda stretta, su ognuno dei quali sono stati individuati i due istanti temporali relativi al primo ciclo vocalico conseguente al rilascio del *burst* (per la consonante) e al centro della vocale (per la vocale). In corrispondenza dei punti rilevati su entrambi gli spettrogrammi, sono state eseguite le due sezioni spettrali (quella delle armoniche per lo spettrogramma a banda stretta e quella delle formanti per lo spettrogramma a banda larga). Dal confronto delle due sezioni spettrali è stato possibile individuare con chiarezza il valore della F2. L’applicazione della procedura suddetta è stata programmata da uno *script* di *PRAAT* (Petracco e Zmarich, 2007). Per Alessandro, invece, l’analisi acustica della coarticolazione è stata automatizzata grazie alla creazione di appositi *script* di PRAAT che sfruttavano la segmentazione e l’etichettatura delle sole sillabe del tipo CV e CVC. Gli s*cript* di *PRAAT* utilizzati per questa analisi sono stati due: il primo è servito al calcolo automatico dei valori di F1 e F2 tramite l’algoritmo LPC del menu *Formant* sulla porzione stabile della vocale e al calcolo del valore centrale e dei valori medi delle stesse formanti (al 30%, 50% e 70% della sua durata), oltre che delle durate di tutti i foni e di tutte le sillabe indicati nei *Tier*. Col secondo script, invece, è stato calcolato il V.O.T. delle sillabe iniziali di enunciato e la F2 della consonante (rilevata un ciclo glottico dopo l’inizio della transizione formantica da C a V). Una volta calcolati in modo automatico i valori di F2 relativi alla consonante e alla vocale (quello centrale) di ciascuna sillaba, essi sono stati messi in relazione per calcolare, attraverso le ‘Equazioni del Locus’ il grado di sovrapposizione spazio-temporale della vocale sulla consonante, per ogni luogo articolatorio.

4. RISULTATI

4.1 Indice percettivo: Il Profilo delle disfluenze

In Tabella 3 vengono riportati i punteggi relativi alla valutazione della gravità della balbuzie di Anna, e i punteggi relativi al profilo delle SLD.

ANNA	**TAPPA -1**	**TAPPA 0** (esordio)	**TAPPA +12**
SSI-3 (Riley, 1994)		Moderata (41°-60° percentile)	Moderata (41°-60° percentile)
Scala gravità *Illinois*		Moderata (4.99)	Moderata (4.99)
Profilo delle *SLD*	Fluente (0.25%)	Moderata (12%)	Moderata (6%)

Tabella 3: Valutazione della gravità della balbuzie e del Profilo delle disfluenze di Anna relativi alla tappa (-1), alla tappa (0) e alla tappa (+12)

Come si può osservare, secondo il *Profilo delle disfluenze* (*SLD*) il soggetto nella tappa (-1) presenta un punteggio molto basso che, se confrontato coi dati della Tabella 1, è tipico delle produzioni di soggetti normofluenti. Nella fase di esordio del disturbo (tappa 0) il soggetto riceve punteggi di gravità relativamente alti (gravità moderata) sia per l'SSI-3 di Riley (1994) che per la Scala di gravità di Yairi & Ambrose (2005). Anche per il *Profilo delle SLD*, il soggetto riceve un punteggio (12%) che rientra nella media dei soggetti che hanno iniziato a balbettare (*cfr*. Tabella 1). A distanza di un anno il livello di gravità del disturbo si mantiene costante rispetto alla fase precedente, ma il nostro indice predittivo registra una riduzione delle *SLD* (6%)[3], che potrebbe costituire una prognosi favorevole. Purtroppo per questo soggetto non disponiamo dei dati relativi al periodo di fine osservazione (+15) che costituirebbe un termine di confronto più affidabile per la prognosi.

Nella Tabella 4 vengono mostrati i dati raccolti per Giuseppe:

GIUSEPPE	**TAPPA -1**	**TAPPA 0**	**TAPPA +12**	**TAPPA +20**
SSI-3 (Riley, 1994)	Lieve (24°-40° perc.le)	Moderata (41°-60° perc.le)	Moderata (41°-60° perc.le)	Lieve (12°-23° perc.le)
Scala gravità *Illinois*	Lieve (2.92)	Moderata (3.71)	Moderata (3.00)	Moderata (3.00)
Profilo delle *SLD*	Lieve (3%)	Lieve (5.25%)	Lieve (4%)	Lieve (3.5%)

Tabella 4: Valutazione della gravità della balbuzie e Profilo delle disfluenze di Giuseppe relativi alla tappa (-1), alla tappa (0), alla tappa (+12) e alla tappa (+20).

A partire dal momento dell'esordio la gravità del disturbo si attesta su livelli moderati per entrambe le scale e si mantiene tale fino alla tappa (+12): a distanza di un anno, dunque, il livello di gravità non è migliorato. Secondo il *Profilo delle disfluenze* la percentuale di *SLD* è di livello lieve all'esordio (5.25%) e a distanza di un anno (4%). Questo valore però non è un indice prognostico positivo perché a distanza di 20 mesi (ben oltre dunque il periodo massimo di osservazione) la percentuale di *SLD* è ancora a 3.5%. Confrontando que-

[3] Tuttavia questo punteggio è ambiguo: possiamo osservare che esso rientra sia nel *range* di soggetti destinati a persistere che di quelli destinati a guarire spontaneamente.

sto valore con quelli della Tabella 1, osserviamo che si pone al limite superiore del *range* di soggetti che recuperano spontaneamente (tappa 19-24 mesi: media 1.99 e DS 1.51), ma è più caratteristico dei soggetti destinati a cronicizzare (media 7.34 e DS 6.75). Infatti, a tutt’oggi, secondo i genitori Giuseppe balbetta ancora, anche se in forma lieve.

Per Alessandro i risultati della valutazione della gravità e del *Profilo delle disfluenze* sono riassunti nella Tabella 5:

ALESSANDRO	**TAPPA +2**	**TAPPA +10**	**TAPPA +13**
SSI-3 (Riley, 1994)	Moderata (61°-77° perc.le)	Severa (89°-95° perc.le)	Severa (89°-95° perc.le)
Scala gravità *Illinois*	Moderata (3.7)	Moderata (4.5)	Moderata (4.8)
Profilo delle *SLD*	Moderata (20%)	Moderata (24%)	Severa (28%)

Tabella 5: Valutazione della gravità della balbuzie e Profilo delle disfluenze di Alessandro relativi alla tappa (0), alla tappa (+10) e alla tappa (+13).

Nella tappa (+2) il soggetto riceve per le due scale di gravità il punteggio “moderato” che tende a rimanere costante o a peggiorare (SSI-3, Riley,1994) a distanza di 10 mesi. Il *Profilo delle SLD* non lascia dubbi sulla prognosi di cronicità: già a partire dalla tappa (+2) il soggetto riceve un punteggio “moderato” di *SLD* (20%) che aumenta a distanza di 10 mesi registrando un valore ben superiore a quello medio dei soggetti destinati a persistere (Tabella 1). La prognosi di cronicità per questo soggetto è confermata dai dati ottenuti per la tappa +13 (periodo di fine osservazione): non solo per entrambe le scale di gravità la balbuzie del soggetto non è migliorata, ma anche per il *Profilo,* a distanza di 13 mesi, il soggetto riceve un punteggio di balbuzie ‘severa’ con un aumento della percentuale di *SLD* dal 20% al 28%. Il valore prognostico di cronicità del *Profilo* viene dunque confermato.

4.2 Indice acustico: Analisi della coarticolazione

Per quanto riguarda l’analisi della coarticolazione sono state prodotte, per ogni luogo articolatorio e per tutte le tappe selezionate di ciascun soggetto, le rette di regressione lineare. Nella Tabella 6 riportiamo i valori delle pendenze (*k)* e delle intercette *(c)* dei tre soggetti balbuzienti analizzati e quelli di 25 bambini normofluenti selezionati per fasce d’età confrontabili con quelle di Anna, Giuseppe ed Alessandro (4 dai 28 ai 30 mesi; 17 dai 39 ai 47 mesi; 4 dai 48 ai 51 mesi). Questi soggetti sono stati registrati nell’ambito del progetto “Migrazioni” del C.N.R, che comprende la somministrazione di una serie di test a bambini di diverse fasce di età, le cui risposte vengono registrate; i bambini sono stati scelti sulla base di un questionario compilato dai genitori, che esclude anomalie psicofisiologiche e accerta la crescita del piccolo in ambiente italofono. Uno di questi test riguarda la ripetizione dei 23 foni consonantici della lingua italiana, compresa la produzione di 36 consonanti occlusive (24+12) in posizione iniziale di parola, organizzate in 2 sessioni di foni occlusivi con 6 ripetizioni (4+2) per ogni fono, seguito dalle vocali /a/ e /i/.

4.3 Luogo di articolazione bilabiale

Per il luogo di articolazione bilabiale Anna a 29 mesi (tappa precedente la comparsa della balbuzie) presenta un grado di coarticolazione molto alto, sia in relazione alle sue tappe successive (30, 31 e 43 mesi) che rispetto ai soggetti normofluenti pari età (fascia d’età 28-29 mesi). Nella fase di comparsa della balbuzie (30-31 mesi) e a distanza di un anno (43

mesi) il valore medio di *k* (pendenza) si abbassa leggermente e si mantiene costante su livelli relativamente alti rispetto a quelli dei suoi coetanei normofluenti.

Per Giuseppe, invece, si osserva un *pattern* coarticolatorio leggermente diverso: nel mese precedente la comparsa della balbuzie il valore di *k* è molto basso, sia rispetto alle sue tappe successive, sia rispetto ai soggetti normofluenti pari età (fascia d'età 39-47 mesi). Quando la balbuzie insorge, invece, il grado di coarticolazione s'innalza rispetto alla fase precedente attestandosi, però, sempre su livelli bassi rispetto al gruppo di controllo. A partire dai 40 mesi il grado di coarticolazione del nostro soggetto continua ad aumentare: a 41 mesi (dopo un mese dall'insorgenza del disturbo) il valore medio della pendenza (*k*) inizia ad attestarsi vicino ai valori medi registrati per il gruppo di controllo, ma a distanza di 20 mesi (tappa +60) Giuseppe presenta un grado di coarticolazione relativamente alto rispetto ai soggetti normofluenti pari età (fascia d'età 48-51 mesi).

		BILABIALI			**DENTALI**			**VELARI**		
Soggetti	**Mesi**	**N. casi**	**k**	**c**	**N.casi**	**k**	**c**	**N.casi**	**k**	**c**
N. 4	**28-30**	31 (7.75)	0.626 (0.099)	592.8	35 (8.75)	0.471 (0.118)	1405.3	12 (4)	0.642*	747.5
N. 17	**39-47**	197 (11.58)	0.688 (0.195)	525.2	238 (14)	0.624 (0.124)	948.3	102 (14.05)	0.710 (0.186)	870.8
N. 4	**48-51**	48 (12)	0.497 (0.126)	988.8	51 (12.75)	0.566 (0.106)	1050.4	27 (6.75)	0.701 (0.258)	877.0
Anna	**29**	50	0.795	191.0	99	0.596	1011.0	41	0.862	326.5
	30	87	0.704	406.6	118	0.451	1354.9	51	0.466	1422.1
	31	92	0.699	408.4	71	0.560	1089.7	63	0.731	778.2
	43	71	0.701	356.8	102	0.498	1234.2	47	0.607	1088.8
Giuseppe	**39**	66	0.457	858.1	113	0.578	1090.7	66	0.742	556.5
	40	149	0.592	620.9	136	0.726	682.1	91	0.805	550.4
	41	123	0.662	452.3	116	0.741	642.1	120	0.933	311.3
	60	89	0.603	514.5	87	0.773	516.4	52	0.852	513.9
Alessandro	**33-35**	160	0.444	1291.0	301	0.300	2026.8	114	0.624	944.0
	41	64	0.342	1513.3	237	0.341	2008.2	65	0.547	1187.6
	44	60	0.463	955.8	128	0.271	2170.1	66	0.433	1483.6

Tabella 6: La parte in alto della tabella, relativa ai soggetti normofluenti, è divisa in tre fasce d'età per ognuna delle quali è indicato il numero di soggetti per fascia, il numero delle produzioni totali e la media (tra parentesi) per ogni luogo articolatorio, media e deviazione standard (tra parentesi) di *k* (pendenza) e la media di *c* (intercetta), v. testo. Per i soggetti sperimentali sono riportati il numero delle occorrenze, i valori medi delle pendenze *(k)* e delle intercette per i luoghi di articolazione bilabiale, dentale e velare (* valore di un singolo soggetto).

Per quanto riguarda Alessandro è possibile osservare che nei mesi immediatamente successivi alla comparsa della balbuzie (33-35 mesi) il bambino presenta un grado di coarticolazione relativamente basso rispetto ai suoi coetanei normofluenti (fascia d'età 28-30 mesi). A distanza di 10 mesi dalla comparsa, il valore medio di *k* si riduce ulteriormente attestandosi ben al di sotto della media del gruppo di controllo (fascia d'età 39-47 mesi, nella quale

soltanto un soggetto, presenta un valore per la pendenza inferiore). Nella tappa (+13) il valore della pendenza aumenta leggermente rispetto alle tappe precedenti rimanendo, però, sempre ben al di sotto della media dei controllo.

4.4 Luogo di articolazione dentale

Come per il luogo di articolazione bilabiale Anna a 29 mesi (prima di iniziare a balbettare) presenta un grado di coarticolazione relativamente alto sia rispetto ai suoi coetanei (gruppo di controllo della fascia d’età 28-30 mesi) che rispetto alle sue tappe successive. Quando compare la balbuzie (a 30 mesi) il grado di coarticolazione diminuisce e rientra perfettamente nei *ranges* normativi. A distanza di un mese il grado di coarticolazione si alza leggermente per poi ridiscendere (a 43 mesi) fino ad assumere un valore medio per *k* inferiore rispetto a quelli normativi (fascia d’età 39-47 mesi).

Per quanto riguarda Giuseppe, si osserva che nel mese precedente la comparsa della balbuzie, il bambino presenta un grado di coarticolazione più basso sia rispetto alle sue tappe successive che rispetto ai coetanei normofluenti (fascia d’età 39-47 mesi) e che esso aumenta bruscamente nelle fasi successive di comparsa del disturbo (40 e 41 mesi) per attestarsi su livelli superiori rispetto ai controllo. A distanza di 20 mesi il valore di *k* aumenta ulteriormente risultando maggiore della media dei coetanei normofluenti.

All’esordio della balbuzie Alessandro presenta un grado di coarticolazione molto basso (che si attesta su valori ben al di sotto della media dei soggetti normofluenti). Il grado di coarticolazione aumenta leggermente a distanza di 10 mesi rimanendo, però, sempre al di sotto della media dei controllo (fascia d’età 39-47 mesi). A distanza di 13 mesi il grado di coarticolazione diminuisce nuovamente assumendo valori fra i più bassi in assoluto rispetto a quelli rilevati per il gruppo di controllo.

4.5 Luogo di articolazione velare

Per il luogo di articolazione velare Anna segue lo stesso *pattern* rilevato per le dentali: nel mese precedente la comparsa della balbuzie il grado di articolazione è elevato non solo rispetto all’unico soggetto di controllo coetaneo, ma anche rispetto alle sue tappe successive. Quando la balbuzie insorge, il grado di coarticolazione diminuisce e si avvicina al valore medio del soggetto di controllo. A partire dal mese successivo il valore del grado di coarticolazione aumenta nuovamente rientrando nei *ranges* dei dati normativi.

Giuseppe, per questo luogo articolatorio, presenta nel mese precedente l’esordio della balbuzie, un grado di coarticolazione nella norma che aumenta nei due mesi successivi attestandosi su livelli ben al di sopra della media del gruppo di controllo (fascia d’età 39-47 mesi). A distanza di 20 mesi dall’esordio della balbuzie, il valore della pendenza rimane alto anche rispetto ai coetanei normali (fascia d’età 48-51 mesi).

Al manifestarsi della balbuzie Alessandro presenta un grado di coarticolazione confrontabile con quello dell’unico coetaneo normofluente. Col progredire del disturbo il grado di coarticolazione diminuisce e a distanza di 13 mesi Alessandro presenta un grado di coarticolazione di molto inferiore rispetto ai suoi coetanei (fascia d’età 39-47 mesi).

5. CONCLUSIONI

Scopo del presente contributo è innanzitutto fornire una presentazione accurata e aggiornata del progetto “Indici predittivi della balbuzie cronica in età prescolare” in corso all’ISTC-CNR. Le conclusioni delle analisi sperimentali fin qui condotte hanno semplicemente un valore esemplificativo. Benché i dati di 3 soggetti (peraltro incompleti) siano

chiaramente insufficienti per stimare il valore prognostico degli indici predittivi sperimentali, al momento ci sentiamo di riporre le nostre maggiori speranze sul *Profilo di disfluenza*, che oltre ad essere di più facile e rapida applicazione rispetto al *grado di coarticolazione intrasillabica*, è più in accordo con le misure di gravità, in particolare per quanto riguarda Alessandro, che da esse congiuntamente verrebbe qualificato come il soggetto maggiormente a rischio di cronicizzazione, mentre in base al grado di coarticolazione egli risulterebbe il soggetto a rischio minore. Però, a questo punto, prima di ridimensionare l'utilità dell'analisi della coarticolazione intrasillabica nello studio della balbuzie, è opportuno notare come i bassi valori manifestati da Alessandro, soprattutto nell'ultima tappa, per i luoghi di articolazione dentale e velare, siano comunque anomali rispetto a quelli dei coetanei normofluenti. Ricordiamo come la coarticolazione può essere definita come una strategia articolatoria per consentire all' apparato pneumo-fono-articolatorio di articolare il maggior numero di foni nella minor unità di tempo, cioè di essere più veloci. Una bassa coarticolazione, come quella manifestata da Alessandro, è compatibile con una bassa velocità di elocuzione, e infatti, proprio questo è il dato che emerge nello studio delle durate (Allegri, 2011), cioè un notevole rallentamento dell'eloquio, soprattutto nell'ultima tappa. C'è anche da dire che è probabile che Alessandro, per riuscire a mantenersi fluente, applichi solo una delle due strategie ipotizzabili, cioè il rallentamento, a scapito del cambiamento del grado di coarticolazione. Ricordiamo, inoltre, come l'allungamento delle durate segmentali possa mettere a disposizione degli articolatori un tempo maggiore per poter raggiungere il bersaglio articolatorio, e, dunque, per ridurre la coarticolazione. Un'altra ipotesi, non necessariamente alternativa, che potrebbe spiegare il basso grado di coarticolazione, è che siano state erroneamente selezionate tra le sillabe fluenti, anche sillabe che fluenti non sono.

Il parametro 'velocità d'elocuzione' in rapporto al grado di coarticolazione non è stato studiato e descritto sistematicamente in questo studio per due ragioni: la prima è che disponiamo dei dati relativi alle durate sillabiche soltanto per uno dei tre soggetti (Alessandro); la seconda ragione è che secondo Sussman et al. (1998), il valore di *k* è 'resistente' ai cambiamenti di velocità: " Speaking rate is another aspect of speaker-induced variation that appears to exert a limited effect on locus equation parameters" (Sussman et al., 1998, p. 248). Anche uno studio di Kugel et al. (1995) ha mostrato che non esiste alcun effetto significativo della velocità d'elocuzione sul grado di coarticolazione indicizzato dal valore di *k*. Tuttavia ci proponiamo, in futuro, di tener conto delle durate segmentali che sappiamo avere delle ricadute sulla fluenza: una minore velocità d'elocuzione, infatti, migliora la fluenza del parlato.

Per concludere, ci auguriamo che nel futuro possano nascere più ricerche sul comportamento linguistico di bambini balbuzienti prescolari di lingua madre diversa da quella inglese, (per poter osservare eventuali differenze interlinguistiche) anche se riteniamo che i risultati trovati finora negli studi americani siano validi anche per i soggetti italiani. Infatti i bambini balbuzienti che imparano a parlare lingue in cui siano presenti i tre *locus* consonantici considerati (come l'italiano e l'inglese), devono tutti, inevitabilmente rispondere ai vincoli neurofisiologici e biomeccanici richiesti per tali articolazioni (Petracco & Zmarich, 2007).

BIBLIOGRAFIA

Allegri, S. (2011), Indici acustici predittivi di balbuzie cronica in età prescolare: studio di un caso, Tesi di Laurea in Logopedia, Università degli studi di Padova, a.a 2010-2011.

Bernardini, S., Vanryckeghem, M., Brutten. G., Cocco, L. & Zmarich, C. (2009), Communication attitude of Italian children who do and do not stutter, J. of Communications Disorders, 42, 155-161.

Bloodstein, O & Bernstein Ratner, T. (2008), A Handbook on Stuttering, Chicago, Thomson Delmar Learning.

Boersma, P., & Weenink, D. (2010). Praat: doing phonetics by computer [Computer program]. Version, 5.1.32.

Bortolini, U. (1995), PFLI Prove per la valutazione fonologica del linguaggio infantile, Padova: Edit Master Srl.

Caselli, MC., Pasqualetti, P. & Stefanini, S. (2007), Parole e frasi del “primo vocabolario del bambino”: Nuovi dati normativi fra i 18 e i 36 mesi e forma breve del questionario, Milano, Franco Angeli.

Chang, S. E., Ohde, R. N. & Conture, E. G. (2002), Coarticulation and formant transition rate in young children who stutter, Journal of Speech, Language and Hearing Research, 45, 676-688.

Fant, G. (1970), Acoustic Theory of speech production, The Hague: Mouton.

Kang C., Riazzudin S., Mundorff J., Krasnewich D., Friedman P. Mullikin J.C. e Drayna D. (2010), Mutations in the Lysosomal Enzyme-Targeting Pathway and Persistent Stuttering, The New England Journal of Medicine, vol. 362, n. 8, 677-685.

Kloth, S.A.M., Kraaimaat, F.W., Janssen, P. & Brutten, G.J. (1999), Persistence and remission of incipient stuttering among high-risk children, J. of Fluency Disorders, 24, 253-265.

Kraft, S.J. & Yairi, E. (2012), Genetic Bases of Stuttering: The State of the Art, 2011, *Folia Phoniatrica et Logopedica*, 64, 34-47.

Kugel, K., Leishman, L.I., Bahr, R.H. & Montgomery, A. (1995), Procedural influences on the measurement of locus equations. Paper presented at the annual meeting of the American Speech-Language-Hearing Association, Orlando, Florida, December 7-10 [aHMS].

Lindblom, B. (1963), On vowel reduction. Stoccolma, The Royal Institute of Technology Speech Transmission Laboratory, Report No 29.

Lisker, L. & Abramson, A.S. (1964), A cross-language study of voicing in initial stops: acoustic measurements, Word, 20, 384-422.

Namasivayam A.K., van Lieshout P. (2011), Speech motor skill and stuttering, J. Motor Behavior, 43, 477-489

Petracco, A. & Zmarich, C. (2007), La quantificazione della coarticolazione nello sviluppo fonetico, in “Atti del III Convegno Nazionale AISV- Associazione Italiana di Scienze della Voce (ITC-IRST Povo - Trento, 29 Nov. – 1 Dic. 2006), Torriana (RN): EDK”, 135-150.

Pisciotta C., Marchiori M., Zmarich C., (2010a), Balbuzie e coarticolazione, in “Atti del V convegno AISV – Associazione Italiana di Scienze della Voce (Università di Zurigo, 4-6 Feb. 2009), Torriana (RN): EDK, 351-372.

Pisciotta C., Bernardini S., Agazzi A., Crivelli N., Manni F., Perosa R., Stocco D., Zmarich C. (2010b), Indici fonetici predittivi di balbuzie cronica in età prescolare: studio di un caso**,**

in “Atti del VI convegno AISV – Associazione Italiana di Scienze della Voce (Univ. di Napoli, 3-5 febbraio 2010), Torriana (RN): EDK, 293-317.

Recasens D. (1999), Acoustic analysis, in Coarticulation: Theory, Data and Techniques (W.J. Hardcastle & N. Hewlett), Cambridge (UK): Cambridge University Press, 322-336.

Riley, G. D. (1994), Stuttering Severity Instrument for Children and Adults-3 (SSI-3), (3rd ed.) Austin Tx.

Robb, M. & Blomgren, M. (1997), Analysis of F2 transitions in the speech of stutters and nonstutters, Journal of Fluency Disorders, 22, 1-16.

Salza, P. L. (1990), La problematica della segmentazione del segnale vocale, in Atti della 1° Giornata di Studio di G. F. S, Padova, 3-6 novembre.

Starkweather, C.W. (1993), Issues in the efficacy of treatment for fluency disorders, Journal of Fluency Disorders, 18, 151-168.

Subramanian, A., Yairi, E. & Amir, O. (2003), Second formant transitions in fluent speech of persistent and recovered preschool children who stutter, Journal of Communication Disorders, 36, 59-75.

Sussman, H. M., Byrd, C. T., Guitar, B. (2011), The integrity of anticipatory coarticulation in fluent and non-fluent tokens of adult who stutter, Clinical Linguistics & Phonetics, 25, 169-186.

Sussman, H.M., Duder, C., Dalston, E., Cacciatore, A. (1999), An acoustic analysis of the developmental of CV coarticulation: A case study, Journal of Speech, Language and hearing Research, 42, 1080-1096.

Sussman, H.M., Fruchter, D., Hilbert, J. & Sirosh, J. (1998), Linear correlates in the speech signal. The orderly output constraint, in Behavioral and Brain Sciences, 21, 241-299.

Van Lieshout P. H. H. M., Hulstijn, W., Peters H. F. M. (2004), Searching for the weak link in the speech production chain of people who stutters: A motor skill approach, in Speech motor control in normal and disordered speech (Maassen B., Kent R.D., Peters H. F. M., van Lieshout P. H. H. M., Hulstijn W.), Oxford: Oxford University Press, 313-355.

Vanryckeghem, M. & Brutten, G. (2006), The KiddyCAT: Communication attitude test for preschool and kindergarten children who stutters, Plural Publishing, San Diego, CA.

W.H.O. (2007), International Statistical Classification of Diseases and Related Health Problems (ICD), http://apps.who.int/classifications/apps/icd/icd10online/

Yairi E. & Ambrose N. (2005), Early Childhood Stuttering: for clinicians by clinicians, Austin, Pro-Ed.

Zmarich C., Fava I., Del Monego G., Bonifacio S. (in preparazione), Verso un “Test Fonetico per la Prima Infanzia”, VIII Convegno AISV, Roma, 25-27 gennaio 2012.

Zmarich, C., Marchiori, M. (2005), L’influenza del focus contrastivo sulla coarticolazione anticipatoria di sillabe “CV” prodotte fluentemente da balbuzienti e non balbuzienti, in “Atti del I convegno AISV– Associazione Italiana di Scienze della Voce (Univ. di Padova, 2-4 Dic. 2004), Brescia: EDK”, 231-250.

PRIMI DATI SULL'ACQUISIZIONE FONETICO-FONOLOGICA DELL'ITALIANO L2 IN PRESCOLARI RUMENI

*°Vincenzo Galatà, §Giada Meneguzzi, §Laura Conter, °Claudio Zmarich[1]
*Istituto di Ricerche sulle Attività Terziarie (IRAT), C.N.R. di Napoli
°Istituto di Scienze e Tecnologie della Cognizione (ISTC), C.N.R. di Padova
§Università di Padova
{vincenzo.galata, claudio.zmarich}@pd.istc.cnr.it

1. SOMMARIO

Nel presente lavoro riportiamo i primi risultati di un progetto di ricerca sull'acquisizione dell'italiano come Lingua Seconda (L2) da parte di bambini pre-scolari figli di immigrati, che si propone di chiarire i meccanismi che governano la percezione fonetico-fonologica e l'influenza esercitata dalla struttura della lingua materna di tali bambini sull'acquisizione dell'italiano, guardando alle loro abilità nel discriminare e produrre suoni consonantici dell'italiano (*cfr*. Galatà & Zmarich, 2011a, 2011b)[2]. In questa sede presentiamo i risultati di un'analisi quali-quantitativa delle produzioni verbali (ripetizione di non-parole) di un gruppo di 10 bambini rumeni di età tra 60 e 72 mesi che abbiamo confrontato con un campione di coetanei italiani, con l'intento primario di: a) verificare la validità dei test per la valutazione della competenza fonetico-fonologica dell'italiano L2 messi a punto all'interno del summenzionato progetto e somministrati ai bambini; b) verificare la capacità del gruppo di bambini rumeni di produrre e discriminare fonemi italiani non presenti in rumeno (nemmeno come allofoni); c) verificare le potenzialità del *software* PHON in ambito di L2 al fine di estendere le procedure di codifica dei dati all'intero campione dei bambini reclutati nell'ambito del progetto. L'analisi preliminare, sebbene condotta su un campione limitato, ha evidenziato un'effettiva influenza delle variabili linguistiche di L1 nell'acquisizione dell'Italiano L2 da parte dei bambini rumeni, supportando e giustificando la proposta di strumenti di valutazione specifici. Da questo punto di vista, i test proposti sembrano rivelarsi un valido strumento nel mettere in luce le difficoltà dei bambini stranieri in termini di acquisizione di L2: in base all'analisi degli errori segmentali e dall'esame degli esiti dei processi fonologici messi in atto dai bambini, i due gruppi evidenziano sostanziali differenze nei pattern di preferenza per determinate classi di fonemi. Infine, per le sue potenzialità, PHON si è confermato un ottimo ausilio alle analisi dei processi fonologici rivelandosi uno strumento valido anche nel campo dello studio dell'acquisizione di una L2.

2. INTRODUZIONE

Come già riportato in altre sedi (*cfr*. Galatà & Zmarich, 2011a, 2011b) la crescente presenza di bambini figli di genitori stranieri nei servizi educativi per l'infanzia e negli asili nido pone in primo piano il problema del loro sviluppo linguistico, dato che sono esposti,

[1] Benché il lavoro sia frutto di una stretta collaborazione tra gli autori, tuttavia, ai soli fini accademici, l'impostazione della ricerca è da attribuire a V. Galatà e C. Zmarich; il § 2 è da attribuire a CZ, il §§ 3 e 7 sono da attribuire a VG e CZ, i §§ 5 e 6 sono da attribuire a VG, CZ, GM e LC; i restanti §§ a VG.

[2] Progetto CNR IC.P10 "Migrazioni", attualmente *in fieri.*

sin dalle prime fasi di acquisizione della lingua parlata, a sistemi linguistici diversi in contesti sociali e culturali differenti. Sebbene le più recenti indicazioni europee e i risultati di diversi studi condotti in vari ambiti di ricerca rivelino la necessità di realizzare contesti educativi attenti alle diversità linguistiche e culturali già nei primi anni di vita (Abdelilah-Bauer, 2008), le esperienze di ricerca e di pratica educativa e psico-sociale, destinate ad accogliere i bambini di lingue e culture diverse all'interno dei servizi per l'infanzia, sono assai ridotte.

Il periodo che va dai 3 ai 6 anni rappresenta una fascia di età che, a livello neurobiologico, è considerata privilegiata per l'acquisizione delle lingue (Bates, 1995; Birdsong, 1999).[3] A livello di organizzazione cerebrale, infatti, la fissazione dei circuiti neuronali di una lingua determinata non è, almeno fino agli 8 anni, irreversibile (*cfr*. Pallier *et alii*, 2003): anche nei casi in cui un bambino si venisse a trovare nella situazione eccezionale di dover abbandonare la lingua madre e apprenderne una nuova (per esempio nei casi di adozione), questi sarebbe in grado, tra i 3-8 anni, di impararla come se si trattasse della lingua madre.[4]

A seconda dell'età e dell'ambiente di appartenenza, si usa distinguere tra: bilinguismo *precoce e simultaneo*, se due lingue sono presenti nell'ambiente di un bambino dalla sua nascita; bilinguismo *precoce e consecutivo*, se la L2 è introdotta nell'ambiente del bambino dopo i 3 anni; infine, se il contatto con la L2 avviene dopo i 6 anni, si parla di bilinguismo *tardivo*. Nel bilinguismo si parla poi di *equilibrio* o *disequilibrio*, considerando il fatto che una lingua può essere più o meno forte e prevalente, a scapito dell'altra che può diventare debole e minoritaria. Si avrà quindi un bilinguismo *additivo* se tutte le competenze cognitive sono sviluppate in entrambe le lingue, *neutro* o *sottrattivo* se tali competenze sono sviluppate in modo ineguale (*cfr*. Bettoni, 2001).

I soggetti maggiormente studiati nelle ricerche condotte in questo ambito, nonostante il fattore età sia tra quelli ritenuti più importanti per la variabilità intersoggettiva, sono gli adulti. Ciò è in parte dovuto al fatto che gli studi che esaminano i bambini vengono fatti rientrare di più nel campo del bilinguismo *precoce* che in quello dell'apprendimento di L2. Gli studi che si sono occupati del bilinguismo dalla nascita, sebbene ancora pochi rispetto a quelli relativi al bilinguismo *precoce e consecutivo* e a quello *tardivo*, hanno tuttavia evidenziato che l'acquisizione delle due lingue è simultanea (Werker & Byers-Heinlein, 2008) e non l'una conseguente all'altra, come accade per definizione per la L2. Bettoni (2001) individua, infatti, tre criteri per i quali solitamente L2 si differenzia da L1: la *cronologia* (la si impara dopo L1), la *competenza* (la si conosce meno bene) e l'*uso* (la si usa meno spesso).[5]

[3] Mentre gli adulti, già cognitivamente e socialmente maturi, apprendono una L2 grazie a strategie tipo *problem solving* e utilizzando la grammatica della L1, per giungere a una approssimazione della competenza linguistica del parlante nativo, i bambini sfruttano le loro capacità innate di acquisizione di una lingua (Bley-Vroman, 1989).

[4] Questo dimostrerebbe anche il perché i bambini siano in grado di acquisire l'inventario consonantico di una L2 in modo così rapido e con prestazioni che si avvicinano a quelle della L1 (*cfr.* Anderson, 2004; Gilhool, Burrows, & Goldstein, 2009).

[5] Nella realtà dei fatti tali differenze non sono così nette, a causa, per esempio, dell'età di acquisizione di L2.

Tra i fattori che condizionano l'acquisizione di L2[6] troviamo sicuramente il grado di parentela linguistica e/o affinità culturale e quindi il sistema della L1 sottostante: la distanza strutturale tra L1 e L2, può infatti fungere da elemento di facilitazione o di difficoltà (Valentini, 2005). Analoga considerazione vale per il grado di apprendimento della L2: la vicinanza strutturale facilita l'apprendimento, ma favorisce l'interferenza, mentre la distanza strutturale implica l'acquisizione di strutture diverse (con tempi iniziali più lunghi), ma riduce l'interferenza.[7]

I fattori di tipo linguistico ed extralinguistico che possono condizionare la fluenza di L2, possono combinarsi in svariati modi, per cui il loro effetto complessivo sulla pronuncia di un individuo deriva da combinazioni diverse da caso a caso. C'è da aggiungere, inoltre, che gli esiti di questi studi dipendono non solo dalla difficoltà di gestire tutte queste variabili (che portano ad avere campioni di soggetti studiati con caratteristiche diverse e quindi non più confrontabili), ma anche dalle diverse tecniche di elicitazione dei campioni di parlato e dalle varie tecniche di valutazione usate (Piske *et alii*, 2001).

3. LO SVILUPPO FONETICO-FONOLOGICO NEI BAMBINI

È provato che per la formazione delle categorie fonetiche della lingua madre i bambini si basano sulle regolarità distribuzionali della lingua, ma i bambini bilingui devono affrontare contemporaneamente i problemi distribuzionali di due lingue, ognuna con le proprie regole. Sul modo in cui si formano, nei soggetti bilingui, queste categorie fonetiche esistono ancora pochi studi, tra l'altro giunti a risultati talvolta difficili da interpretare se confrontati tra di loro, a causa della grande eterogeneità che esiste tra gruppi di bilingui e tra i singoli individui (Werker & Byers-Heinlein, 2008). È tuttavia certo che la costruzione di queste categorie fonetiche avviene nel corso del primo anno di vita e che i bilingui adulti fluenti sanno discriminare i suoni di ogni loro lingua, quando le hanno acquisite entrambe dalla nascita, anche se spesso la prestazione è migliore nella lingua dominante. Tuttavia, se una lingua è stata acquisita dopo un'altra, per L2 vi sono distinzioni fonetiche per cui dimostrano scarse abilità di discriminazione (Bosch, Costa, & Sebastián-Gallés, 2000; Sebastián-Gallés & Bosch, 2005).

3.1 Lo sviluppo percettivo

Affinché il bambino possa imparare le unità di significato del linguaggio è necessario che sia in grado di decodificare e associare le caratteristiche acustiche, e visive, della produzione linguistica dell'adulto. Tali capacità sono presenti già dalla nascita e permettono la formazione di categorie cognitive di suoni e unità linguistiche (*cfr*. Zmarich, 2010).

Il processo percettivo più strettamente correlato con lo sviluppo articolatorio, e che sembra comparire per primo, riguarda gli indici di discriminazione acustica utilizzati per distinguere i suoni linguistici. I bambini, a differenza degli adulti che li discriminano in modo categoriale, sono dotati della capacità di discriminare molti, se non tutti, gli indici acustici presenti nei suoni linguistici. Questa abilità di estrarre dal segnale acustico le informazioni

[6] Altra variabile molto studiata (negli adulti) è l'esperienza in L2, intesa come numero di anni trascorsi in una comunità dove L2 rappresenta la lingua dominante, ma le numerose ricerche condotte hanno però portato a risultati contrastanti: alcuni la ritengono importante per l'accuratezza nella pronuncia di L2, altri meno (Piske, Mackay, & Flege, 2001).

[7] L'interferenza è più rilevante in fonologia e semantica-lessico, meno in sintassi e morfologia (*cfr.* Mioni, 2005).

corrispondenti a caratteristiche fonetiche universali, diminuisce sensibilmente verso gli 8 mesi e scompare quasi del tutto verso la fine del primo anno di vita; perciò quando i bambini cominciano a produrre le prime forme di linguaggio, la capacità di discriminare differenze fonetiche si indirizza verso la specifica fonologia della lingua madre (Mattock, Amitay, & Moore, 2010). Recenti ricerche hanno dimostrato che la finestra temporale per l'acquisizione del linguaggio ha in realtà un inizio, un'ampiezza e una fine molto più variabili di quanto si pensasse inizialmente (quando veniva definita "periodo critico"): oggi si preferiscono i termini "periodo sensibile" o "periodo ottimale" intendendo un periodo biologicamente (ed esperienzialmente) determinato, in genere presto nella ontogenesi, durante il quale alcuni aspetti del funzionamento neurale e comportamentale dell'organismo sono *particolarmente* sensibili a certi fattori ambientali (Tees, 2001). Esistono inoltre prove che supportano l'ipotesi dell'esistenza di uno o più periodi ottimali per l'acquisizione del linguaggio. Tra i vari domini linguistici risultano esserci, per esempio, differenze nell'impatto dell'età di acquisizione: alcuni aspetti del linguaggio, come la sintassi, la morfologia e la fonologia, mostrano tempi di apertura e chiusura alle influenze esperienziali ristretti, mentre altri, come l'acquisizione del lessico, mostrano una relativa apertura per tutta la durata della vita (Johnson & Newport, 1989). Per quanto riguarda la fonologia, è poi probabile che esistano diversi periodi ottimali per l'acquisizione delle sue diverse componenti (es. discriminazione dei segmenti fonetici, degli allofoni, delle regole fonotattiche). Lo sviluppo percettivo è condizionato da due fattori: la progressiva maturazione della porzione del sistema nervoso preposto alla percezione linguistica e l'influenza esercitata dall'ambiente linguistico circostante il bambino. Infatti, quando una particolare abilità coincide con la stimolazione linguistica che il bambino riceve, egli manterrà o migliorerà tale abilità attraverso la pratica nella percezione e produzione dei fonemi specifici; quando, invece, l'abilità di discriminare (pur presente alla nascita) non coincide con il linguaggio dei genitori, questa necessiterà di un aggiustamento che Aslin & Pisoni (1980) definiscono *riallineamento* (*attunement*). Un'altra possibilità è rappresentata dall'*attenuazione* o *perdita* della capacità di discriminare una determinata opposizione fonologica, nel caso in cui questo manchi nella lingua del contesto in cui è inserito il bambino. Le capacità discriminanti non presenti alla nascita devono, invece, essere acquisite. Attraverso questo processo di categorizzazione, che rinforza le connessioni fra i tratti presenti nei fonemi della lingua adulta, il bambino verso i 12 mesi ha sviluppato un grado di categorizzazione fonemica vicino a quello dell'adulto. Il declino della *performance* sulla discriminazione dei contrasti fonetici non appartenenti alla lingua nativa non è comunque assoluto, ma una certa sensibilità latente continua a esistere, anche se non ai livelli che mostrano i parlanti nativi (Polka, 1992): il mantenimento dato dall'esperienza dovrebbe perciò essere inteso come il risultato di un'organizzazione, piuttosto che una perdita della sensibilità percettiva iniziale (Lalonde & Werker, 1995).

La capacità di riconoscere ciò che è familiare da ciò che non lo è in stimoli di tipo linguistico, è sicuramente un punto di partenza importante nel processo di acquisizione del linguaggio, ma non è sufficiente per consentire di ottenere tutte le informazioni che sono indispensabili per cominciare a parlare: un'altra abilità che si sviluppa parallelamente alla capacità di discriminazione fonetica è quella di segmentazione del *continuum* dei suoni percepiti nelle unità che costituiscono il linguaggio (parole, sintagmi, frasi). Dopo ripetute esposizioni a una lingua però, è possibile riconoscere delle regolarità nella frequenza con cui certi insiemi di suoni si presentano, si inizia cioè a capire quelle che sono le *proprietà distribuzionali* della lingua. Secondo alcune recenti prospettive, i nostri meccanismi di rile-

vazione delle regolarità sono così potenti che possono guidare efficacemente il bambino nell'apprendimento del linguaggio e nella scoperta delle regole *fonotattiche* della propria lingua madre. La sensibilità del bambino a queste strutture gli fornisce una regola, nella maggior parte dei casi valida, per ricavare il confine di parola anche in assenza di qualsiasi informazione sul significato: il confine di parola si colloca dove c'è una violazione di una regola fonotattica (*cfr*. Werker & Tees, 2005). È solo intorno ai 5-6 anni che inizia ad apparire la conoscenza dei fonemi come unità discrete, combinabili e commutabili, che rappresenta contemporaneamente un prerequisito e un prodotto dell'alfabetizzazione. Per Werker & Tees (2005) ciascuno dei passi appena descritti, relativi all'evoluzione della sensibilità e all'uso delle categorie fonologiche, rafforza l'organizzazione percettiva iniziata nella prima infanzia, rendendola più resistente al cambiamento. Gli stessi autori sostengono, inoltre, che i cambiamenti indotti dall'esperienza a ogni livello (categorie fonetiche, categorie fonologiche, items lessicali-semantici, lettura e scrittura) influenzano sia quei componenti già sviluppati che quelli non ancora emersi.

3.2 Lo sviluppo della produzione orale

L'evoluzione delle capacità di produzione dei suoni può essere considerata parallela a quella delle capacità percettive, anche se sussistono molteplici e reciproche influenze. É però possibile identificare delle linee generali nello sviluppo fonologico e riferirle alle più generali tappe dello sviluppo linguistico e cognitivo, sebbene vi sia una certa variabilità individuale nei tempi di inizio di ciascuno stadio (inteso come periodo in cui si palesa l'insorgenza di comportamenti vocali non osservati in precedenza): tali stadi non sono però tra loro discreti e le vocalizzazioni tipiche di un periodo possono continuare anche in quello successivo (*cfr*. Zmarich, 2010).

Dai 18 mesi fino ai 4 anni si assiste alla fase dello *sviluppo fonemico* (*cfr*. Stoel-Gammon & Dunn, 1985; Vihman, 1996; Zmarich, 2010). Tra i 18 e i 20 mesi ha luogo lo sviluppo del vocabolario che si associa a un marcato cambiamento del sistema fonologico del bambino: il bambino, costretto dalla rapida crescita del vocabolario, adotta un approccio basato sulle corrispondenze segmentali con le parole adulte. Iniziano quindi ad apparire parole multisillabiche, gruppi consonantici e il numero dei suoni prodotti cresce, anche se il sistema fonetico adulto non è ancora completamente acquisito. In questo stadio i bambini smettono di evitare di produrre quelle parole contenenti strutture fonologiche non presenti nelle parole del loro lessico concettuale mentale (fenomeno noto come "word avoidance", *cfr*. Schwartz & Leonard, 1982). Cercando di produrre parole più complesse compaiono gli errori (o processi) di semplificazione, che essendo simili in tutti i bambini suggeriscono uno sviluppo di una organizzazione simile.[8] Benché esistano numerose classificazioni di questi errori, tutte includono processi che fondamentalmente possono essere distinti in: *processi di struttura*, che semplificano la struttura fonotattica, cioè i foni in contesto, e *processi di sistema*, che semplificano il sistema fonologico eliminando i contrasti (Bortolini, 1995). *I processi di struttura*[9] non riguardano il modo in cui viene articolato il singolo fono, ma ri-

[8] Secondo Ingram (1976) lo sviluppo fonologico deriva dalla naturale soppressione dei processi fonologici innati, ossia di quei meccanismi che neutralizzano certe sillabe o distinzioni fonologiche tipiche dell'adulto.

[9] I tipi di errori più diffusi nella semplificazione della struttura fonotattica sono: cancellazione della sillaba non accentata, riduzione di gruppi consonantici, processi di assimilazione o armonia, riduzione di dittonghi a un solo elemento vocalico, cancellazione di conso-

guardano le operazioni che vengono compiute all'interno della parola in termini di restrizioni fonotattiche, ossia restrizioni che vengono applicate alle normali combinazioni dei segmenti in relazione alla loro co-occorenza nella sillaba o nella parola. Questo tipo di processo si riscontra facilmente quando i bambini tentano di produrre sequenze sillabiche più complesse di CV (consonante/vocale): nei bambini l'organizzazione sillabica progredisce da strutture semplici CV, o sue duplicazioni secondo il modello CVCV, fino ad arrivare a strutture più complesse del tipo CVC, gruppi consonantici (CCV) e parole multisillabiche. Con i *processi di sistema*[10]si fa riferimento alla sostituzione di una certa classe di fonemi con un'altra. Tutti gli errori di articolazione frequenti in questo periodo potrebbero derivare dalla forza di precedenti connessioni già stabilite nel sistema cognitivo del bambino. Non appena il bambino impara nuove parole, configurazioni sillabiche e fonemi, gli errori possono coinvolgere anche ciò che è conosciuto bene.

Dai 4 agli 8 anni ha luogo la fase della *stabilizzazione del sistema fonologico*: durante questo periodo si stabilizza la produzione di quei fonemi che precedentemente era ancora variabile e si completa l'inventario fonetico. Inoltre, grazie all'apprendimento della letto-scrittura intorno ai 6 anni, i bambini imparano che le parole possono essere segmentate in unità discrete e che i suoni possono essere rappresentati da simboli grafici e sviluppano la competenza metafonologica, vale a dire la capacità di trattare il linguaggio nel suo aspetto più astratto indipendentemente dall'uso. Inoltre, vengono costantemente migliorate la capacità di organizzare ed eseguire programmi motori associati a proposizioni sempre più complesse. Vari studi (Hawkins, 1984; Kent, 1976) hanno infatti evidenziato che i bambini fino alla pubertà non riescono a coordinare completamente i vari livelli del sistema di comunicazione verbale necessario per produrre parole con le caratteristiche temporali dell'adulto, risultando così più lenti rispetto agli adulti nella coarticolazione nei gruppi consonantici, presentando una certa stereotipicità nella produzione di alcuni particolari segmenti e dimostrando una maggiore variabilità nella ripetizione di una stessa espressione.

4. OBIETTIVI

Abbiamo visto sopra come, sotto il profilo fonetico-fonologico, l'acquisizione di una L2 risulti fortemente condizionata dal sistema della L1 sottostante. Il bambino di famiglia straniera che apprende l'italiano come L2 si ritrova ad affrontare situazioni in cui, per esempio, fonemi L1 non esistono in L2 (con necessità di sopprimerli) e altre in cui allofoni contestuali in L1 sono fonemi in L2, e viceversa (con necessità rispettivamente di distinguerli o non distinguerli). Allo stesso tempo il bambino deve superare altre problematiche a livello cognitivo (categorizzazione dei foni e della fonotassi di L2) e a livello di controllo motorio (acquisizione di abitudini articolatorie più o meno nuove).

Ci siamo quindi posti una serie di interrogativi sull'apprendimento dell'italiano L2 nei bambini di famiglia straniera in età prescolare.

A partire dai dati raccolti somministrando ai bambini i due test messi a punto all'interno del progetto e descritti in Galatà & Zmarich (2011a, 2011b), con il presente studio ci siamo innanzitutto proposti di verificare la validità dei predetti test per la valutazione della compe-

---

nante e/o vocale, metatesi, migrazione, epentesi di consonante (per una trattazione esaustiva si veda Bortolini, 1995).

[10] I più frequenti sono: stopping, affricazione, fricazione, gliding, anteriorizzazione, posteriorizzazione, desonorizzazione/sonorizzazione.

tenza fonetico-fonologica dell'italiano L2 e di verificare le potenzialità del *software* PHON[11] in ambito di L2 al fine di estendere le procedure di codifica ai dati all'intero campione dei bambini reclutati nell'ambito del progetto.

Ci siamo inoltre interrogati, in via preliminare e con i primi dati alla mano, sulle capacità di un piccolo gruppo di bambini rumeni di produrre e discriminare i fonemi consonantici italiani non presenti nella loro lingua madre (nemmeno come allofoni) rispetto ai loro coetanei italiani. In ultima istanza ci siamo chiesti se, e in che misura, le variabili linguistiche ed extralinguistiche possano influire sull'acquisizione di L2: per questo abbiamo analizzato e codificato i questionari riguardanti la biografia linguistica dei bambini compilati dagli stessi genitori.

5. MATERIALI E METODI

5.1. Il campione di indagine

Il campione di indagine è costituito da un gruppo di 10 bambini rumeni pre-scolari, 5 maschi e 5 femmine, con un profilo di sviluppo nella norma[12], di età compresa tra i 61 e i 73 mesi (età media: 68 mesi, ds ±4.2 mesi). Per meglio valutare le loro prestazioni, i loro risultati sono stati successivamente confrontati con quelli di un campione di 10 bambini italiani coetanei (5 maschi e 5 femmine, età media: 65 mesi, ds ±2.5 mesi), estratti a caso dalla fascia superiore (> 61 mesi) di un campione di 122 bambini italiani di età tra 36 e 76 mesi.

Poiché i bambini a cui ci rivolgiamo col nostro progetto provengono da famiglie di stranieri immigrati nel nostro paese (famiglie in cui l'italiano è spesso appreso da entrambi i genitori in età adulta), i bambini intervistati ricadono in quello che Bettoni (2001) definisce bilinguismo *precoce e consecutivo*: nella maggior parte dei casi, infatti, nei primi anni di vita questi crescono parlando esclusivamente la loro lingua madre. Ciò vale sia per i bambini che nascono in Italia sia per quelli che giungono nel nostro paese nella prima infanzia (ad es. a seguito di ricongiungimenti familiari). Per tutti loro, salvo casi specifici, l'acquisizione dell'italiano ha inizio solo con l'ingresso nella scuola dell'infanzia all'età di ca. 3 anni.

5.2 Le consonanti del rumeno

Il rumeno appartiene, come l'italiano, alla famiglia linguistica dell'Indoeuropeo (vedi descrizione fornita da Mioni, 2005: 7). L'inventario fonetico delle consonanti del rumeno consta di 22 fonemi: rispetto all'italiano[13] sono assenti in rumeno i fonemi /ʤ/, /ʎ/ e /ɲ/, come anche la geminazione delle consonanti con funzione distintiva. Sono presenti anche diverse varianti allofoniche che qui non riportiamo (per maggiori dettagli si veda Chitoran, 2001): ciononostante, possiamo ritenere i due sistemi abbastanza simili. Si osservi anche che in italiano /ʤ, ʎ, ɲ/ in posizione intervocalica sono considerate geminate *intrinseche*, da

[11] PHON (Rose & MacWhinney, in press) è un software gratuito *open-source* utile a quanti si occupano di linguaggio dal punto di vista segmentale (trascrizione fonetica) e offre la possibilità di comparare le effettive realizzazioni fonetiche con i target fonologici (download disponibile da http://phon.ling.mun.ca/phontrac/).

[12] Riportato come tale dagli stessi genitori che hanno risposto a un apposito questionario.

[13] Limitatamente alle consonanti, l'inventario fonetico dell'italiano, che rappresenta il nostro punto di partenza per i confronti e le analisi, consta di 23 fonemi consonantici (comprese le approssimanti /w/, /j/; per approfondimenti *cfr*. Bertinetto & Loporcaro, 2005).

non confondere con le geminate *lessicali* a cui più avanti ci riferiremo quando parleremo di consonanti geminate (*cfr*. Schmid, 1999).

	bilab.	lab.dent.	dent.	post.alv.	palat.	vel.	lab.vel.	glott.
occl.	p b		t d			k g		
nasali	m		n		ɲ			
polivibr.			r					
fricative		f v	s z	ʃ ʒ				h
affricate			ʦ ʣ	ʧ ʤ				
appross.					j		w	
lat. appr.			l		ʎ			

Tabella 1. Inventario fonetico del rumeno (Chitoran, 2001). Nelle caselle in grigio vengono riportate per completezza i fonemi dell'italiano assenti in rumeno.

5.3 Le prove somministrate e la raccolta dei dati

I dati su cui ci siamo basati in questa sede sono di due tipi: 1) risposte (uguale/diverso) dei bambini a un test di discriminazione (di tipo AX) di non parole; 2) produzione di alcune non parole prodotte dagli stessi bambini in un compito di ripetizione.

Le due prove, quella di discriminazione e quella di produzione, contengono stimoli (non parole, appunto) di tipo 'CV.CV: le consonanti sono quelle della lingua italiana, mentre per le vocali sono state utilizzate le vocali cardinali /a, i, u/. Per la definizione dei contrasti nelle coppie di non parole del test di discriminazione, si è deciso di contrapporre al fonema target, presente in italiano ma non presente nella L1, uno dei fonemi italiani condivisi con la L1, scelto in base alla somiglianza o prossimità data dal numero di tratti distintivi su base acustica (tratti di Jakobson, 1966[1963]) secondo la matrice ridondante del sistema fonologico italiano redatta da Mioni (1983: 64). Nella coppia di non parole sono stati quindi messi in opposizione quei fonemi italiani non presenti nelle L1[14] con i fonemi invece presenti e che si differenziavano rispetto ai primi per un numero massimo di 3/4 tratti acustici.[15] Come protocollo di somministrazione ci siamo ispirati a Roy & Chiat (2004): nella prova di ripetizione gli stimoli sono stati presentati ai bambini su entrambi i diffusori acustici, chiedendo loro di fare "il gioco del pappagallo" e di ripetere le parole; nella prova di discriminazione la presentazione degli stimoli è stata effettuata in modalità stereofonica con i diffusori acustici camuffati uno da *Grillo Parlante* e uno da *Pinocchio*. Tutti gli stimoli, prodotti da un parlante di sesso femminile, sono stati presentati ai bambini con una intensità normalizzata a 65dB in modalità randomizzata all'interno di ciascuna sessione (una sessione di prova e tre sessioni di test in entrambe le prove sperimentali). La somministrazione, gestita tramite computer attraverso procedure automatizzate e due distinti script di *Praat*, è stata effettuata in uno dei locali messi a disposizione dalla scuola dell'infanzia. Tutte le risposte del bambino, sia quelle del test di discriminazione che quelle del compito di ripetizione, sono state registrate con microfono *AKG Perception 120* collegato a una scheda audio *Edirol UA-101* (a 44.1kHz, 16bit-mono) con monitoraggio in cuffia.[16]

[14] Le prove in oggetto sono state create per più L1: ovvero le comunità rumena, arabo-marocchina, albanese e nigeriana Igbo e Edo.
[15] Per la geminazione è stato utilizzato il tratto prosodico di lunghezza.
[16] Maggiori dettagli disponibili in Galatà & Zmarich (2011a, 2011b).

In questa sede abbiamo selezionato dalla prova di produzione solo quelle non parole (10 in tutto) i cui fonemi in posizione intervocalica sono fonemi dell'italiano non presenti in rumeno: /dadzdza, tadzdzi, naʎʎa, saʎʎu, siɲɲi, tiɲɲa/; per la geminazione, che in italiano ha funzione distintiva, abbiamo selezionato le non parole /bitti, daffi, jalli, tissa/. Allo stesso modo, per la prova di discriminazione abbiamo estrapolato solo le risposte dei bambini riferite alle opposizioni presentate e contenenti uno dei suddetti fonemi target, ovvero le tre coppie (una per sessione) di non parole /sadzdza, saza/, /dzaki, daki/, /tadzdzi, tazi/ per /dz/; le coppie /tina, tiɲɲa/, /nabi, ɲabi/, /sini, siɲɲi/ per /ɲ/; le coppie /nala, naʎʎa/, /ʎaki, laki/, /saju, saʎʎu/ per /ʎ/; per la geminazione le coppie /daffi, dafi/, /tisa, tissa/, /biti, bitti/.

5.2 Codifica e analisi delle risposte dei bambini

Tutte le non parole prodotte sono state segmentate ed etichettate. La segmentazione e l'etichettatura è stata effettuata in *Praat* da un operatore mediante ispezione della forma d'onda e del sonogramma: sia la parola *target* che la parola prodotta dal bambino (*actual*) sono state trascritte ortograficamente e foneticamente (rispettivamente *IPA target* e *IPA actual*). Per l'*IPA actual* è stata effettuata anche una segmentazione ed etichettatura a livello di foni. Nei casi di dubbio o incertezza sull'etichettatura della produzione è stato richiesto a due giudici esterni di ascoltare lo stimolo e di fornire una etichettatura dello stesso: nei casi di ulteriore incertezza si è deciso a maggioranza con le risultanze oggettive dell'ispezione del sonogramma.

Sfruttando le potenzialità di *PHON*, con l'etichettatura effettuata in *Praat* abbiamo generato un database contenente le suddette annotazioni, ovvero la trascrizione ortografica delle non parole, la codifica e la trascrizione in IPA della forma canonica dello stimolo presentato al bambino (*target*), e la forma concretamente pronunciata dal bambino su ripetizione (*actual*): a partire da tali dati è stato possibile calcolare l'inventario fonetico, classificare e conteggiare i processi fonologici di semplificazione.[17]

Con riferimento alla prova di discriminazione, in questa sede sono state invece conteggiate le risposte fornite dai bambini agli stimoli selezionati per il presente lavoro e raggruppate per contrasto (foni /dz, ʎ, ɲ/ e geminazione).

5.3 La biografia linguistica dei bambini

Per interpretare le prestazioni dei bambini stranieri e valutare le loro capacità fonetico/fonologiche alla luce di eventuali variabili significative per il raggiungimento della fluenza nell'italiano L2 (come variabili di tipo linguistico ed extralinguistico), ai genitori dei bambini è stato chiesto di compilare un questionario. Il questionario includeva una biografia linguistica del bambino per raccogliere una serie di informazioni inerenti: a) lo stato di salute del bambino (in rif. al parto, alla nascita e a eventuali patologie post-parto); b) le tappe evolutive e di sviluppo linguistico del bambino; c) il nucleo familiare e le persone che il bambino frequenta abitualmente; d) informazioni anagrafiche e socio-economico-demografiche sui genitori del bambino. Rispetto ai bambini italiani, per i bambini stranieri è stata prevista un'integrazione al questionario riguardante aspetti più specifici come: a) grado e tipo di stimolazione linguistica; b) pratiche narrative; c) competenze linguistiche nella lingua prevalente; d) grado di acculturazione e sfera socio-culturale dei genitori stra-

[17] Operazioni possibili grazie a funzioni di ricerca predefinite o ulteriori funzioni di ricerca costruibili attraverso PhonEx, un linguaggio dalla sintassi semplice che opera su un insieme di oggetti simili ai tratti distintivi delle teorie fonologiche.

nieri; e) aspettative sulla scuola. È stata quindi richiesta autorizzazione e consenso scritto a partecipare alla sperimentazione, oltre alla sottoscrizione dell'informativa sulla *privacy* (legge 196/2003).

6. RISULTATI E DISCUSSIONE DEI DATI

In Figura 1 riportiamo la *performance* dei bambini rumeni e dei bambini italiani. Se ci si limita a osservare la percentuale delle produzioni corrette dei *target* "problematici per i rumeni" di tutti i bambini, sembrerebbe che le nostre attese siano confermate solo per le non-parole /naʎʎa, saʎʎu, siɲɲi, tiɲɲa, daffi, jalli/, escludendo l'ipotesi che i bambini rumeni abbiano difficoltà per gli stimoli contenenti il fonema /ʤ/ e mettendo in dubbio l'esistenza o meno di una certa difficoltà per le geminate (Figura 1).

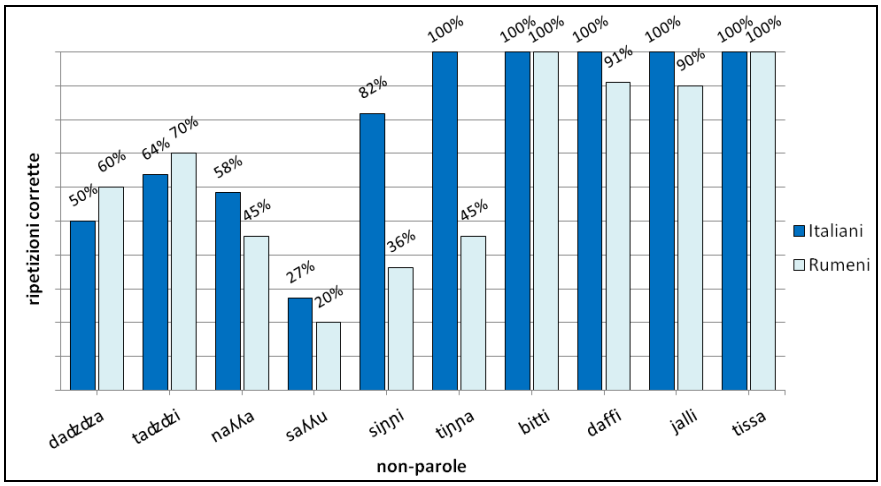


Figura 1: Produzioni corrette delle 10 non-parole esaminate.

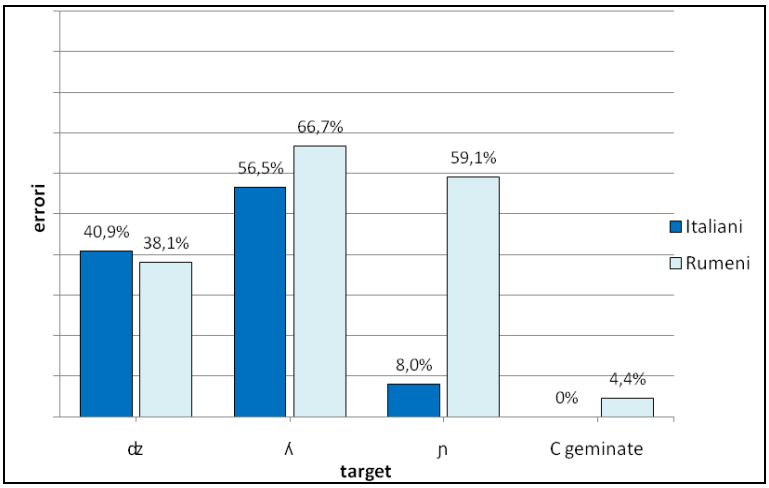


Figura 2: Produzioni errate dei fonemi assenti in rumeno e gestione del tratto di lunghezza.

Da una più attenta analisi, che mette a confronto in Fig. 2 gli errori dei due gruppi di bambini nella produzione dei fonemi non presenti nella lingua rumena e nella gestione del tratto di lunghezza, emergono dei risultati interessanti (sulle geminate ritorneremo più avanti). Viene confermata la maggiore difficoltà dei bambini di L1 rumena nella realizzazione dei fonemi /ʎ/ e /ɲ/; in particolare emerge una spiccata difficoltà proprio per quest'ultimo fonema. Si evince anche come non sussista una particolare differenza tra italiani e rumeni nella realizzazione di /ʣ/.

Il ruolo delle variabili linguistiche è invece messo in luce dal tipo di errori commessi dai bambini italiani e rumeni nel produrre le non-parole contenenti i fonemi assenti in rumeno (analisi delle sostituzioni in Fig. 1, Fig. 2 e Fig. 3) che evidenziano come la sensibilità alle discriminazioni fonetiche sia guidata dal sistema della L1 di tali bambini (Mattock, Amitay, & Moore, 2010). Nei tentativi di realizzare l'affricata alveolare sonora /ʣ/ (vedi Fig. 3), i due gruppi di bambini esibiscono processi diversi: nei rumeni si rileva una preferenza per il processo di posteriorizzazione (/ʣ/ realizzata come /ʤ/), mentre gli italiani commettono più spesso una fricazione (/ʣ/ realizzata come /z/). È ipotizzabile che questa preferenza per /ʤ/ dei bambini stranieri sia dovuta all'alta frequenza di questo fonema nella loro lingua[18], mentre il tipo di errore commesso dagli italiani è sicuramente riconducibile al fenomeno dell'assibilazione: ciò potrebbe essere riconducibile al fatto che i bambini italiani del gruppo di controllo sono tutti di origine veneta, regione nella quale sovente le consonanti affricate vengono realizzate come sibilanti con esiti di /ʣ/ in /z/ o /s/ (Telmon, 1993).

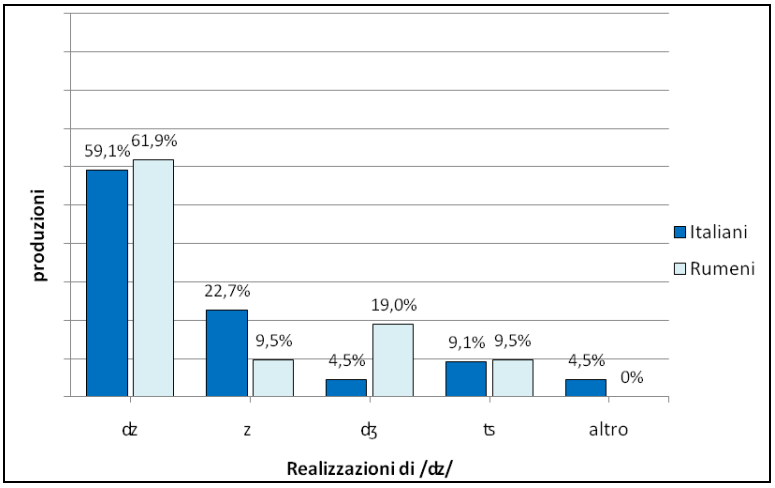


Figura 3: Realizzazioni del fonema /ʣ/.

Analogamente, nelle produzioni scorrette della nasale palatale /ɲ/ (Figura 4), i bambini del gruppo di controllo sbagliano esclusivamente riconducendo il fonema *target* ad altri fonemi nasali, ma di diverso luogo di articolazione: errore anch'esso riconducibile all'origine veneta di questi bambini, dove è frequente la difonizzazione delle consonanti nasali e laterali palatali, e viceversa (Telmon, 1993). I bambini rumeni sbagliano o per difonizzazione (31,8%) o sostituendo la nasale palatale con una nasale più una semiconsonante (22,7%).

[18] Si tratta solo di un'ipotesi che merita, tuttavia, un opportuno approfondimento.

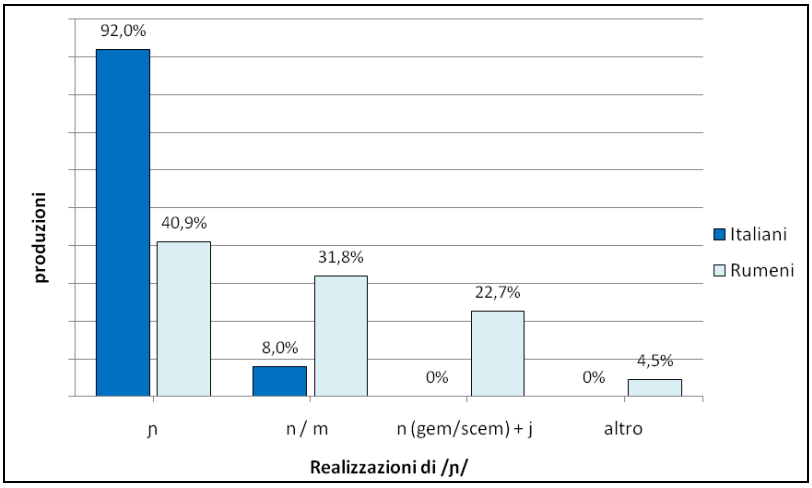


Figura 4: Realizzazione del fonema /ɲ/.

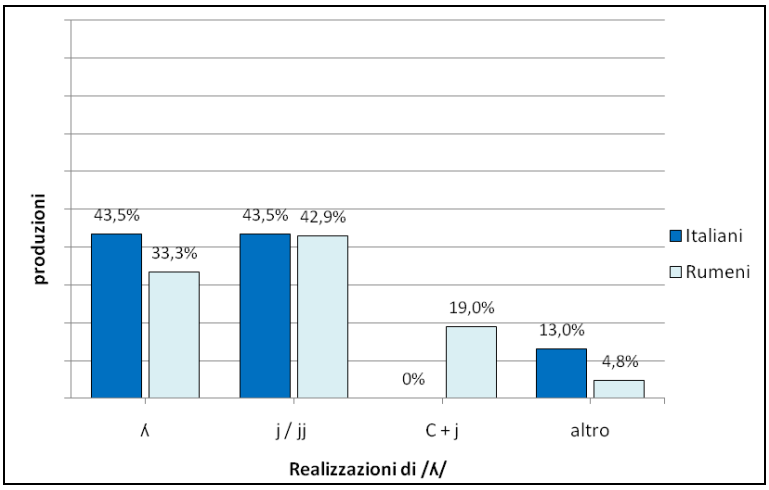


Figura 5: Realizzazione del fonema /ʎ/.

In Figura 5 riportiamo gli errori commessi dai due gruppi di bambini nella pronuncia della laterale palatale /ʎ/: nei bambini rumeni si registra una percentuale di errore piuttosto alta con una maggiore variabilità nel tipo di processo di semplificazione; tale dato si dimostra alto anche nei bambini italiani e non è per nulla anomalo se si considera che si tratta di uno degli ultimi foni a comparire nell'inventario fonetico dei bambini italiani[19] e che, a maggior ragione, lo si possa considerare un suono "difficile" per i bambini rumeni.

In Figura 6 abbiamo, infine, comparato la capacità dei bambini rumeni di produrre correttamente i fonemi esaminati e le consonanti geminate con la loro capacità di categorizzazione che sappiamo essere un prerequisito per la corretta produzione. Da questo confronto

[19] Bortolini (1995) riporta il fono come attestato nell'inventario fonetico in meno del 50% dei bambini fino ai 48 mesi da lei analizzati.

emerge che le palatali /ɲ/ e /ʎ/ vengono percepite meglio di quanto siano prodotte, mentre i risultati relativi alla percezione e produzione dell'affricata alveolare sonora sembrano dovuti a risposte casuali (il punteggio si aggira attorno al 60%).

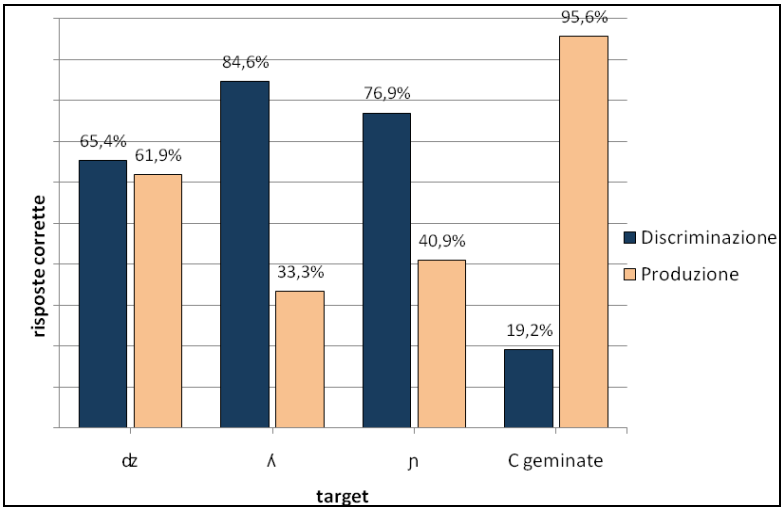


Figura 6: Media delle risposte corrette fornite nelle due prove dai bambini rumeni.

Merita, tuttavia, particolare attenzione il caso delle consonanti geminate, dove a fronte di una scarsa capacità dei bambini nel percepirle emergerebbe una buona capacità nel produrle. Sebbene ciò possa sembrare un risultato incompatibile (nella prova di produzione è normale attendersi un punteggio minore o uguale a quello di percezione, non il contrario), in questo caso troviamo la dimostrazione al fatto che i bambini rumeni sono insensibili al tratto di lunghezza delle consonanti. Infatti, il buon punteggio ottenuto nella prova di produzione è frutto del fatto che questi bambini producono tutte le consonanti con una durata compresa tra quella con cui noi pronunciamo le consonanti scempie e le geminate, e tale valore intermedio viene da noi spesso categorizzato come più simile alla durata delle consonanti geminate piuttosto che quella delle scempie: l'attribuzione di un giudizio positivo da parte di chi valuta questo aspetto è facilmente giustificabile dal fatto che variando il tratto di lunghezza di un fono, il fonema non cambia. Questo dato è in linea con quanto evidenziato in altri studi (vedi Hawkins, 1984; Kent, 1976) sull'incapacità dei bambini fino alla pubertà di coordinare completamente i vari livelli del sistema di comunicazione verbale necessario per produrre parole con le caratteristiche temporali dell'adulto, manifestando stereotipicità nella produzione di alcuni particolari segmenti e dimostrando una maggiore variabilità nella ripetizione di una stessa parola o espressione.

Infine, correlando le percentuali di errore commesse da ciascun bambino nella prova di produzione con il rispettivo tempo di esposizione all'italiano (variabile extralinguistica su cui ci si è maggiormente concentrati, in quanto attesa come la più importante), non abbiamo riscontrato una relazione valida per tutti i bambini: ciò potrebbe suggerire che la loro prestazione sia stata condizionata anche da altri fattori, come per esempio la frequenza d'uso di L2.

7. CONCLUSIONI

In linea generale esiste un'effettiva difficoltà per i bambini di L1 rumena a realizzare quegli aspetti non condivisi con l'italiano dalla loro lingua madre, per cui è possibile affermare che le loro competenze fonetico-fonologiche in italiano sono influenzate, a differenza dei coetanei italiani, da una diversa sensibilità alle discriminazioni fonetiche dettate dal sistema della loro L1 (Mattock, Amitay, & Moore, 2010).

I primi risultati che qui abbiamo esposto, sebbene meritino maggiori approfondimenti statistici, evidenziano e sostengono l'ipotesi secondo cui i bambini non siano in grado di coordinare a pieno i vari livelli del sistema di comunicazione verbale necessario per produrre parole con le caratteristiche temporali dell'adulto, soprattutto per quanto riguarda le geminate (*cfr*. Hawkins, 1984; Kent, 1976).

Riguardo alla bontà dei test messi a punto dal primo e dall'ultimo autore di questo contributo si può certamente affermare come questi risultino utili per valutare la competenza fonetico-fonologica in italiano dei bambini stranieri; essi infatti aiuterebbero a distinguere i casi di difficoltà legati solo all'acquisizione di L2 da possibili casi di reale ritardo di linguaggio, attraverso il confronto con i coetanei italiani. Permangono, tuttavia, una serie di limitazioni insite nei due test: non vengono per esempio considerati i nessi consonantici e, sebbene si sia tentato di coprire l'inventario fonetico dell'italiano, non è stato possibile testare ogni fono in tutti i contesti per evidenti limitazioni legate ai tempi e alle difficoltà insite nella somministrazione di un compito sperimentale a bambini piccoli.

Per concludere, poiché gran parte delle analisi svolte sono state condotte con l'ausilio delle funzioni del *software* PHON, si conferma la validità di questo strumento oltre che nel campo dello studio dello sviluppo fonetico-fonologico dei bambini, anche nel campo dello studio dell'acquisizione di una L2 in cui, come abbiamo dimostrato, risulta spesso utile il confronto tra *target* ed effettiva produzione del soggetto (produzione che può facilmente differire dal *target* per interferenza della lingua madre).

RINGRAZIAMENTI

Si ringraziano in particolare un revisore anonimo e la dr.ssa Cinzia Avesani per i preziosi suggerimenti che ci hanno permesso di migliorare il presente lavoro.

BIBLIOGRAFIA

Abdelilah-Bauer, B. (2008). *Il bambino bilingue. Crescere parlando più di una lingua*. Milano: Cortina Raffaello.

Anderson, R. T. (2004). Phonological acquisition in preschoolers learning a second language via immersion: a longitudinal study. *Clinical Linguistics & Phonetics*, *18*(3), 183-210.

Bates, E. (1995). Conclusioni. In M. C. Caselli & P. Casadio (Eds.), *Il primo vocabolario del bambino. Guida all'uso del questionario MacArthur per la valutazione della comunicazione e del linguaggio nei primi anni di vita* (pp. 93-98). Milano: Franco Angeli.

Bertinetto, P. M., & Loporcaro, M. (2005). The sound pattern of Standard Italian, as compared with the varieties spoken in Florence, Milan and Rome. *J. Int. Phon. Ass.*, *35*(02), 131-151.

Bettoni, C. (2001). *Imparare un'altra lingua*. Bari: Laterza.

Birdsong, D. (1999). *Second language acquisition and the critical period hypothesis*. (David Birdsong, Ed.). Mahwah (NJ), London: Lawrence Erlbaum Associates.

Bley-Vroman, R. (1989). What is the Logical Problem of Foreign Language Learning? In S Gass & J. Schachter (Eds.), *Linguistic Perspectives on Second language Acquisition* (pp. 41-68). Cambridge: Cambridge University Press.

Boersma, P., & Weenink, D. (2010). Praat: doing phonetics by computer [Computer program]. *version*, 5.1.32.

Bortolini, U. (1995). *PFLI Prove per la valutazione fonologica del linguaggio infantile*. *Edit Master, Venezia*.

Bosch, L., Costa, A., & Sebastián-Gallés, N. (2000). First and second language vowel perception in early bilinguals. *European Journal of Cognitive Psychology*, *12*(2), 189-221. Psychology Press.

Chitoran, I. (2001). *The Phonology of Romanian: A Constraint-Based*. Berlin, New York: Mouton de Gruyter.

Galatà, V., & Zmarich, C. (2011a). Le non-parole in uno studio sulla discriminazione e sulla produzione dei suoni consonantici dell'italiano da parte di bambini pre-scolari. In B. Gili Fivela, A. Stella, L. Garrapa, & M. Grimaldi (Eds.), *Atti del VII Convegno Nazionale dell'Associazione Italiana di Scienze della Voce* (pp. 118-129). Roma: Bulzoni Editore.

Galatà, V., & Zmarich, C. (2011b). Una proposta per valutare l'influenza fonetico-fonologica della lingua di origine dei bambini figli di immigrati sull'acquisizione dell'italiano. In G. C. Bruno, I. Caruso, M. Sanna, & I. Vellecco (Eds.), *Percorsi migranti* (pp. 301-317). Milano: McGraw-Hill.

Gilhool, A., Burrows, L., & Goldstein, B. (2009). English Phonological Skills of English Language Learners. *Poster session presented at the annual American Speech Language Hearing Association Convention*. New Orleans, LA.

Hawkins, S. (1984). On the development of motor control in speech: Evidence from studies of temporal coordination. In N. Lass (Ed.), *Speech and Language: Advances in Basic Research and Practice* (Vol. 11, pp. 317-374). New York: Academic Press.

Ingram, D. (1976). *Phonological disability in children*. Edward Arnold.

Jakobson, R. (1963). *Essais de linguistique générale,* Paris, Minuit (trad. it. *Saggi di linguistica generale,* Milano, Feltrinelli, 1966).

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. (H. D. Brown & S. Gonzo, Eds.)*Cognitive Psychology*, *21*(1), 60-99. Elsevier.

Kent, R. D. (1976). Anatomical and neuromuscular maturation of the speech mechanism: evidence from acoustic studies. *Journal Of Speech And Hearing Research*, *19*(3), 421-447.

Lalonde, C. E., & Werker, J. F. (1995). Cognitive influences on cross-language speech perception in infancy. *Infant Behavior and Development*, *18*(4), 459–475. Elsevier.

Mattock, K., Amitay, S., & Moore, D. R. (2010). Auditory development and learning. In C. Plack (Ed.), *Oxford Handbook of Auditory Science* (pp. 297-324). Oxford University Press.

Mioni, A. (1983). Fonologia. In L. Croatto (Ed.), *Trattato di Foniatria e Logopedia, Vol. 2, Aspetti linguistici della comunicazione* (pp. 51-87). Padova: La Garangola.

Mioni, A. (2005). *Immigrati e comunicazione interetnica in Italia*. Università di Padova.

Pallier, C., Dehaene, S., Poline, J.-B., LeBihan, D., Argenti, A.-M., Dupoux, E., & Mehler, J. (2003). Brain imaging of language plasticity in adopted adults: can a second language replace the first? *Cerebral cortex*, *13*(2), 155-61.

Piske, T., Mackay, I. R. A., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics*, *29*(2), 191-215. Elsevier.

Polka, L. (1992). Characterizing the influence of native language experience on adult speech perception. *Attention, Perception & Psychophysics*, *52*.

Rose, Y. & MacWhinney, B. (to appear). The PhonBank Initiative. In Jacques Durand, Ulrike Gut & Gjert Kristoffersen (eds.), *Handbook of Corpus Phonology*. Oxford: Oxford University Press.

Roy, P., & Chiat, S. (2004). A prosodically controlled word and nonword repetition task for 2- to 4-year-olds: evidence from typically developing children. *JSLHR*, *47*(1), 223-234.

Schmid, S. (1999). *Fonetica e fonologia dell'italiano*. Paravia Scriptorium.

Sebastián-Gallés, N., & Bosch, L. (2005). Phonology and bilingualism. In J. F. Kroll & A. M. B. Groot (Eds.), *Handbook of bilingualism: Psycholinguistic Approaches* (pp. 68-87). Oxford: Oxford University Press.

Schwartz, R. G., & Leonard, L. B. (1982). Do children pick and choose? an examination of phonological selection and avoidance in early lexical acquisition. *Journal of Child Language*, *9*, 319-336.

Stoel-Gammon, C., & Dunn, C. (1985). Normal and disordered phonology in children. Baltimore: University Park Press.

Tees, R. C. (2001). Critical and sensitive periods. In P. Winn (Ed.), *Dictionary of biological psychology* (p. 195). London: Routledge Press.

Telmon, T. (1993). Varietà regionali. In A. A. Sobrero (Ed.), *Introduzione all'italiano contemporaneo. La variazione e gli usi* (pp. 93-149). Roma-Bari: Editori Laterza.

Valentini, A. (2005). Lingue e interlingue dell'immigrazione in Italia. *Linguistica e Filologia*, *21*, 185-208.

Vihman, M. M. (1996). Phonological Development. Oxford: Blackwell Publishers.

Werker, J. F., & Byers-Heinlein, K. (2008). Bilingualism in infancy: first steps in perception and comprehension. *Trends in cognitive sciences*, *12*(4), 144-51.

Werker, J. F., & Tees, R. C. (2005). Speech perception as a window for understanding plasticity and commitment in language systems of the brain. *Developmental psychobiology*, *46*(3), 233-51.

Zmarich, C. (2010). Lo sviluppo fonetico/fonologico da 0 a 3 anni. in S. Bonifacio, L. Hsvastja Stefani (Eds.), *L'intervento precoce nel ritardo di Linguaggio. Il modello INTERACT per il bambino parlatore tardivo* (pp. 17-39), Franco Angeli, Milano.

VERSO UN "TEST FONETICO PER LA PRIMA INFANZIA"

*Zmarich C., °Fava I., °Del Monego G., #Bonifacio S.
*CNR-ISTC, Padova, °Università di Padova, #IRCCS "Burlo Garofolo", Trieste
claudio.zmarich@pd.istc.cnr.it, logopedia@burlo.trieste.it

1. SOMMARIO

Il presente contributo si propone di descrivere le capacità fonetico-fonologiche di un gruppo di 30 bambini veneti e marchigiani distribuiti a gruppi di 10 in tre fasce d'età: 18-23 mesi; 24-29 mesi; 30-36 mesi. Per tale scopo è stata utilizzata la versione aggiornata al mese di aprile 2011 del Test Fonetico per la Prima Infanzia (TFPI, cfr. Zmarich *et alii*, 2010). La conoscenza sullo sviluppo fonetico-fonologico in queste età è tuttora carente, e la mancanza è critica soprattutto a livello di pratica clinica, per diagnosticare un possibile ritardo o disordine fonetico-fonologico. Inoltre, benché esista almeno un test, le Prove per la Valutazione Fonologica del Linguaggio Infantile (o PFLI, Bortolini, 1995), che si propone come scientificamente fondato e adatto per bambini dai 24 mesi in poi, esso risulta fortemente deficitario per i motivi che illustreremo. Il TFPI, attualmente ancora in fase di perfezionamento, ma non lontano da una forma definitiva, è diviso in tre subtest in base all'età dei soggetti ed è costruito secondo tre criteri: 1. Fonetico: i fonemi consonantici della lingua italiana devono essere attestati in almeno due parole diverse per ciascuna posizione lessicale; 2. Semantico/Frequenziale: le parole sono sostantivi concreti relativamente frequenti nel lessico infantile; 3. Gradualità nella Complessità Fonetica: le parole utilizzate mostrano un aumento della complessità dalla prima alla terza fascia, per numero e tipi di sillabe.

In questo campione, i foni attestati in oltre l'80% dei bambini della prima fascia d'età sono soprattutto occlusivi sordi e nasali. Nella seconda fascia vengono consolidate tutte le occlusive e compaiono le fricative. Nell'ultima fascia l'inventario fonetico è piuttosto completo sia in posizione iniziale che mediana, avendo consolidato la maggior parte delle consonanti che nelle precedenti età erano tra il 50% e l'80%. I dati relativi ai processi di semplificazione vedono primeggiare la riduzione dei gruppi consonantici ed evidenziano una riduzione progressiva, in base all'età cronologica, di quasi tutti i processi.

Per quanto riguarda il test, possiamo dire che presenta vari punti di forza, tra i quali: 1) la facilità e la velocità di somministrazione; 2) la semplicità e la familiarità del materiale visivo utilizzato nelle diverse fasce d'età; 3) la validità di costrutto, perché la sua applicazione produce risultati che concordano con quelli di strumenti simili (PFLI) e 4) la buona correlazione con altri aspetti importanti dello sviluppo linguistico (PVB; TPL, Axia, 1995).

2. INTRODUZIONE

2.1. L'importanza di un test fonetico

Il corretto sviluppo del linguaggio può rappresentare una delle preoccupazioni maggiori per un genitore. Si ritiene che la diffusione dei disturbi specifici del linguaggio (disturbi del linguaggio senza causa organica apparente) si aggiri attorno al 15%-20% in età prescolare (10%-15% disturbi espressivi, 5% disturbi espressivo-ricettivi (*DSM IV*, APA, 2001). Una componente importante riguarda i cosiddetti "disturbi della fonazione" (leggi "fonologici", 2-3% a 6/7 anni). Questi disturbi presentano dei problemi di diagnosi differenziale sia ri-

spetto alla normalità, date le differenze individuali nell'espressione linguistica, sia per la difficoltà di discriminare le diverse componenti nelle varie tappe evolutive, biologiche e psico-relazionali, necessarie allo sviluppo della produzione linguistica. È importante, quindi, in ambito clinico, avere a disposizione degli strumenti valutativi in grado di fornire utili indicazioni sulle capacità linguistiche del bambino, specie se molto piccolo (18-36 mesi).

L'aspetto più indagato nell'ambito delle ricerche sullo sviluppo fonetico-fonologico è quello segmentale, per motivi pratici e teorici. Il motivo pratico è dovuto alle difficoltà nell'utilizzo di strumentazioni e procedimenti non adatti ai bambini che, soprattutto quando sono piccoli, sono scarsamente collaborativi, e alla necessità di valutare la normalità dello sviluppo con un metodo semplice e veloce, come quello basato sulla trascrizione fonetica del percetto uditivo. Il motivo teorico è la forte dipendenza della fonetica e ancor più della fonologia dal concetto di segmento/fono/fonema, individuabile attraverso la trascrizione fonetica (Zmarich, 2010). Nella pratica clinica è fondamentale raccogliere la produzione segmentale di un soggetto per confrontarla con i dati normativi per la stessa fascia d'età, al fine di stabilire la normalità o meno di quel percorso individuale. Anche se, da un punto di vista teorico, raccogliere il parlato spontaneo infantile prodotto senza alcuna sollecitazione può risultare attraente per la sua ecologicità, è opportuno sottolineare che tale procedura è soggetta a dei rischi. In particolare, come evidenziato da Shriberg & Kwiatkowski (1985) e Eisenberg & Hitchcock (2010), i bambini potrebbero comunicare in modo inintelligibile, essere riluttanti a comunicare all'interno del tempo a disposizione, produrre un parlato che differisce strutturalmente dai dati normativi (per es. si sa che i bambini evitano di produrre le parole per loro "difficili") o esprimersi in modo innaturale (per es. in tono canzonatorio o cantilenante), senza contare la variabilità indotta dall'esaminatore e dalla tipologia dell'eventuale materiale utilizzato, che potrebbe rendere inconfrontabili i dati raccolti in situazioni e\o momenti e\o da persone diverse (Zmarich *et alii*, 2010).

Molti dei test esistenti nella letteratura internazionale fanno riferimento a tre modalità principali di compiti utili alla valutazione delle capacità fonetico-fonologiche dei bambini, specialmente quelli più piccoli, ovvero la denominazione di figure, la ripetizione di parole e la ripetizione di non-parole. Vance *et alii* (2005) fanno riferimento a un modello dell'elaborazione linguistica in età evolutiva allo scopo di descrivere i processi e gli stadi di elaborazione di queste tipologie di compiti: per quelli di ripetizione si va dall'elaborazione uditiva dell'*input* acustico, che implica la percezione, la discriminazione e il riconoscimento del segnale linguistico, alla ritenzione nella memoria a breve termine delle informazioni relative ai singoli stimoli. La ripetizione di non-parole richiede poi la costruzione ex-novo di un programma motorio mentre la ripetizione di parole può richiamare la rappresentazione fonologica e il programma motorio collegato. Questi compiti vengono usati soprattutto per testare il buon funzionamento della memoria fonologica, per la verifica del buon funzionamento della programmazione motoria (per le non-parole), e per la verifica della correttezza dell'informazione fonologica (per le parole), perché un bambino che non discrimina bene non riproduce bene. Pur essendo relativamente facili e veloci da somministrare, il rischio che ambedue comportano è quello di sovrastimare le capacità fonetico-fonologiche del bambino nel contesto ecologico quotidiano. Per riprodurre il contesto ecologico di produzione linguistica, che va dalla rappresentazione semantica alla produzione articolata, si ricorre al compito di denominazione di figure. In letteratura questo tipo di test viene designato anche con la qualifica "articolatorio", ma siccome in genere le produzioni verbali

vengono codificate in trascrizione fonetica "larga", è più corretto usare il termine "fonetico". Un test fonetico è il tipico prodotto della cosiddetta analisi fonetica indipendente (dal modello adulto) e fornisce un indice delle capacità fonetico-fonologiche del bambino ad una certa età, individuando i foni con cui costruisce le parole, e confrontandoli con i foni prodotti da un campione rappresentativo di coetanei, così da poter valutare la produzione del soggetto come appropriata, in ritardo o deviante. Inoltre, attraverso l'analisi fonologica (sempre di tipo indipendente) del sistema di contrasti in termini di classi fonologiche naturali tra foni presenti in una stessa posizione di parola, il test permette di ricostruire l'emergere di una organizzazione fonologica (Stokes *et alii*, 2005). Infine, con lo stesso test si può condurre un'analisi cosiddetta relazionale (perché mette in relazione la produzione del bambino con quella adulta), studiando gli errori di semplificazione degli stimoli che il bambino commette, che essendo simili in tutti i bambini suggeriscono di essere in presenza di un fenomeno sistematico e "naturale". Il test fonetico viene allora chiamato fonologico e informa sui *pattern* di errore che i bambini commettono, in termini quali-quantitativi, attraverso l'analisi dei cosiddetti "processi fonologici". Senza qui prendere posizione sulla loro natura (cfr. Zmarich, 2010), esistono numerose categorie di processi, che fondamentalmente possono essere ricondotti ai processi di struttura, che semplificano il contesto combinatorio dei foni, e ai processi di sistema, che semplificano il sistema fonologico riducendo il numero dei fonemi in contrasto su una data posizione (Bortolini, 1995). Generalmente il test fonetico indaga la produzione consonantica, e non quella vocalica, sulla base del fatto che il sistema vocalico risulta tipicamente già acquisito entro la fine del terzo anno di vita e la maggior parte dei test non scende sotto i 3 anni (Eisenberg & Hitchcock, 2010). Per l'italiano non ci sono studi sistematici, ma, considerando che il sistema vocalico italiano è più semplice di quello inglese, e che i contrasti morfologici sono spesso affidati a variazioni vocaliche, si ritiene che venga acquisito più precocemente (cfr. Giulivi et alii, 2010). Per Eisenberg & Hitchcock (2010), un buon test fonetico deve inoltre rispondere ai seguenti requisiti: testare tutti i fonemi in tutte le posizioni in cui occorrono, attestare la presenza di un fonema solo dopo che è stato prodotto in almeno due parole diverse, utilizzare parole foneticamente controllate (per accento, lunghezza, struttura sillabica), e utilizzare vocali diverse per ogni consonante (per evitare che il bambino produca non tanto la consonante quanto olisticamente l'intera sillaba).

Attualmente in Italia, diversamente dal mondo anglosassone, non sono disponibili in commercio strumenti scientificamente solidi in grado di fornire una valutazione delle capacità fonetiche e fonologiche dei bambini piccoli, a partire dai 18-24 mesi, cioè l'età in cui generalmente inizia a comparire e a consolidarsi il cosiddetto Primo Vocabolario, fatta eccezione, forse, per le "Prove per la Valutazione Fonologica del Linguaggio Infantile" (PFLI) di Bortolini (1995). Questo test, rivolto a bambini dai 2 ai 5 anni con disordine fonologico, secondo l'autrice può essere utilizzato anche con bambini "con tutti i tipi di linguaggio infantile". Tale test utilizza delle vignette colorate che rappresentano persone, oggetti ed azioni, che il bambino deve descrivere. Esso presenta molti punti deboli. Alcune vignette possiedono uno scarso potenziale di elicitazione, alcuni fonemi sono sovrastimolati, mentre altri non lo sono affatto o troppo poco, l'aspetto psicometrico è inconsistente (la prestazione del bambino non riceve punteggio, il campione di riferimento è dubbio (non viene spiegato come è stato costruito). Infine, ci sono tutti gli svantaggi relativi alla sollecitazione di linguaggio connesso, quali ad esempio che a) la produzione è poco controllata; b)

tutto il materiale deve essere trascritto in IPA (*International Phonetic Alphabet*) implicando un grande dispendio temporale, poco compatibile con i tempi clinici; c) il materiale iconografico utilizzato è poco adatto a bambini piccoli; d) la riuscita del test dipende troppo dalla loquacità del bambino (*talkativeness*) ed infine, e) l'operatore potrebbe riscontrare delle difficoltà nell'individuazione dei target prodotti da un bambino poco intelligibile (poiché ogni vignetta rappresenta diversi target, cfr. Zmarich *et alii*, 2010).

L'obiettivo del presente lavoro consiste nel presentare dati preliminari sulla validità e l'affidabilità (cfr. Pedrabissi e Santinello, 1997) della versione aggiornata al mese di aprile 2011 del Test Fonetico per la Prima Infanzia (TFPI, cfr. Zmarich et al. 2010), e di fornire la descrizione delle capacità fonetico-fonologiche di un gruppo di 30 bambini veneti e marchigiani distribuiti a gruppi di 10 in tre fasce d'età: 18-23 mesi; 24-29 mesi; 30-36 mesi.

2.2.Lo sviluppo fonetico-fonologico dei bambini italiani

La tab. 1, tratta da Zmarich & Bonifacio (2005), illustra gli inventari fonetici (IF) di un gruppo rappresentativo di 13 bambini italiani (di area veneto-giuliana), analizzati longitudinalmente dai 18 ai 27 mesi e confrontati con gli IF calcolati sulle parole del PVB.

18 IN	**p* b t k m**
18 IV	**p t k m**
21 IN	**p* b t* k m n**
21 IV	**p* b t* d k* m n* l***
24 IN	**p* b* t* d k m* n f l**
24 IV	**p* b t* d* k* m* n* f v s l***
27 IN	**p* b* t* d* k* g m n* f v s l* kw**
27 IV	**p* b t* d* k* g m* n* v s l***
targ IN	**p* b* t d* k* g m n f s l kw**
targ IV	**p* b* t* d* k* g m n* f v s l* r st**

Tabella 1: Inventari Fonetici in posizione iniziale (IN) e interna (IV) di parola a 18, 21, 24 e 27 mesi di età e nel target adulto (Targ) per le consonanti attestate in oltre il 50% dei 13 bambini di Zmarich e Bonifacio (2005) (*in oltre il 90%)

Dai 18 ai 27 mesi il sistema fonetico di questi bambini cresce sistematicamente. A 18 mesi sono presenti solamente le occlusive, prevalentemente sorde, e una nasale, articolate anteriormente e inserite in sillabe del tipo CV. Dai 21 mesi, invece, l'inventario fonetico risulta più completo in posizione mediana e si possono osservare i primi influssi della lingua nativa, ovvero l'italiano (cfr. anche Zmarich, 2008). In particolare, si afferma il contrasto di sonorità, compaiono l'approssimante laterale e l'affricata palatoalveolare sorda che riflette la capacità di prolungare un fono o una sua fase. A 24 mesi si può osservare il rafforzamento di tutti i foni occlusivi e la comparsa delle fricative, mentre a 27 mesi, aumentano i tipi sillabici complessi, del tipo CVC e CCV. Per l'ultima fascia, Zanobini *et alii* (2012) forniscono gli IF di un gruppo di 30 bambini dai 36 ai 42 mesi, che aggiungono [v] e [ ] in posizione iniziale e [z], [r], [j] e [w] in posizione intervocalica.

3. MATERIALI E METODI

3.1. Soggetti

Il campione della ricerca è costituito da 30 bambini, di cui 15 maschi e 15 femmine, di età compresa tra i 18 e i 36 mesi con sviluppo linguistico normale (superiori per n. parole al 10° centile della popolazione nel PVB, vedi paragrafo 3.2.3), le cui famiglie hanno firmato il consenso alla partecipazione allo studio. 19 bambini sono stati selezionati da un asilo nido di Selvazzano Dentro (PD) e 11 bambini da un asilo nido di Monte San Vito (AN). I soggetti sono equamente suddivisi nelle tre fasce d’età previste dal test: 10 bambini (4 maschi e 6 femmine) tra i 18 e i 23 mesi (media: 20,06 mesi), 10 bambini (5 m. e 5 f.) tra i 24 e i 29 mesi (media: 26,57 mesi) e 10 bambini (6 m. e 4 f.) tra i 30 e i 36 mesi (media: 31,11 mesi). Gli altri criteri di inclusione per la partecipazione allo studio sono stati l’assenza di disturbi linguistici, l’assenza di problemi di udito accertati o supposti, meno di 4 episodi di otite all’anno, l’italiano come lingua madre. Tali informazioni sono state ricavate da un questionario relativo alla salute del bambino compilato dai genitori.

3.2. Procedura sperimentale

Alcuni giorni prima della registrazione, i genitori dei bambini potenzialmente reclutabili sono stati contattati dagli autori ed informati sulle finalità del progetto e sul suo svolgimento. Ai genitori che hanno deciso di partecipare è stato fatto firmare il modulo per il consenso informato e per la *privacy*, e consegnato il questionario PVB. Dopo il ritiro e l’analisi del PVB, i bambini selezionati sono stati testati all’asilo nido in due giornate differenti: nella prima è stato sottoposto il TFPI (vedi paragrafo 3.2.1) e nella seconda il TPL (Axia, 1995, vedi paragrafo 3.2.2). I test sono stati somministrati in una stanza accogliente del nido, nota al bambino e priva di rumori e oggetti che potevano essere fonte di distrazione. E' stato usato un registratore digitale hi-fi (Edirol R-09, con microfono ad alta sensibilità) per la registrazione dell’intera produzione, collocato il più vicino possibile al bambino. Con i bambini della prima fascia d’età (dai 18 ai 23 mesi) sono stati utilizzati gli oggetti, che l’esaminatore estraeva uno alla volta da un sacchetto chiedendo al bambino “Che cos’è?” o “Come si chiama questo?”. Con i bambini delle altre due fasce d’età, invece, sono state utilizzate le immagini colorate, contornate da una sottile cornice bianca, su cartoncini di 21x15 cm. L’ordine di presentazione degli oggetti e delle immagini non era immutabile, ma dipendeva dagli spostamenti d’interesse del bambino. La durata del test è stata di circa 10-15 minuti, a seconda della fascia e del bambino. A distanza di pochi giorni, gli stessi bambini sono stati valutati, per le capacità comunicative e linguistiche, tramite il TPL.

3.2.1. Il Test Fonetico della Prima Infanzia

Il test qui usato per la valutazione dello sviluppo delle capacità fonetiche di bambini tra i 18 e i 36 mesi, denominato “Test Fonetico della Prima Infanzia” (TFPI), è diviso in tre sottotest, in base all’età cronologica dei soggetti (18-23 mesi; 24-29 mesi; 30-36 mesi)[1]. Attualmente risulta in fase di perfezionamento, ma non è lontano da quella che dovrebbe essere la sua forma definitiva. La versione presentata risale ad aprile 2011 (per una versione precedente, cfr. Zmarich *et alii*, 2010). Esso è stato costruito in base a tre criteri principali:

1) Criterio Fonetico: i fonemi consonantici della lingua italiana devono essere attestati in almeno due parole del test diverse per ciascuna posizione lessicale: singola iniziale

[1] Abbiamo volutamente fatto riferimento ai 3 sottogruppi dell’app. A di Caselli *et alii* (2007)

di parola, singola intervocalica. Per i gruppi consonantici, poiché sono moltissimi e difficilmente reperibili in almeno due nomi concreti frequenti, il test propone non tanto la sequenza degli stessi due o tre foni in due parole diverse, ma la sequenza delle stesse classi fonologiche naturali, che possono essere rappresentate da foni diversi. Il gruppo consonantico iniziale è sempre omosillabico, il gruppo consonantico intervocalico è di preferenza eterosillabico. Le due parole, intese come possibilità offerte al bambino, devono garantire la produzione di almeno una parola con il fono target;

2) Criterio Semantico/Frequenziale: laddove possibile, le parole utilizzate nel test sono tratte dall'appendice A del libro *Parole e Frasi del "Primo Vocabolario del Bambino* (PVB) di Caselli *et alii* (2007), selezionate nella categoria dei sostantivi concreti (criterio semantico, per poterli rappresentare tramite immagini) in base al valore percentuale più alto all'interno di ciascuna delle 3 fasce di età (criterio frequenziale: i valori percentuali si riferiscono al campione normativa del PVB).

3) Criterio della Gradualità nella Complessità Fonetica: le parole utilizzate mostrano un aumento di complessità, dalla prima alla terza fascia, per numero e tipi di sillabe (cioè parole più lunghe e con gruppi consonantici via via più complessi), così come queste emergono da alcuni studi sullo sviluppo fonetico dei bambini italiani (Zmarich & Bonifacio, 2005; Zanobini *et alii*, 2012). Inoltre, poiché la mancata pronuncia di una certa consonante potrebbe essere dovuta più alla lunghezza della parola che la contiene che all'effettiva incapacità del bambino a produrla, le strutture foniche che non compaiono nella fascia precedente sono inserite in parole corte e semplici nella fascia successiva (per es., di forma *CVCV*). Viceversa, per costruire le parole lunghe vengono utilizzati i fonemi già presenti nella fascia precedente. In questo modo le due possibili fonti di difficoltà (strutture foniche nuove/maggiore lunghezza) non sono mai compresenti nella stessa parola.

La costruzione di questo nuovo strumento mira, perciò, a raggiungere i seguenti obiettivi: disporre di un test di articolazione in grado di fornire una valutazione delle capacità fonetiche e fonologiche dei bambini piccoli, in termini di foni prodotti e di processi di semplificazione delle parole target (i cosiddetti processi fonologici), già a partire dai 18 mesi; raccogliere attraverso questo strumento dati normativi sulla popolazione italiana, estendendo la raccolta a campioni più numerosi, e variati per caratteristiche socioeconomiche e geografiche; progettare l'analisi anche di popolazioni con linguaggio atipico (per es. ritardi di linguaggio, disordini fonologici, disturbo specifico di linguaggio ecc…). Le parole - stimolo del test sono riportate in tabella 2.

La capacità fonetica di un bambino viene indicizzata rapportando in percentuale il numero dei foni\gruppi consonantici da lui prodotto al n. complessivo producibile: per dare come attestato un fono o un nesso consonantico è sufficiente che produca almeno la metà dei target (50%) che contengono il fono o il nesso in questione. Inoltre, lo stesso target viene testato in due posizioni distinte di parola, all'inizio e in posizione mediana.

PAROLE	18-23 mesi	24-29 mesi	30-36 mesi
Animali	cane, gallo, gatto	coniglio, gallina, gatto, giraffa, lupo, maiale, pesce, rana, scimmia, tartaruga	coniglio, elefante, gallina, gatto, giraffa, lupo, pesce, pinguino, rana, scimmia, scoiattolo, tartaruga, zanzara, zebra
Verso degli animali	bubu/baubau	/	/
Giocattoli	cubo, lego, palla	cubo	martello, tromba
Veicoli	bici	bici	barca, bici, moto
Cibo	latte, pappa, pomodoro, succo, torta	biscotti, ciliegie, formaggio, fragola, gelato, pasta, pizza, pomodoro, succo	biscotti, caramella, ciligiege, cioccolata, formaggio, fragola, pasta, pizza, pomodoro, succo, torta
Abbigliamento	ciabatta	scarpe, ciabatte	ciabatte, guanti, sciarpa
Parti del corpo	bocca, capelli, denti,dito, gambe, mano, piede	bocca, braccio, denti, dito, lingua, naso, piede, unghie	denti, dito, lingua, naso, piede
Oggetti d'uso familiare	biberon, chiave, ciuccio, tappo	biberon, chiave, ciuccio, cuscino, cucchiaio, scopa, tazza, telefono, vasino	biberon, chiave, ciuccio, cuscino, cucchiaio, giornale, sapone, spazzolino, straccio, vasino
Persone	mago, mamma	mamma	/
All'aperto	luna, sasso	foglia, scivolo, sole	foglia, sasso, spiaggia, strada
Routine	ciao, nanna, no, si	nanna	/
Oggetti della casa	/	cassetto	finestra, tavolo
Aggettivi	/	rosso, verde	rosso, verde
Posti dove andare	/	/	giostra, scuola

Tabella 2: Parole stimolo suddivise per fasce di età e per categorie semantiche

La produzione del target riceverà, dunque, un punto (cioè sarà attestato) per ciascuna delle due posizioni. Le ripetizioni “immediate” non sono considerate. Ad esempio, per attestare [g] in posizione iniziale di parola per la prima fascia d’età vengono proposte le parole “gatto” e “gallo”: per dare attestato il fonema è sufficiente che il bambino produca la [g] di una delle due parole target. Per indicizzare la prestazione del bambino, si rapporta il numero dei target realizzati al numero totale di quelli producibili, e lo si esprime in percentuale. Per es., il numero totale di foni e nessi consonantici per la fascia dei 18-23 mesi è di 24, perciò se un bambino ne produce solamente 12 otterrà un punteggio pari al 50%. La scelta di attestare una data struttura fonica in una data posizione di parola sulla base di una sola produzione sulle due a disposizione (come minimo), può essere criticata poiché preclude la possibilità di graduare le prestazioni tra chi produce due volte su due la stessa struttura fonica (produzione consolidata) e chi risponde producendola solo una volta, non rispondendo o rispondendo con un altro fono alla seconda (produzione parziale, cfr. Eisenberg & Hitchcock, 2010). A suo vantaggio, questa soluzione prevede che un bambino possa non produrre semplicemente per stanchezza o disattenzione, semplifica l’attribuzione del punteggio individuale, e scinde la valutazione quantitativa (quanto “bravo” è il bambino?) da quella qualitativa (i foni sono acquisiti pienamente o parzialmente?). L’informazione resta infatti sempre recuperabile. La tabella 3 indica le parole utilizzate per l’elicitazione dei foni nelle diverse posizioni per la prima fascia, e viene riportata come esempio. Le parole in tabella, racchiuse tra le parentesi tonde, rappresentano i target alternativi ancora in fase di verifica. La trascrizione fonetica usata è di tipo SAMPA, vedi *www.phon.ucl.ac.uk/home/sampa/*). Si noterà che una stessa parola è tipicamente rappresentata in più colonne, perché si è cercato nei limiti del possibile di utilizzare lo stesso item per testare più consonanti.

	Posizione			
	iniziale	iniziale gruppo consonantico	mediana	mediana gruppo consonantico
p	palla, pappa, pomodoro	pjede	pappa, (kapelli), tappo	
b	bitSi, biberon, bokka		biberon, (kubo), (ciabatta), (bubu/baubau)	gambe
t	tappo, torta		gatto, latte, dito	denti
d	denti, dito		pjede, pomodoro	
k	kane, (kubo), (kapelli)	kjave	sukko, bokka	
g	gatto, gambe, (gallo)		lego, mago	
m	mamma, mano, mago		mamma, pomodoro	gambe
n	nanna, no		nanna, kane, mano, luna	denti
f				
v			kjave	
s	sukko, si, sasso	sasso		
z				
S				
ts				
dz				
tS	tSuttSo, tSao		tSuttSo, bitSi	
dZ				
l	lego, latte, luna		palla, (kapelli), (gallo)	
L				
r			biberon, pomodoro	
j		kjave, pjede		
w				

Tabella 3: Parole stimolo relative alla fascia d'età 18-23 mesi, ripartite per posizione di parola

Per i gruppi consonantici, presentiamo a titolo di esempio la tabella 4:

	INIZIALE (omosillabici)	MEDIANA (eterosillabici)
OCCLUSIVA + SEMICONSONANTE	**pj**ede **kj**ave	-
OCCLUSIVA NASALE + OCCLUSIVA	-	ga**mb**e de**nt**i

Tabella 4: Parole stimolo relative alla fascia d'età 18-23 mesi, ripartite per posizione di parola

3.2.1. Il Test del Primo Linguaggio

Il Test del Primo Linguaggio (TPL, Axia, 1995) è uno strumento di valutazione dello sviluppo linguistico, sia comunicativo che verbale in senso stretto. Il test è rivolto ai bambini dai 12 ai 36 mesi, con un intervallo normativo di tre mesi in tre mesi. Il TPL si articola in tre scale, con tre sub-scale di comprensione e tre di produzione linguistica, che valutano tre aree: le abilità comunicative e pragmatiche, la conoscenza dei nomi e di semplici concetti linguistici e, infine, le prime abilità sintattiche.

3.2.2. Il Primo Vocabolario del Bambino

Il Primo Vocabolario del Bambino (PVB) è la versione italiana del *MacArthur-Bates Comunicative Development Inventories* (CDI) ed è un questionario per i genitori di bambini tra gli 8 e i 36 mesi, utilizzato sia in ambito di ricerca, sia nella pratica clinica per lo studio e la valutazione delle capacità comunicative e linguistiche nei primi anni di vita in bambini con sviluppo tipico e atipico (Caselli *et alii*, 2007). Delle due schede di cui consiste ("Gesti e Parole", per bambini tra gli 8 e i 17 mesi e "Parole e Frasi", per quelli tra i 18 e i 30 mesi), qui è stata utilizzata sola la seconda, nella forma lunga, che richiede un tempo di compila-

zione di circa 30-40 minuti. La sezione principale della scheda consiste in una lista di 670 parole che i genitori devono attestare come prodotte.

3.3. Trascrizione fonetica e codifica in Phon

Il segnale acustico relativo alla produzione linguistica del bambino, una volta registrato, è stato visualizzato al PC con Praat, per facilitare la trascrizione fonetica in IPA. Questa poi è stata informatizzata col programma Phon (*http://childes.psy.cmu.edu/phon/*). Tale programma implementa moltissime funzioni richieste per le analisi dello sviluppo fonologico, quali la connessione ai dati multimediali, la segmentazione delle unità (nel nostro caso le parole), l’etichettatura, la comparazione sistematica tra il target fonologico e le forme fonetiche effettivamente prodotte dal soggetto (Zmarich *et alii*, 2010). Per ognuna delle parole prodotte dal bambino, è stato, quindi, creato un *record* (cioè una scheda dell’archivio elettronico relativo a una registrazione), che riporta il *target*, trascritto al livello d’informazione *Orthography* in caratteri alfabetici come glossa, e trascritto al livello d’informazione *IPA Target* in simboli IPA. La effettiva produzione del bambino è trascritta al livello d’informazione *IPA Actual* in simboli IPA. Inoltre per ciascun record è stato possibile specificare su un livello di annotazione separato se la parola era stata prodotta spontaneamente (S) o su ripetizione (R). Il target lessicale e la produzione infantile hanno quindi ricevuto in modo automatico una suddivisione in sillabe (*target syllables, actual syllables*) e un allineamento delle due trascrizioni (*syllables alignement*), in base alle regole di sillabazione dell’italiano (Zmarich *et alii*, 2011) che ripartiscono i segmenti di una sillaba a Attacco, Nucleo e Coda. Poiché ogni record è associato alla porzione del segnale acustico relativo, è possibile verificare in ogni momento la trascrizione ascoltando l’audio corrispondente. Phon permette di eseguire diversi tipi di analisi di uso comune negli studi sull’acquisizione fonologica, che rispondono a due tipologie: analisi indipendente della frequenza di occorrenza di strutture foniche (dai tratti distintivi ai foni e ai tipi sillabici) e analisi relazionale degli errori (tramite il confronto sistematico tra la forma prodotta dal bambino e la forma del target). Quest’ultimo caso è esemplificato, per esempio, dal tipo di analisi conosciuto tradizionalmente col nome di analisi dei processi fonologici (cfr. Bortolini, 1995). Per questo lavoro si è deciso di fare una scelta sui processi fonologici in base alle caratteristiche linguistiche dell’italiano e alla frequenza dei processi nello sviluppo (Bortolini, 1995). La interrogazione (*query*) relativa ad ogni processo in Phon si basa sulla definizione in tratti di ogni segmento, e permette la ricerca di una relazione di presenza/assenza e/o diversità in uno o più tratti tra il segmento *target* e il segmento prodotto effettivamente. Per quanto riguarda i processi che semplificano il sistema, si sono presi in considerazione, tramite una *Search* del tipo *Aligned Phones* a livello di *Project*, con il linguaggio *Phonex*, i seguenti:

- Anteriorizzazione (*Velar fronting*): [*Matches*: Target **{Consonant,Dorsal}** Actual **{Consonant,Labial|Coronal}**]
- Stopping: [*Matches*: Target **{Consonant,Continuant|Affricate}** Actual **{Consonant,Stop}**]
- Posteriorizzazione (*Coronal-Backing*): [*Matches*: Target **{Consonant,Coronal|Labial}** Actual **{Consonant,Dorsal}**]
- Desonorizzazione (*Devoicing*): [*Matches*: Target **{Consonant,- Continuant,Voiced}** Actual **{Consonant,- Continuant,Voiceless}**]
- Sonorizzazione (*Voicing*): [*Matches*: Target **{Consonant,-Continuant,Voiceless}** Actual **{Consonant,-Continuant,Voiced}**]

- Fricativizzazione (Spirantization): [*Matches*: Target **{Consonant,Stop|Affricate}** Actual **{Consonant,Continuant}**]
- Gliding: [*Matches*: Target **{-Glide}** Actual **{Glide}**]

Per quanto riguarda i processi che semplificano la struttura fonotattica:

- Cancellazione (*Deletion*): [*Matches*: Target **{Consonant}**]
- Riduzione dei gruppi consonantici: [*Matches*: Target**{Consonant}{Consonant}** Actual **{Consonant}**]

Nell'analisi della riduzione dei gruppi sono compresi anche i casi di degeminazione. L'escluderli, per ottenere il numero di processi di riduzione dei gruppi consonantici "puri," richiederebbe infatti ricerche più approfondite che qui non possono essere presentate.

4. RISULTATI

4.1. Sviluppo fonetico e fonologico

Il numero medio strutture foniche (foni + gruppi consonantici) per la fascia dai 18 ai 23 mesi è di 8,9 (DS: 2,51, cfr. tab. 5).

			PVB	PVB	PVB	TPL S. Com&Pra		TPL S.Vocabol.		TPL S.Sintassi	
Fascia	n.F.	%n.F.	n.Par.	Cent	E.Les	Pr	Perc	Prod	Perc	Prod	Per
18-23 M	8,9	37,0	189,4	68,0	22,1	1,2	1,4	12,3	7,7	8,3	3,2
18-23 DS	2,5	10,4	74,7	20,3	1,6	0,6	0,7	5,5	5,7	5,6	2,1
24-29 M	25,9	60,2	459,4	68,3	29,6	1,5	2,4	17,4	14,5	11,5	17,6
24-29 DS	5,8	13,5	107,3	24,7	4,2	0,4	0,5	1,7	4,2	7,8	16,2
30-36 M	33,2	69,6	505,9	51,6	31,0	1,5	2,2	18,2	15,9	14,8	32,7
30-36 DS	4,2	8,0	89,9	28,8	3,7	0,5	0,9	0,6	1,7	5,6	12,3

Tabella 5: Media (M) e Deviazione Standard (DS) del n. assoluto e percentuale dei foni prodotti dai tre gruppi di soggetti (18-23 mesi; 23-29 mesi; 30-36 mesi), insieme con il numero di parole attestato al PVB, il centile corrispondente, l'età lessicale e i punteggi relativi al TPL (Scala "Comunicazione e Pragmatica, scala "Vocabolario", scala "Prima Sintassi)

Generalmente i soggetti hanno prodotto poche strutture foniche rispetto a quelle producibili, e solamente un soggetto è arrivato a produrre la metà delle strutture previste per l'età. Nella fascia "24-29 mesi" il numero medio delle strutture foniche prodotte è 25,9 (DS: 5,8). Oltre la metà dei soggetti produce un numero di strutture foniche superiore alla metà di quelle previste (43). Per la fascia superiore, dai 30 ai 36 mesi, il numero medio di strutture foniche prodotte dai bambini sale a 33,2 (DS: 4,2). In questo caso, risulta evidente come la totalità dei soggetti ha prodotto un numero di strutture foniche superiore alla metà di quelli previste (49). Di seguito vengono mostrate le figure relative alla percentuale di bambini, per ogni fascia d'età, che possiede un determinato fono (fig. 1) o gruppo consonantico (fig.2) nell'inventario fonetico. Tali risultati sono stati ottenuti sommando il numero di bambini che produce un certo fono e/o un gruppo consonantico, ed esprimendolo in percentuale come proporzione sul totale dei 10 bambini per ogni fascia. Si noti che le soglie numeriche sono le stesse delle analoghe figure esposte in Bortolini (1995) e in Zanobini *et alii* (2012) per consentire un confronto diretto con i loro risultati in sede di discussione. E' possibile notare che in posizione iniziale di parola i foni [p], [b] e [t] fanno parte dell'IF di più

dell’80% dei bambini già dalla prima tappa d’età, i foni [k], [m], [n], e [v] dalla seconda tappa, mentre [d], [f] e [s] solo tra i 30 e i 36 mesi. A 18-23 mesi tra i foni in possesso del 50-80% dei bambini vi sono [k] e [m]; mentre per i foni [d], [f] e [s] bisogna aspettare la seconda fascia d’età, e per i foni [ ], [dz] e [ ], la terza. I foni restanti sono posseduti da meno del 50% dei bambini in tutte le fasce d’età. In posizione intervocalica, nella fascia dei bambini più piccoli sono solamente tre i foni ([p], [t] e [n]) che fanno parte dell’IF di più dell’80% dei bambini, mentre solo due ([v] e [ ]) quelli in possesso del 50-80% dei soggetti. Relativamente ai 24-29 mesi è possibile constatare come circa la metà dei foni facciano parte dell’IF di più dell’80% dei bambini, mentre un quarto degli altri foni sono prodotti dal 50-80% dei bambini e un quarto da meno del 50%. A 30-36 mesi la maggior parte dei foni sono prodotti da più dell’80% dei bambini, mentre tre foni ([g], [ ] e [ ]) rientrano nell’IF del 50-80% dei soggetti. Per quanto riguarda i gruppi consonantici (fig. 2), in nessuna fascia d’età sono presenti in posizione iniziale di parola in più dell’80% dei bambini. Il 50-80% dei soggetti inclusi nelle fasce dei 24-29 mesi e 30-36 mesi produce un solo gruppo consonantico (occlusiva + semiconsonante). Tutti gli altri gruppi consonantici sono attestati in meno del 50% dei soggetti. Gli unici gruppi consonantici in posizione mediana prodotti da oltre l’80% dei bambini sono nella fascia dei bambini più grandi: occlusiva più semiconsonante e nasale più ostruente. In questa fascia d’età il gruppo composto da una nasale più una occlusiva più una semiconsonante è posseduto dal 50-80% dei bambini, i restanti gruppi da meno del 50%. Gli unici due gruppi individuabili nel 50-80% dei bambini, a 23-29 mesi, sono quelli omosillabici (sonorante più semiconsonante) e quello eterosillabico (sonorante più ostruente più semiconsonante).

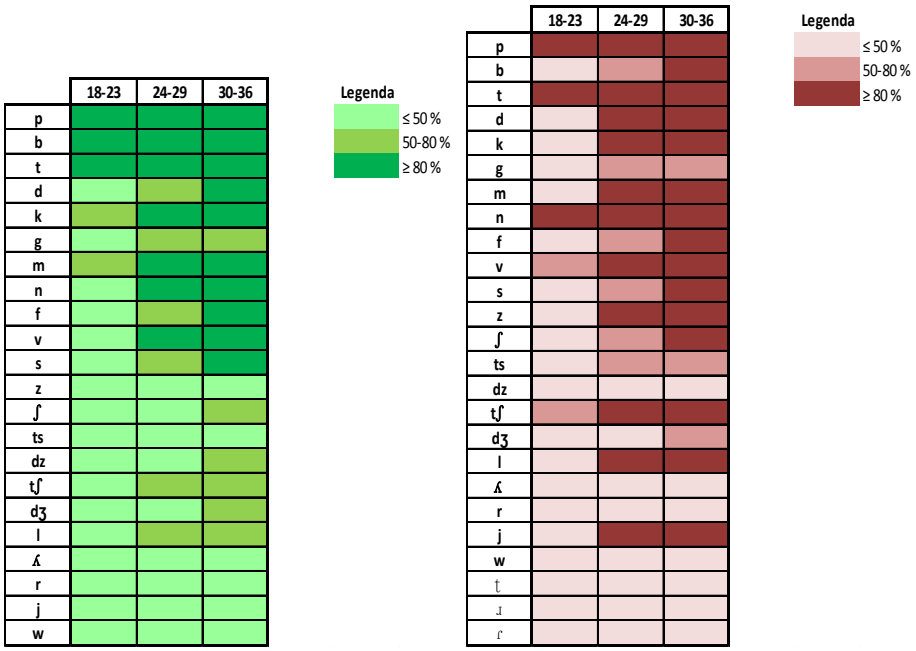


Figura 1: % di bambini che produce un dato fono in posizione iniziale (sx) e intervocalica (dx)

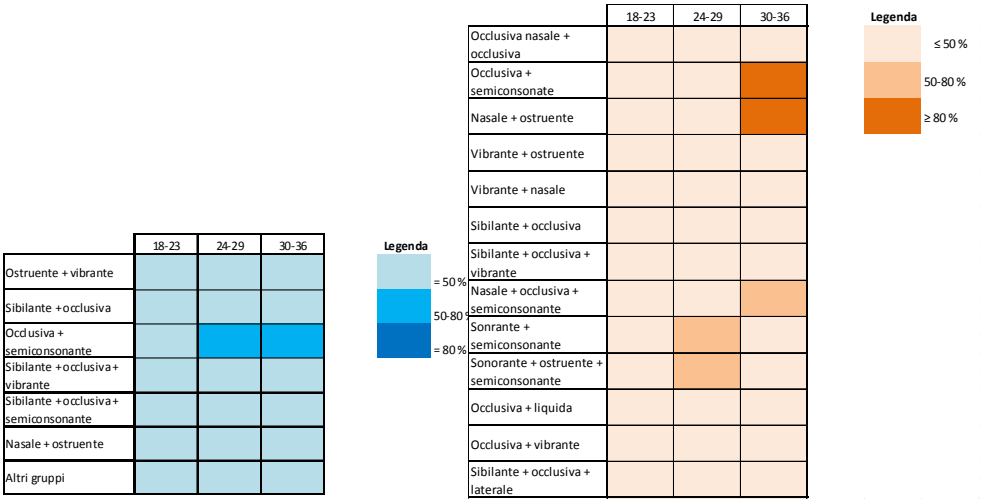


Figura 2: % di bambini che produce un dato Gruppo Consonantico in posizione iniziale (sx) e intervocalica (dx)

Passiamo ora ai risultati dell'analisi dei processi fonologici. Se un segmento è affetto da più di un processo, questi vengono contati tutti, contribuendo ad aumentare il n. totale dei processi. Questi dati assoluti però ancora ci dicono poco se non sappiamo a quanti segmenti i processi potevano potenzialmente applicarsi. Per ciascuno dei processi all'interno di ciascuna fascia d'età, Phon ha calcolato il numero effettivo dei segmenti fonetici colpiti da ogni singolo processo e il numero dei contesti potenziali (cioè dei segmenti) a cui ogni singolo processo poteva teoricamente applicarsi. In fig. 3 vengono mostrate le percentuali di ciascun processo fonologico, espresse come numero di segmenti semplificati rispetto al numero totale possibile di segmenti a cui si poteva teoricamente applicare quel processo.

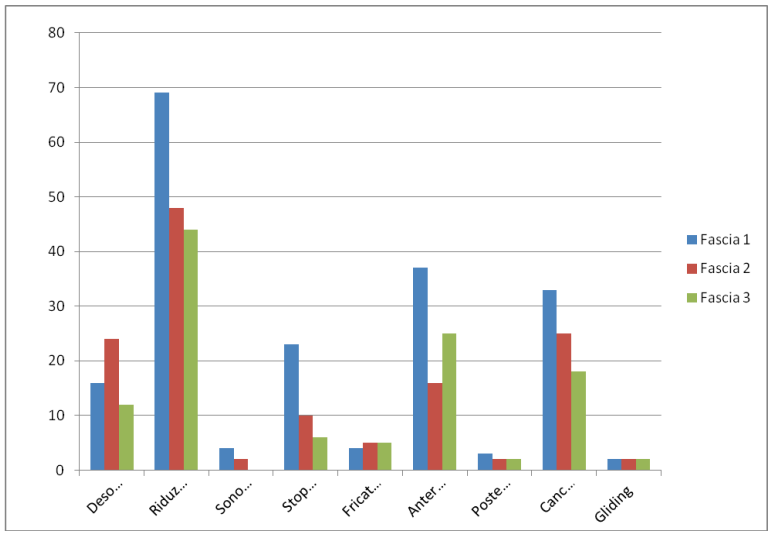


Figura 3: frequenza % di ogni processo fonologico in ciascuna fascia d’età

Per chiarire meglio come sono state calcolate le percentuali, ci si può riferire al caso concreto del processo di “desonorizzazione”, dove si è posto uguale a cento il numero totale di tutti i segmenti sonori non sonoranti (cioè delle ostruenti sonore) nel corpus e si è calcolata la proporzione percentuale su questo totale dei segmenti che risultavano desonorizzati.

Pur con tutte le cautele dovute alla scelta da noi fatta di escludere dalla ricerca alcuni processi, i dati relativi ai processi selezionati sono molto chiari e parlano di una riduzione progressiva, con l’età, di quasi tutti i processi. Aggiungiamo inoltre che questi grafici fanno vedere chiaramente che alcuni processi sono attestati in modo più importante rispetto ad altri, come il processo di riduzione dei gruppi consonantici (che, ricordiamo, qui comprende anche i processi di degeminazione). Considerando i processi riscontrati in questo campione è possibile identificare la cronologia dei processi presenti nello sviluppo fonologico normale. Il primo processo a estinguersi è quello di sonorizzazione, seguito da quelli di posteriorizzazione e di gliding. Gli ultimi invece sono riduzione dei gruppi e anteriorizzazione, rispettivamente con il 44 e il 25%.

Per valutare la validità e affidabilità del TFPI, è stata eseguita una serie di analisi statistiche di correlazione lineare tra i punteggi individuali del test fonetico (percentuale di foni prodotti sul totale dei foni teoricamente producibili) e l’età cronologica, insieme con tre punteggi ricavati dal PVB, e 6 punteggi ricavati dai sub test del TPL (per i punteggi medi e DS si rimanda alla tabella 5). Si ricorda che il TPL usa 3 scale (Scala Comunicativa e Pragmatica, Scala del Vocabolario e Scala della Prima Sintassi), suddivise ciascuna in due subscale di comprensione e di produzione linguistica. Inoltre, poiché la Scala Comunicativa e Pragmatica si compone di varie prove, alcune relative alla comprensione e altre alla produzione, si è deciso di riportare nelle tabelle solo il punteggio medio delle due tipologie di prove (comprensione e produzione).

	VARIABILE DIPENDENTE - FONI PRODOTTI SU QUELLI PRODUCIBILI (%)			
	K di Pearson	Bartlett Chi-square statistic		
VARIABILI INDIPENDENTI	valore	valore	g.d.l.	probabilità
ETÀ CRONOLOGICA	**0,766**	24,324	1	0,000
PVB	0,713	19,557	1	0,000
PVB - PERCENTILE APPROSSIMATO	0,012	0,003	1	0,953
ETÀ LESSICALE	0,814	27,738	1	0,000
TPL - MEDIA PROVE DI PRODUZIONE SCALA COMUNICATIVA E PRAGMATICA	0,237	1,588	1	0,208
TPL - MEDIA PROVE DI COMPRENSIONE SCALA COMUNICATIVA E PRAGMATICA	0,436	5,798	1	0,016
TPL - PROVE DI COMPRENSIONE SCALA DEL VOCABOLARIO	0,514	8,436	1	0,004
TPL - PROVE DI PRODUZIONE SCALA DEL VOCABOLARIO	0,653	15,283	1	0,000
TPL - PROVE DI COMPRENSIONE SCALA DELLA PRIMA SINTASSI	0,447	6,125	1	0,013
TPL - PROVE DI PRODUZIONE SCALA DELLA PRIMA SINTASSI	0,581	11,345	1	0,001

Tab. 5. Analisi della correlazione lineare tra il punteggio individuale al test fonetico (% foni prodotti su producibili) ed alcune misure linguistiche del PVB e del TPL

I valori della Scala del Vocabolario e della Scala della Prima Sintassi sono stati ottenuti sommando il numero di figure riconosciute e nominate dal soggetto rispetto al numero totale delle figure (20) della prova di produzione e di comprensione nella Scala del Vocabolario e solo della prova di comprensione nella Scala della Prima Sintassi, mentre per la prova di produzione di quest'ultima scala il punteggio totale è di 60 in base al punteggio (da 0 a 3) ottenuto dal bambino per ogni figura. Tutti i valori riportati, per essere interpretati, sono stati trasformati in percentili in base alle tabelle del TPL (Axia, 1995: 52-54). Il grado di associazione è stato misurato con il coefficiente di correlazione lineare di Pearson (r), tra la variabile sperimentale (rappresentata dal numero di foni prodotti dai 30 soggetti, senza distinzione di fascia d'età) e le altre variabili, considerate separatamente. La maggior parte delle variabili, eccetto il PVB percentile e la media delle prove di produzione e comprensione della Scala Comunicativa e Pragmatica e della scala di comprensione della Prima Sintassi, sono significativamente correlate alla variabile sperimentale "foni prodotti su quelli producibili (%)", $p \leq .005$, da cui si conclude che i risultati ottenuti col TPFI correlano bene con altri aspetti importanti dello sviluppo linguistico.

4. DISCUSSIONE E CONCLUSIONI

Per quanto riguarda la descrizione delle caratteristiche relative all'acquisizione fonetico-fonologica dei bambini italiani, è possibile istituire un confronto sugli inventari fonetici presentati nella presente ricerca e quelli pubblicati da Bortolini (1995), da Zmarich e Bonifacio (2005) e da Zanobini *et alii* (2012). Volendo fare una prima considerazione generale sui dati ottenuti dalla presente ricerca, e considerando solo i foni attestati per oltre l'80%, possiamo affermare che, mentre per la prima fascia d'età in cui il numero di foni in posizione iniziale e mediana sono in numero uguale, nelle altre due fasce d'età il numero dei foni in posizione mediana è superiore rispetto a quelli in posizione iniziale. Nell'inventario fonetico della prima fascia d'età sono presenti soprattutto occlusive sorde, che sono articolatoriamente più semplici delle corrispettive sonore per quanto riguarda la realizzazione del VOT (la produzione del VOT negativo per le occlusive sonore italiane, che richiede la manovra supplementare dell'abbassamento della glottide, pone una vera sfida ai bambini di questa età). Compaiono inoltre l'approssimante laterale e l'affricata palatoalveolare. Nella seconda fascia si assiste ad un consolidamento di tutti i foni occlusivi e un ingresso massiccio di fricative e nasali. Relativamente all'ultima fascia d'età, è possibile notare che l'inventario fonetico è piuttosto completo sia in posizione iniziale che in posizione mediana. In particolare, la maggior parte delle consonanti che nelle precedenti età erano prodotte tra il 50% e l'80% nella terza fascia vengono consolidate. Questi dati sono in generale accordo con quanto riscontrato anche da Bortolini (1995), Zmarich e Bonifacio (2005) e Zanobini *et alii* (2012). Analizzando i gruppi consonantici si può notare che i bambini più piccoli (18-23 mesi) non possiedono alcun gruppo consonantico. I bambini tra i 24 e i 29 mesi, invece, possiedono solo un gruppo consonantico in sede iniziale di parola, composto da un'occlusiva e una semiconsonante. Infine, nell'ultima fascia d'età i bambini possiedono un solo gruppo consonantico in posizione iniziale (occlusiva più semiconsonante) e due

gruppi in posizione mediana (occlusiva più una semiconsonante e nasale più ostruente, quest’ultimo eterosillabico). Qui il confronto con Bortolini (1995) e Zanobini *et alii* (2012), che non considerano i gruppi consonantici, non è possibile.

Per quanto riguarda l’analisi dei pattern di errore, dovuti a sostituzione o a processi fonologici, si assiste a una riduzione progressiva, in base all’età cronologica, di quasi tutti i processi. Il processo che colpisce di più le produzioni lessicali della stragrande maggioranza dei bambini è il processo di cancellazione, soprattutto dei gruppi consonantici, che vengono ridotti, e poi delle consonanti geminate, che vengono degeminate. Questo dato è in accordo con Zanobini *et alii* (2012). Il primo processo a estinguersi è quello di sonorizzazione, seguito da quelli di posteriorizzazione e di *gliding*. Gli ultimi invece sono quelli di riduzione dei gruppi e di anteriorizzazione, ancora abbondantemente attestati al 36° mese.

In generale, per quanto riguarda il test, possiamo dire che presenta vari punti di forza, tra i quali: 1) la facilità e la velocità di somministrazione; 2) la semplicità e la familiarità del materiale visivo utilizzato nelle diverse fasce d’età, 3) la validità di costrutto, perché la sua applicazione produce risultati che concordano con quelli di strumenti simili (PFLI) e 4) la buona correlazione con altri aspetti importanti dello sviluppo linguistico (TPL, PVB). Un aspetto critico potrebbe riguardare la tripartizione in base all’età cronologica dei soggetti, perché penalizzante rispetto ai bambini eccezionalmente dotati, e ai bambini ai confini di fascia. Una soluzione potrebbe essere quella riportata da Zmarich *et alii* (2010): “allo scopo di valutarli e classificarli correttamente, è stata prevista la possibilità che un bambino che produca in modo accurato tutte, o quasi tutte, le parole di una certa lista (la soglia per poter passare alla lista successiva si potrebbe fissare alla produzione del 90% delle parole), possa accedere alle parole della fascia successiva che contengono le strutture non ancora prodotte (siano essi foni isolati e in gruppo consonantico, o parole lunghe per numero di sillabe)”. Una carenza al cui rimedio gli autori stanno già lavorando riguarda la assenza di parole che possano valutare esplicitamente il contrasto di lunghezza consonantica e quello di sonorità, molto importanti nella lingua italiana e spesso fonte di difficoltà per i bambini. Rispetto ai criteri di Eisenberg & Hitchcock (2010) nel TFPI è stato fatto uno sforzo generico per differenziare i contesti vocalici, e mantenere uno schema accentuale di tipo parossitono, ma questo obiettivo non è stato perseguito in modo esplicito e sistematico.

RINGRAZIAMENTI

Alla prof.ssa Chiara Levorato, relatrice della tesi di Ilaria Fava. A un revisore anonimo, che ci ha permesso di migliorare il testo.

BIBLIOGRAFIA

APA, DSM-IV-TR (2001), Manuale diagnostico e statistico dei disturbi mentali, Elsevier-Masson.

Axia, G. (1995), Test del Primo Linguaggio – TPL: manuale, Firenze: Organizzazioni Speciali.

Bortolini, U. (1995), PFLI Prove per la valutazione fonologica del linguaggio infantile, Padova: Edit Master Srl.

Caselli, M.C., Pasqualetti, P. e Stefanini, S. (2007), Parole e frasi nel “primo vocabolario del bambino”. Nuovi dati normativi fra 18 e 36 mesi e Forma breve del questionario, Milano: FrancoAngeli.

Del Monego, G. (2011), Inventari Fonetici e processi fonologici in bambini dai 18 ai 36 mesi d’età valutati con un nuovo test fonetico, tesi di Laurea in Logopedia, Università di Padova, AA 2010-2011

Eisenberg, S.E. & Hitchcock, E. R. (2010), Using Standardized Tests to Inventory Consonant and Vowel Production: A Comparison of 11 Tests of Articulation and Phonology, Language, Speech, and Hearing Services in Schools, 41, 488-503.

Fava, I. (2011), Un nuovo test fonetico per bambini dai 18 ai 36 mesi, Tesi di Laurea Magistrale in Psicologia, Università di Padova, AA 2010-2011

Giulivi, S., Vayra, M., Zmarich, C. (2010), "Lo sviluppo fonetico di due bambini italiani: dalle caratteristiche universali a quelle linguo-specifiche", in Atti del 6° convegno AISV (F. Cutugno, P. Maturi, R. Savy, G. Abete, I. Alfano, a cura di), Napoli, 3-5 febbraio 2010, Torriana: EDK, 233-248.

Pedrabissi, L. e Santinello, M. (1997), I test psicologici, Bologna: Il Mulino.

Shriberg, L.D. e Kwiatkowski, J. (1985), Continuous speech sampling for phonologic analyses of speech-delayed children, Journal of Speech and Hearing Disorders, 50: 323-334

Stokes, S.F., Klee, T., Carson, C.P., Carson, D. (2005), A Phonemic Implicational Feature Hierarchy of Phonological Contrasts for English-Speaking Children, J. Speech, Language and Hearing Research, 48, 817–833.

Vance, M., Stackhouse, J., Wells, B. (2005), Speech – production skills in children aged 3-7 years, International Journal of Language Communication Disorders, 40, 1, 29-48.

Zanobini, M., Viterbori, P., Saraceno, F. (2012), Phonology and Language Development in Italian Children: An Analysis of Production and Accuracy, J. Speech Lang. Hear. Res., 55, 16-31.

Zmarich, C. (2008), L'emergere dei suoni dell'italiano in una prospettiva interlinguistica, in G. Marotta, L. Costamagna (a cura di), *Acquisizione linguistica e teorie fonologiche*, Pisa: Pacini, 43-65.

Zmarich, C. (2010), Lo sviluppo fonetico/fonologico da 0 a 3 anni, in Bonifacio S., Hsvastja Stefani L., L'intervento precoce nel ritardo di Linguaggio. Il modello INTERACT per il bambino parlatore tardivo, Milano: FrancoAngeli, 17-39.

Zmarich, C. e Bonifacio, S. (2005), Phonetic inventories in Italian children aged 18-27 months: a longitudinal study, in Proceedings of INTERSPEECH'2005 – EUROSPEECH, Lisboa, September 4-8, 2005, 757-760.

Zmarich, C., Bonifacio, S., Bardozzetti, M. P., Pisciotta, C. (2010), Test fonetico della prima infanzia per bambini dai 18 ai 36 mesi: analisi con "Phon" dei primi dati raccolti, in Atti del 5° convegno AISV "La dimensione temporale del parlato" (S. Schmid, M. Schwarzenbach & D. Studer,, Eds), Zurigo, 4-6 febbraio 2009, Torriana: EDK, 567-588.

Zmarich C., Dispaldro M, Rinaldi P., Caselli M. C. (2011), Caratteristiche fonetiche del "Primo Vocabolario del Bambino", Psicologia Clinica dello Sviluppo, XV, 1, 235-256.

MEASUREMENTS OF VIBRATO PARAMETERS IN A PERFORMANCE OF SARDINIAN TRADITIONAL SINGING POETRY

Paolo Bravi
Conservatorio di Musica di Cagliari
pa.bravi@tiscali.it

1. SUMMARY

Artists and Sardinian music experts agree in considering vibrato an essential feature of the traditional singing styles of Sardinia. According to many of them, "Sardinian vibrato" is something 'typical'. It is deemed like a stylistic seal of the traditional vocality of the Island, a kind of ornamentation different from the vibrato of other singing styles.

No formal analysis has been carried out so far to shed light on this aspect of vocal ornamentation. In this paper, measurements of the acoustic features of vocal vibrato in four singing poets are presented and discussed. The research corpus includes 506 segments of vibrato that have been recorded during a real performance held in a small village of Southern Sardinia in Summer 2010, a poetic contest named *cantada campidanesa* performed by the improvising/singing poets (*cantadoris*) with the accompaniment of a two-part choir (*su bàsciu e contra*).

On one hand the results of the analysis show the heterogeneous characteristics of the poets, on the other they stress the tendency towards a vibrato with high *rate* (up to 8 Hz), narrow *extent* (not exceeding a mean value of ±0.4 st) and a notable waving in the overall *intonation*, the latter being measured with reference to the spread between the overall F0 mean and the moving mean between adjacent peaks in each segment. In a second step of the analysis, vocal timbre (i.e. singing on different vowels), pitch (singing on different scale degrees) and duration have been tested as prospective predictors of variability in the examined vibrato features. Whereas vocal timbre does not seem to be a valid explaining factor (apart from one exception), scale degree and duration are able to account for the variance in various respects.

2. INTRODUCTION

Sardinian traditional music comprises many distinct vocal and instrumental genres and styles. While some characteristics are relevant only to some of them, others occur in most or all of them. The attitude towards a strong ornamentation of melodies belongs to the latter ones. It can be remarked all over the Island, both in instrumental and in vocal styles.

Particularly in the Southern part of Sardinia, flourishing art is considered an essential part of the musicians' skill and a major characteristic in the vocal practice. According to Giovanni Meloni, a renowned guitar player of traditional music, the art of *froriduras* ('flourishes, ornamentation') is to be considered an essential aspect of the 'musical idiom' of Southern Sardinia:

> "Nosus, in su Campidanu e in su Sulcis specialmenti, amaus meda, stimaus de fai frodiduras, poita ca funt cosas giai arabas, unu pagheddu. Custas froriduras donant unu sabori diversu, prus sardu, antigun, de sa cantada. Ti portu un'esempru de una boxi. Ddoi est unu cantadori de Teulada [Paolo Mura], chi cantendi a mutetus longus puru,

> oltre ca a sa ghitarrina, fait custas froriduras, ca in sa musica, sa ghitarra naraus 'acciaccatura'. Beh, custu ddu fait cun sa boxi, ma praxit meda. Praxit, parit ca si bieus rapresentaus… su gustu nostu est rapresentau cun custa froridura, fait pròpiu parti de s'idioma nostu"
> [tr. from Sardinian (Campidanese): "Particularly in su Campidanu and in su Sulcis [Southern Sardinia], we love making froriduras because they are a kind of an Arabian thing, somehow. Froriduras give the singing a different flavour, a more Sardinian and more ancient one. Let me take a voice as an exemple. There is a singer from Teulada [a village in the south-west coast of Sardinia; the singer is Paolo Mura] who, while singing either in the a mutetus longus style and in the a sa ghitarrina [with the guitar] style, makes these froriduras, which in music, on the guitar, we call 'acciaccatura'. Well, he makes this with his voice, and people appreciate very much. It seems that we feel represented… our taste is represented by this froridura, it is really a part of our singing idiom" [interview to Giovanni Meloni, 10.07.2011]

Vibrato is an aspect of the art of ornamentation which, as opposed to other forms of embellishments, concerns only the singing voice. A common trend toward a particular style of vibrato may be seen in all traditional Sardinian singing styles[1]. Giorgia Loi, a singer with experience both in popular and Sardinian traditional singing, argues that vibrato is different in either case:

> "il vibrato sardo è molto più fitto… le onde del vibrato sono molto più fitte, il vibrato americano è molto più largo. E poi il vibrato sardo ha dei punti che secondo me non si definiscono neanche vibrato, adesso mi sfugge il termine, dove c'è magari un singhiozzato, ecco, che non viene utilizzato assolutamente nel rock, nel pop"
> [tr. from Italian: "The waves of the Sardinian vibrato are much tighter than those of the American vibrato. Moreover, the Sardinian vibrato in some points – in my opinion – is not really a vibrato, now the word escapes me, there is a sort of "hiccup" which is never used in rock or pop music" [interview to Giorgia Loi, 12.07.2011]

3. MATERIALS

The object of analysis is a real performance of singing poetry in the traditional style of Southern Sardinia, namely the *a mutetus longus* poetical competition which took place in Terraseo-Narcao (Southern Sardinia) on 7 August 2010, in occasion of the festival in honour of Saint James. The performers were the *cantadoris* (semi-professionals improvising and singing poets) Emanuele Saba, Paolo Mura, Antonio Pani, Pierpaolo Falqui (in the following graphs labelled respectively as "Sa", "Mu". "Pa" and "Fq" respectively), with the vocal accompaniment of Antonio Mei (*bàsciu*, 'bass voice') and Antonio Garau (*contra*) (Figure 1). The recording was made with a Tascam DR-100 digital recorder, connected in line to the mixer of the audio service.

[1] According to some authors, a rapid fluctuation of the voice with features similar to those that can be found in the Sardinian traditional song should be defined as *tremolo* (Brodnitz, 1953; Schoultz-Coulon & Battmer, 1981). In this paper this term will not be used: since it has also been employed with different meanings (cfr. Lomax, 1968; Schaum, 1982; Giacomoni, 1996), it is preferable to avoid it.

Figure 1. On the top: Terraseo-Narcao, 7 August 2010: the public place where the *a mutetus* poetic competition took place. On the bottom: from the left to the right, the four improvising / singing poets: Antonio Pani, Emanuele Saba, Pierpaolo Falqui, Paolo Mura.

The selection of the corpus of vibrato samples has been made through manual segmentation of the recording and performed using the software *Praat* (Boersma & Weenink, 2011). The audio segments included in the corpus have been detected through listening and exhibit rapid fluctuations of the pitch, with a more or less regular pattern, during the emission of a vowel.

The identification of vibrato sections in a real performance of *a mutetu* singing is not a straightforward and uncontroversial issue. Two main problems arise.

The first problem concerns the decision of what type of vocal oscillation is to be included in the corpus. In same cases, the pitch oscillation is very slight and just hearable (see example in Figure 2); in other ones, the distinction between vibrato and wider forms of pitch

undulation like series of turns (in Italian musical terminology: *gruppetto*) is not sharp (see examples in Figure 3a and 3b). In both cases, the decision whether to consider them as vibrato (and therefore include them in the corpus) or not is largely subjective.

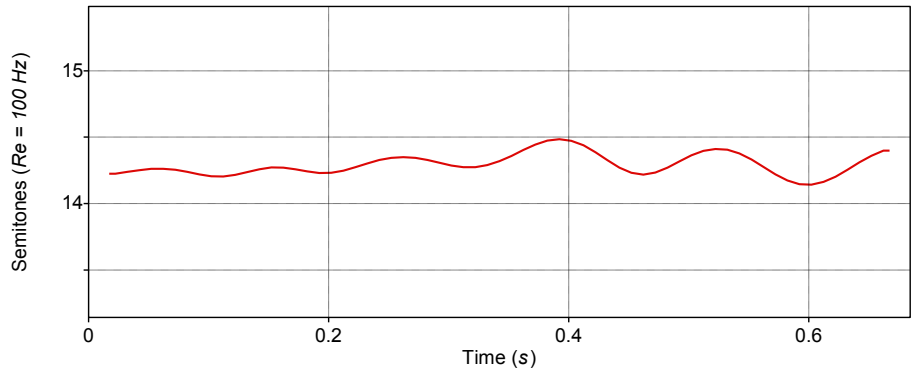


Figure 2. Example of light fluctuation of F0.

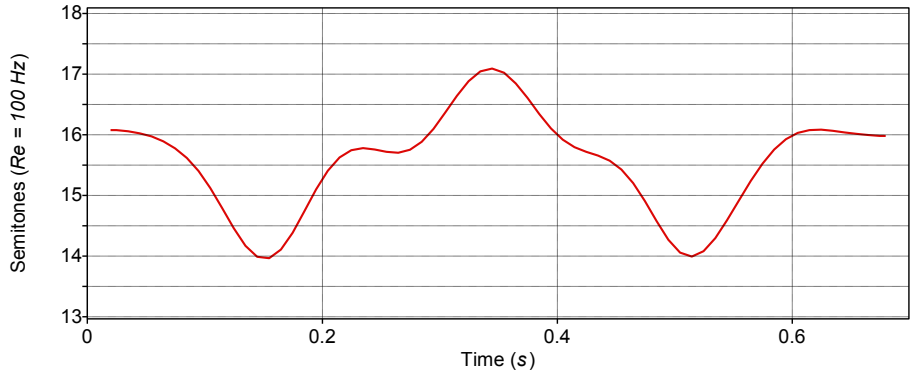


Figure 3a. Example of clearly recognizable turns: the remarkable extent of the pitch fluctuation and the short stops between the downward and upward melodic movements indicate clear shifts from one degree of the scale to the preceding or the following one.

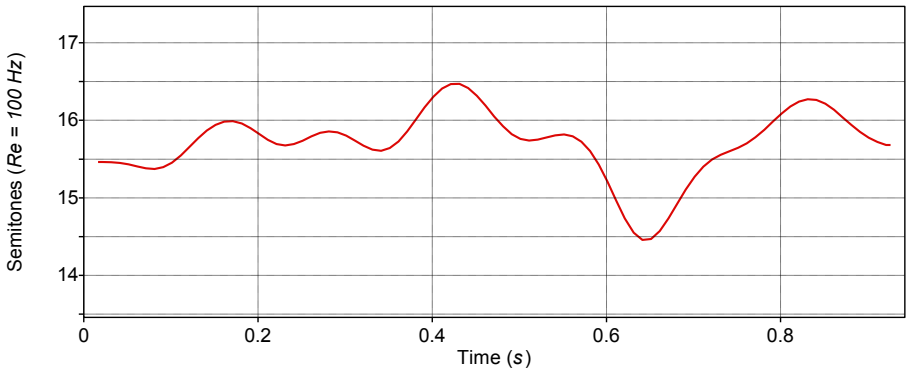


Figure 3b. Example of F0 undulation with no clear boundary between vibrato and turns.

The second problem concerns the exact detection of the starting and end points of the vibrato sections. They are in some cases neither easily to detect by ear nor clearly visible on the pitch tracking (see example in Figure 4).

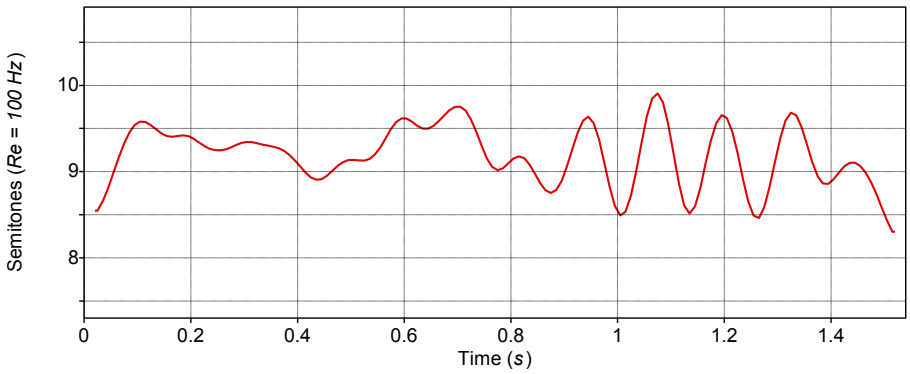


Figure 4. Example of vibrato with non clearly defined starting point.

Every segment of vibrato has been labelled by means of a *Praat* textgrid, with indication of the scale degree and of the vowel used (Figure 5).

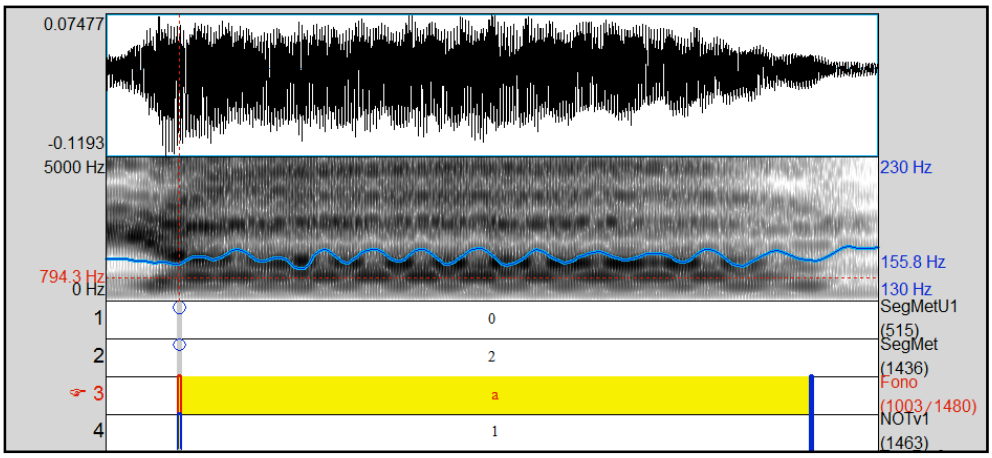


Figure 5. An example of manual annotation of the audio file via *Praat* textgrid.

The fundamental frequency (F0) tracking was carried out via *Praat* and the resulting curves were smoothed by means of the relevant *Praat* function[2]. In order to obtain reliable results, the vibrato segments have undergone a two-step filtering process. These preliminary steps have been performed, as well as the following statistical analysis, by means of the software *R* (RDCT, 2011). The first filter has aimed at excluding from the corpus the segments of very short length. The duration threshold has been arbitrarily set at 0.5 sec. As a consequence, 162 out of 720 (22%) segments have been filtered out. The second filter has been applied after a visual inspection of the data and has aimed at a further selection of the samples by excluding spurious measurements and anomalous values due to inaccuracy in the procedure of peak detections in the F0 curves. This second step has led to the exclusion of 52 out of 558 (9%) segments.

The resulting corpus comprises 506 audio segments, whose distribution is unbalanced across the four singing poets. Almost half of the total number of segments belongs to Paolo Mura, whereas Pierpaolo Falqui only has only a small number of them (Table 1). Concerning the duration of the audio segments, it is noteworthy that in most cases the length of the segments is much shorter than that of the excerpts usually employed in the research on vibrato (up to 5 / 6 seconds long). In this case, vibrato segments reaching or close to the duration of 2 seconds occur only exceptionally (overall mean and SD: 0.76 sec. ± 0.29). Durations of vibrato segments are different in the four poets (longer in Mura than in the other poets) and their distribution is positively skewed (again, particularly in the case of Mura; see Table 1).

[2] Parameters for pitch tracking have been set as follows: time step: 0.001 s; floor / ceiling: 150 / 300 Hz; remaining parameters: default values (Boersma, 1993). The smoothing has been applied with parameter 'bandwidth' set to 10 Hz (*Praat* procedure for smoothing pitch tracks is shortly described here: http://uk.groups.yahoo.com/group/praat-users/message/5529).

Poet	Number of collected items	Per-centage	Duration				
			Mean	SD	Q .25	Median	Q .75
Fq	35	7	0.72	0.26	0.55	0.65	0.80
Mu	247	49	0.81	0.30	0.59	0.70	0.91
Pa	106	21	0.72	0.24	0.54	0.65	0.76
Sa	118	23	0.70	0.28	0.53	0.61	0.72

Table 1. Number and distribution of durations of samples.

4. ANALYSIS

The features of the vibrato that have been taken into account are rate (VR), extent (VE) and intonation (MF0). The results of the analysis are synthesized in Table 2 and discussed in Par. 4.1. Par. 4.2 and Par. 4.3 deals with the effects of vocal timbre, scale degree and duration on the aforementioned vibrato features.

Poet	VRmn	VRsd	VEmn	VEsd	MF0mn	MF0sd
Fq	6.36	1.37	0.19	0.06	0.08	0.06
Mu	7.81	0.51	0.26	0.09	0.10	0.07
Pa	7.66	0.44	0.23	0.05	0.10	0.07
Sa	6.05	0.37	0.40	0.09	0.14	0.09

Table 2. Vibrato features per poet: rate (VRmn: mean; VRsd: SD), extent (VE: mean; VEsd: SD), intonation (MF0mn: mean; MF0sd: SD).

4.1. Vibrato parameters

The typical Sardinian vibrato, as one can hear from Mura's or Pani's singing voices, has a high and relatively constant vibrato rate, close to 8 Hz (Table 2 and Figure 6). Saba's vibrato is, in this respect, quite different, with a mean rate of 6 Hz, which results higher than the one observed in most contemporary male opera singers (Large et alii, 1971; Shipp et alii, 1980) but far from the very high values of Mura and Pani. Falqui's vibrato is more a weak and unstable undulation of voice than a vibrato in the strictest sense of the word, that is to say a pitch oscillation characterised by a relatively regular periodicity. The main datum in his case is related to the standard deviation of the vibrato rate. The mean of the SDs is three times or so higher than those of the other three poets. This means that the rate of the pitch fluctuation in his case is extremely irregular compared to the others'.

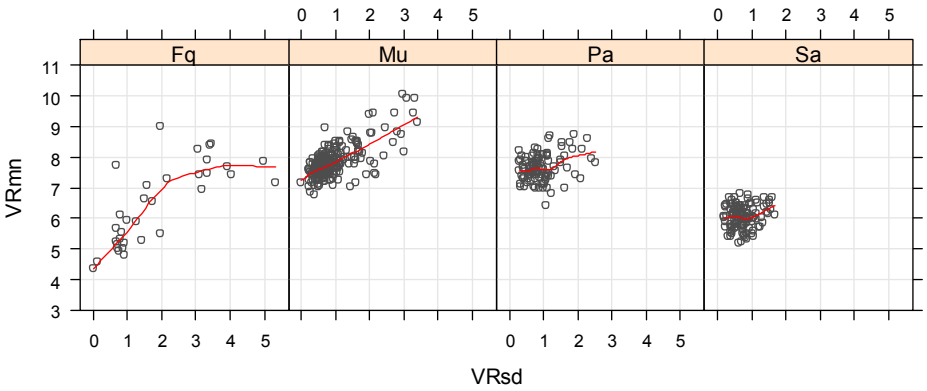


Figure 6. Vibrato rate mean (VRmn) and SD (VRsd), per poet.

The vibrato extent is usually small. The highest value (± 0.4 st) is shown by Saba, while the other three poets' mean vibrato extent does not reach ± 0.3 st (Table 2). A relevant feature is the SD of the values, which is considerably higher in Saba and Mura than in Pani and Falqui (Table 2 and Figure 7).

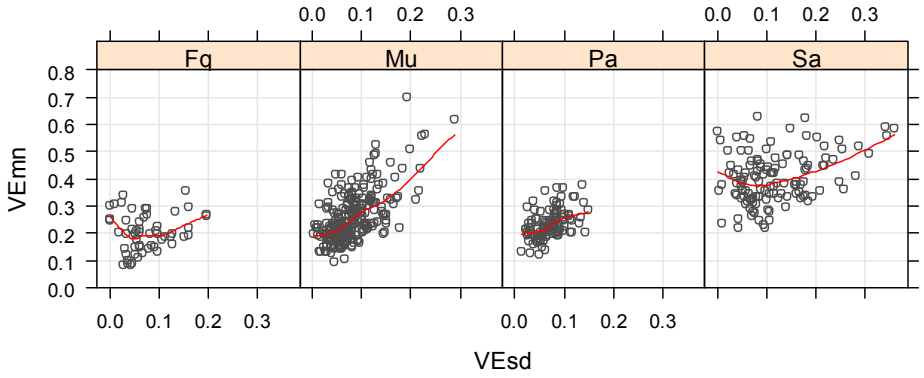


Figure 7. Vibrato extent mean (VEmn) and SD (VEsd), per poet.

The third feature of vibrato which has been taken into account is intonation. Change in intonation has been evaluated referring to the average differential mean (MF0), i. e. by measuring the absolute difference between the overall F0 mean in the segment and the dynamic mean running between pairs of subsequent peaks. In all the four poets mean and SD are strongly correlated, and the change in intonation is remarkably high, particularly in Saba's vibrato where the mean of the means is 0.14 st and the SD is 0.9 st (Table 2 and Figure 8).

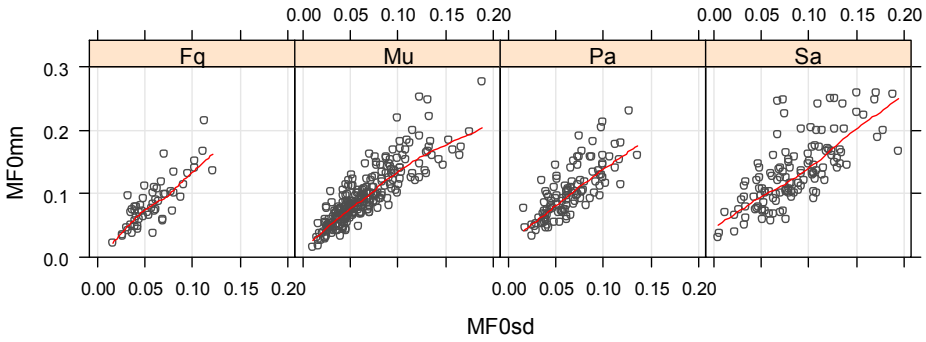


Figure 8. Average differential mean (MF0mn) and SD (MF0sd), per poet.

4.2. Vowels and scale degrees

Vibrato rate is generally considered to be constant within a singer. However, some evidence have been put forward that the rate may be affected by the emotional involvement of the singer (Shipp et alii, 1980), by the age of the singer (Damsté et alii, 1982), by the position of the vibrato cycle in the vibrato emission (Bretos & Sundberg, 2003). Vibrato extent has been described as varying with loudness of phonation (Winckel, 1953; Schoultz-Coulon & Battmer, 1981) and with variation of pitch (Bennett, 1981).

Granted that no previous analysis has been carried out on the vibrato in Sardinian traditional singing styles, some prospective factors of variation have been tested. In particular, potential effects due to vocal timbre (vowel) and pitch (scale degree) have been taken into account and analysed by means of two-way Anova tests, with the following results.

Vowel does not have any significant effect on vibrato rate and extent, with the only exception of Mura's vibrato extent ($F=3.04$, $df=4$, $p=0.017$), with [i] significantly higher than in [a] (Tukey HSD, $p=0.003$).

Scale degree has a significant effect on the vibrato rate in Mura ($F=6.91$, $df=3$, $p=<0.001$) and Saba ($F=4.27$, $df=4$, $p=0.002$). In both cases, the difference is significant for the pair I-IV, but whereas the rate is higher in the I degree than in the IV one in Mura, the opposite happens in Saba (Tukey HSD, $p=<0.001$ [Mura], $p=0.003$ [Saba]). In these two poets scale degree also has significant effects on the vibrato extent ($F=31.96.$, $df=3$, $p=<0.001$ [Mura]; $F=4.20$, $df=4$, $p=0.003$ [Saba]). In both, the difference of the means is significant in the pairs between low and high degrees (I/III and IV/V respectively – Tukey HSD, $p=<0.001$ for all pairs [Mura]; I and IV/V respectively – Tukey HSD, $p=0.04$ (I-IV) and $p=0.006$ (I-V) [Saba]), with low degrees showing a higher mean vibrato extent than the high ones.

4.3. Duration

It has been often observed that vibrato starts and ends with irregular fluctuations. In order to examine if duration may affect the results, a closer examination has been done on the three parameters by dividing the corpus in four equal duration classes with breaks on

the 1°, 2° and 3° quartile and by applying one-way Anova tests. The four classes have been labelled as L (long), ML (medium long), MS (medium short), S (short).

As Figure 9 shows, duration affects the results in various ways. As far as SD is concerned, a common trend in the three parameters emerges. Deviation is significantly higher (or near to the significance level, set at 0.05, in the case of the parameter MF0) in longer segments. As far as mean is concerned, conversely, opposite trends emerge in the vibrato rate and extent. Whereas in the former a rise (near to the significance level) is present in longer segments, in the latter a significant rise ($F=6.63$, $df=3$, $p<0.001$) is present in shorter ones. Finally, no significant difference regards the mean in the parameter MF0.

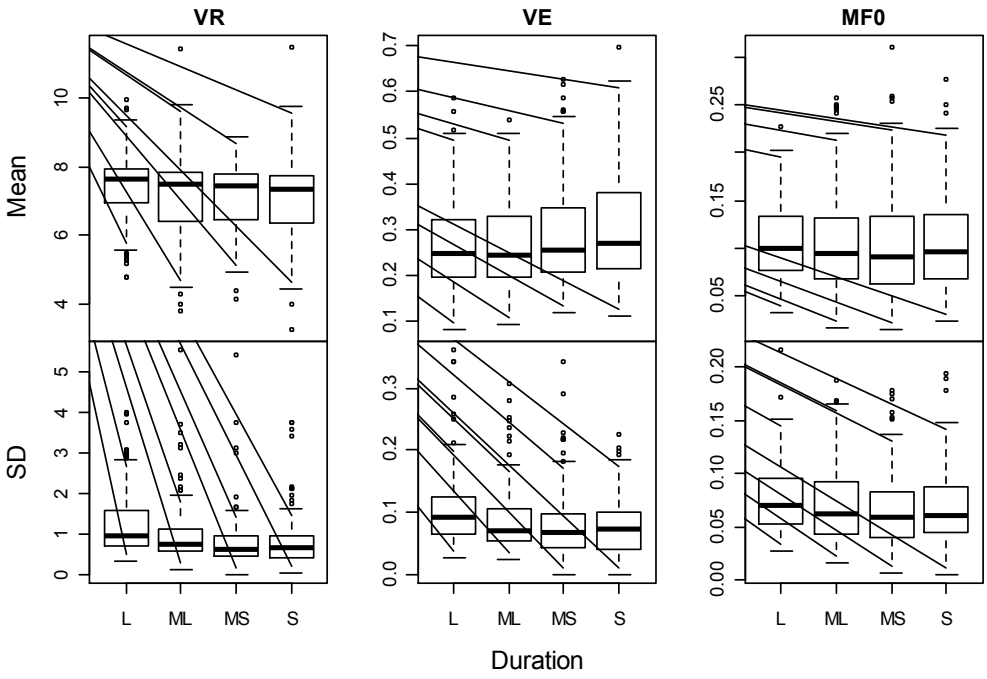


Figure 9. Vibrato rate, vibrato extent and intonation means and SDs in four duration classes (from left to right on each panel: Long, Medium Long, Medium Short, Short).

5. CONCLUSION

Vibrato is an important aspect of the Campidano singing style and of Sardinian song on the whole. The Sardinian vibrato has characteristics that set it apart from analogous vocal ornamentation of other singing styles. Particularly, the results of the analysis carried out on a performance of *a mutetu* singing poetry show that the vibrato in this style has, at least in the cases of some renowned voices, a higher rate and a lower extent than the vibrato commonly used by contemporary male opera singers. Moreover it is characterized by a notable waving of the overall intonation, which has been measured as the spread between the global mean and the moving mean between adjacent peaks of the vibrato fluctuation.

As far as factors that can account for the vibrato rate and extent variability within each singer are concerned, vocal timbre (i.e. singing on different vowels) and pitch (i.e. singing on different scale degrees) have been first tested. Vocal timbre does not seem to be a valid explaining factor: only in one case the difference of vowels gives rise to a significant difference in the rate means. Scale degree seems to have a greater importance, at least in two poets out of four. In particular, in these two poets a notable difference between low degrees and high degrees has been noticed in the vibrato extent. Finally, segment duration is a factor that affects the results, particularly the mean values of the vibrato extent, which are lower in longer segments, and the standard deviation in all features (rate, extent and intonation), which is higher in longer segments.

ACKNOWLEDGEMENTS

The author wishes to thank Antonio Pani, Emanuele Saba, Paolo Mura and Pierpaolo Falqui, the poets who took part in the performance of Terraseo on 7 August 2010, Antonio Dessì, Carlo Schirru, Giorgia Loi, Giovanni Meloni, Ivo Murgia, Paolo Zedda and the anonymous reviewers for their cooperation and help.

BIBLIOGRAPHY

Bennett, G. (1981). Singing synthesis in electronic music. In Research Aspects of Singing (Vol. n. 33, p. 34-50). Stockholm: Royal Swedish Academy of Music.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. IFA Proceedings, 17, p. 97-110.

Boersma, P., & Weenink, D. (2011). Praat: doing Phonetics by computer. Retrieved from http://www.fon.hum.uva.nl/praat/

Bravi, P. (2010). A sa moda campidanesa. Pratiche, poetiche e voci degli improvvisatori nella Sardegna meridionale. Nuoro: ISRE.

Bretos, J., & Sundberg, J. (2003). Measurements of Vibrato Parameters in Long Sustained Crescendo Notes as Sung by Ten Sopranos. Journal of Voice , 17 (3), 343-352.

Brodnitz, F. S. (1953). Keep Your Voice Healthy. New York: Harper and Row.

Damsté, H., Reinders, A., & Tempelaars, S. (1982). Why should voices quiver? In Vox Humana. Studies presented to Aatto Sonninnen (p. 26-34). Institute of Finnish Language & Comm., Univ. Jyväskylä.

Giacomoni, G. (1996). Elementi di teoria musicale. Parma: Azzali.

Large, J., & Ivata, S. (1971). Aerodynamic study of vibrato and voluntary 'straight' pairs in singing. Folia Phoniatrica et Logopaedica , 23, 50-65.

Lomax, A. (1968). Folk Song Style and Culture. New Brunswick, U.S.A.: Transaction Publishers.

Prame, E. (1994). Measurements of the vibrato rate of ten singers. Journal of the Acoustical Society of America (96), 1979-1984.

RDCT. (2011). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved 2011, from http://www.R-project.org

Schaum, W. (A cura di). (1982). Schaum Dictionary Of Musical Terms. Mequon, WI: Schaum Publications, Inc.

Schoultz-Coulon, H., & Battmer, R. (1981). Die quantitative Bewertung des Sängervibratos. Folia Phoniatrica et Logopaedica , 1-14.

Shipp, T., Leanderson, R., & Sundberg, J. (1980). Some acoustic characteristics of vocal vibrato. Journal of Research in Singing , IV (1), 18-25.

Sundberg, J. (1994). Acoustic and psychoacoustic aspects of vocal vibrato. STL-QPSR , 35, 45-68.

Winckel, F. (1953). Physikalische Kriterien für objektive Stimmbeurteilung. Folia Phoniatrica et Logopaedica , 231-252.

PERCEZIONE I

"QUANDO PARLO ITALIANO CAPISCONO SUBITO CHE SONO STRANIERO". PARAMETRI SOPRASEGMENTALI E TEMPI DI REAZIONE AL GRADO DI ACCENTO STRANIERO

Elisa Pellegrino
Università degli Studi di Napoli "L'Orientale"
epellegrino@unior.it

RIASSUNTO

Alla luce del rinnovato interesse verso la dimensione ritmico-prosodica negli studi sulla percezione dell'accento straniero, il presente lavoro si è posto un duplice obiettivo: individuare i parametri soprasegmentali che rendono l'eloquio di apprendenti cinesi di italiano L2 percettivamente deviante da quello dei madrelingua e definire i tempi necessari all'ascoltatore nativo per qualificare la provenienza del proprio interlocutore. Per lo scopo, 16 apprendenti cinesi di italiano L2, con un livello di competenza linguistica intermedio e due madrelingua con funzione di gruppo di controllo sono stati coinvolti in un task di parlato letto. Le performance degli informanti italiani e stranieri sono state oggetto di un test percettivo: a 56 ascoltatori è stato chiesto di valutare il grado di accento straniero di ciascuno speaker. Del corpus di parlato è stato effettuata un'analisi spettro-acustica e sono stati calcolati i seguenti indici ritmico-prosodici: velocità di articolazione e di eloquio, fluenza, range tonale, composizione del parlato e durata media dei silenzi. Dal confronto tra i dati delle analisi spettro acustiche e i risultati del test percettivo è emerso che i parametri più rilevanti a differenziare i tre gradi di accento straniero sono: la percentuale di parlato, la quantità dei silenzi, la fluenza e il range tonale. Nel passaggio dal parlato con accento straniero molto forte a quello percepito come lievemente deviante si assiste ad un aumento progressivo del range tonale, della fluenza e della percentuale di parlato rispetto alla componente silenzio e disfluenze.

Per definire i tempi di reazione all'accento straniero è stato somministrato un secondo test percettivo a 20 ascoltatori campani. Il materiale oggetto della valutazione consisteva nelle sequenze lette dai nativi e dagli speaker cinesi che al test precedente avevano ricevuto un giudizio netto. Dai risultati dell'indagine è emerso che per l'ascoltatore madrelingua riconoscere la provenienza del proprio interlocutore è un compito meno impegnativo rispetto al valutarne l'intensità dell'accento straniero. Sono sufficienti in media 5 secondi per determinare l'appartenenza di un parlante alla propria comunità linguistica; ne servono oltre 10 invece per formulare il giudizio di forestierismo. Confrontando i dati emersi dalla prima e dalla seconda fase della ricerca è emerso che tra i parametri analizzati, il range tonale è rilevante sia per determinare la provenienza del parlante, sia l'intensità dell'accento straniero. La velocità di articolazione e di eloquio invece sono influenti per qualificare l'origine dell'interlocutore; la fluenza invece entra in gioco solo per giudicare il grado di accento straniero.

1. INTRODUZIONE

È opinione condivisa in letteratura che l'acquisizione di una seconda lingua (L2) in età adulta può pregiudicare il raggiungimento di una competenza nella lingua target paragonabile a quella del parlante nativo (Perry & Harris, 2002; Birdsong, 2006). Un apprendente di L2 con un livello di competenza tale da poter essere compreso alla stregua di un madrelingua (Munro & Derwing, 1995), può tuttavia essere percepito come 'straniero' se nel suo eloquio si manifestano gli effetti del transfer fonologico dalla sua L1. Sostituzioni, cancellazioni, aggiunta di foni, spostamenti di accenti (Major, 2001), così pure ridotte velocità di articolazione e di eloquio, fluenza bassa e range tonale monotono (Pettorino et alii, 2011) sono tutti segnali che informano l'ascoltatore nativo del mancato o parziale controllo da parte del proprio interlocutore delle caratteristiche segmentali e sovrasegmentali della L2.
Sebbene vi sia un'ampia letteratura di lavori tesi a determinare i fattori che influenzano la percezione dell'accento straniero, manca un consenso diffuso intorno al ruolo svolto dalla componente segmentale e da quella prosodico-intonativa. Copiosi studi sull'argomento si sono per lo più concentrati sui tratti fonetici che deviano dalla pronuncia dei nativi. Lavori specifici infatti sono stati condotti sull'articolazione delle vocali (Kuhl, 1991; Flege et alii, 1997; Pallier et alii, 1997; Walley & Flege, 1998) e delle consonanti (Flege, 1991; Flege et alii, 1995; Tsukada et alii, 2004; Tsukada, 2005) da parte di apprendenti di L2. Analogamente molti dei recenti modelli proposti per spiegare la produzione e la percezione del parlato di L2 - si considerino ad esempio lo *Speech Learning Model* di Flege (2003) il *Perceptual Assimilation Model* di Best (1995) e l'*Ontogeny Philogeny Model* di Major (2001) – si sono focalizzati esclusivamente sulla produzione e percezione dei segmenti e sugli effetti del transfer dalla L1 alla L2 a livello segmentale. Tuttavia, dal campo stesso della linguistica acquisizionale (De Meo & Pettorino, 2011, 2012; Horgues, 2005), dagli studi sul parlato sintetizzato (Flege et alii, 1997; Grover et alii, 1987; Jilka, 2000; Magen, 1998; Munro, 1995; Ramus & Mehler, 1999) e dai progettisti di sistemi di riconoscimento automatico del parlante e dell'accento straniero (Piat et alii, 2008) provengono dati incoraggianti per quanti considerano il ruolo dei parametri prosodici di primaria importanza nella percezione dell'accento straniero e nel riconoscimento dell'origine dei parlanti di L2. Nella direzione della rivalutazione dei tratti prosodici si inserisce anche la produzione scientifica nazionale e internazionale sulla percezione dell'accento straniero in italiano e sul riconoscimento della L1 del parlante non nativo. (Boula de Mareüil et alii, 2004, Boula de Mareüil & Vieru-Dimulescu, 2006; De Meo et alii, 2012).

2. LO STUDIO

2.1. Obiettivi e articolazione della ricerca

Alla luce del rinnovato interesse verso la dimensione ritmico-prosodica negli studi sulla percezione dell'accento straniero, questo lavoro si è posto come obiettivo quello di indagare i parametri soprasegmentali che rendono il parlato di apprendenti cinesi di italiano L2 percettivamente deviante da quello dei madrelingua. A tale scopo, la ricerca è stata articolata in due sezioni principali: nella prima parte, l'attenzione è stata posta sui correlati acustici del grado di accento straniero. L'intento era stabilire se alcuni parametri prosodici fossero più influenti rispetto ad altri nell'orientare il giudizio di forestierismo dell'ascoltatore madrelingua. Nella seconda parte, invece, sono stati considerati i tempi ne-

cessari al nativo per riconoscere la provenienza del proprio interlocutore e a valutare la qualità della sua performance orale attraverso un giudizio di accento straniero.

3. SEZIONE 1: CORRELATI SOPRASEGMENTALI DELL'ACCENTO STRANIERO

3.1. Partecipanti

Per determinare i correlati soprasegmentali dell'accento straniero dei cinesi, sono stati coinvolti nello studio 16 apprendenti sinofoni di Italiano L2. Gli informanti rappresentavano un gruppo molto omogeneo per età (20-22 anni), livello di competenza linguistica (intermedio), anni di studio della lingua italiana, tempo di permanenza in Italia e percorso di studi intrapreso. Tutti i soggetti avevano già studiato l'italiano in Cina per circa 2-3 anni, al momento del test risiedevano a Napoli da circa un mese e frequentavano un corso di studi di tipo umanistico presso l'Università degli Studi "L'Orientale". Per l'età di prima esposizione alla lingua italiana, i locutori confluivano nella categoria dei bilingui tardivi. Alla ricerca hanno partecipato anche due italiani madrelingua, coetanei dei cinesi, con funzione di gruppo di controllo.

3.2. Materiali e metodi

Per evitare che l'argomento, la lunghezza delle parole e la struttura sintattica delle frasi potesse influenzare l'esito del test sia i partecipanti italiani sia gli studenti cinesi sono stati sottoposti ad uno task di parlato letto. Per annullare la variabile input, quindi, tutti gli informanti hanno letto uno stesso brano di circa 50 parole sulla sindrome del jet lag. Il testo era stato tratto da una rivista di bordo di una compagnia aerea italiana ed era stato leggermente riadattato nel lessico e nella sintassi per facilitarne la comprensione ai lettori cinesi. In seduta di registrazione, il testo è stato letto prima silenziosamente e poi ad alta voce per l'acquisizione del corpus. Le performance degli italiani e degli stranieri sono state registrate in sessioni singole, in una camera anecoica con il software Sony Sound Forge 7.0, ad una frequenza di campionamento di 44.100Hz.

Il corpus di parlato raccolto è stato oggetto di valutazione percettiva da parte di 56 ascoltatori campani, diplomati e laureati, di età compresa tra i 20 e i 50 anni, con e senza esperienza dell'accento straniero dei cinesi. Il protocollo sperimentale prevedeva che ogni ascoltatore, prima dell'inizio del test, compilasse una scheda sociolinguistica in cui indicare i propri dati anagrafici, le lingue straniere conosciute, la familiarità con l'italiano pronunciato dai cinesi. Il test percettivo consisteva quindi nell'ascoltare le sequenze lette dai singoli speaker e nel formulare poi, per ciascun locutore, un giudizio di accento straniero su una scala di crescente intensità a quattro punti (0= parlante nativo; 1= accento lieve; 2= accento forte; 3= accento molto forte).

Il corpus di italiano L1 ed L2 è stato esaminato per singole catene foniche. Di ciascuna catena fonica è stato misurato il numero di sillabe realmente prodotte, la loro durata, i valori di f_0 max e f_0 min per catena fonica, la presenza di disfluenze, la durata delle pause silenti successive alla catena fonica. Sulla base delle misure effettuate sono stati calcolati i seguenti indici ritmico-prosodici:

- velocità di articolazione (VdA), intesa come il rapporto fra il numero delle sillabe e il tempo in cui esse sono state prodotte (sill/s);
- velocità di eloquio (VdE) misurata come il rapporto fra il numero delle sillabe e il tempo totale, comprensivo di pause vuote e piene (sill/s);

- fluenza (F) calcolata come il rapporto fra il numero delle sillabe e il numero delle catene foniche (sill/CF);
- range tonale (RT) espresso in semitoni;
- le componenti di silenzio, disfluenza e fonazione calcolate in misura percentuale rispetto alla durata complessiva del testo;
- il numero e la durata media delle pause silenti (PS).

Le analisi del segnale vocale sono state condotte con l'ausilio del software open source *Wavesurfer* 1.8.8. I dati ottenuti dalle analisi spettro-acustiche sono stati poi messi in relazione con quelli emersi dal test percettivo, al fine di indagare l'esistenza di una correlazione tra parametri soprasegmentali e grado di accento straniero.

3.4. Risultati del test percettivo

I risultati del test percettivo riferiscono della puntualità con cui gli ascoltatori italiani hanno riconosciuto i loro connazionali. I due madrelingua infatti sono stati giudicati come nativi nel 96% dei casi; solo il 4% (pari a due ascoltatori) si è astenuto dal rispondere. Passando ai giudizi espressi sui i cinesi, l'unanimità di pareri rilevata per i madrelingua cede il passo ad una maggiore incertezza nel valutare il grado di accento straniero (figura 1).

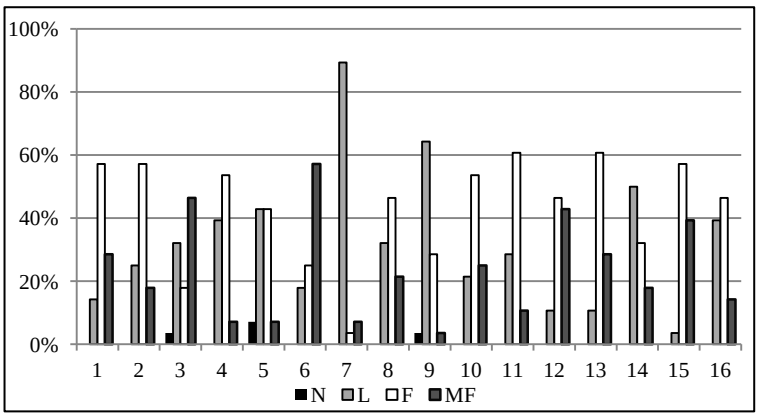


Figura 1: Risultati del test percettivo sui parlanti cinesi
(N= nativo, L=accento straniero lieve; F= accento forte; MF=accento molto forte).

Ad eccezione degli speaker 7 e 9, il cui parlato è stato giudicato come connotato da accento straniero lieve rispettivamente dall'89 e dal 64% degli ascoltatori, i giudizi espressi sugli altri cinesi non sono stati altrettanto netti. Il parlante 6 è l'unico a poter rappresentare la categoria "accento straniero molto forte". La percentuale di ascoltatori che ha fornito tale giudizio (57%), infatti, è il doppio di quanti lo hanno giudicato "forte" (25%) e addirittura supera di tre volte quella degli ascoltatori che hanno risposto "accento lieve" (18%). Solo l'accento di cinque cinesi (1, 2, 10, 11, 13) è stato considerato "forte" da più del 50% degli ascoltatori, con percentuali di scarto di circa il 30% rispetto all'alternativa "molto forte" e dal 30 al 40% rispetto a quella "lieve".

In questo studio i dati degli speaker 4, 5, 8, 12, 14, 15, 16 non sono stati oggetto di analisi per l'eccessiva disomogeneità dei risultati ottenuti. Per alcuni informanti infatti nessuno dei quattro giudizi (N, L, F, MF) aveva raggiunto il 50% delle preferenze (fig. 2).

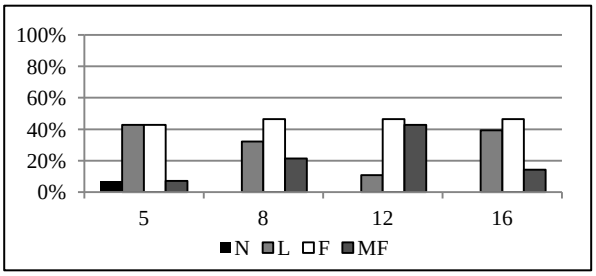


Figura 2: Risultati del test percettivo dei parlanti cinesi 5, 8, 12, 16.

Gli speaker 4, 14 e 15 sono stati esclusi dallo studio per l'esiguo scarto rilevato tra le quattro valutazioni (fig. 3).

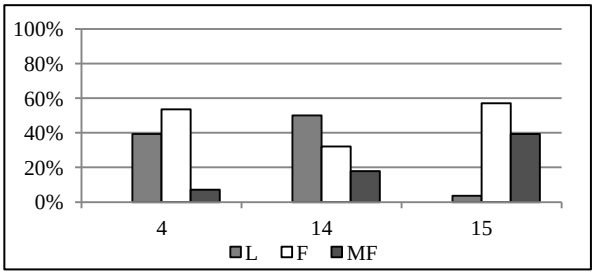


Figura 3: Risultati del test percettivo dei parlanti 4, 14, 15.

3.5. Correlati soprasegmentali del grado di accento straniero

Al fine di accertare l'esistenza di una correlazione tra parametri soprasegmentali e grado di accento straniero, i risultati del test percettivo sono stati comparati con quelli emersi dalle analisi spettro-acustiche. Ad un primo confronto tra i due tipi di dati è emerso che è stato giudicato "nativo" l'eloquio più veloce, più fluente e dal range tonale più ampio (fig. 4), (tab.1).

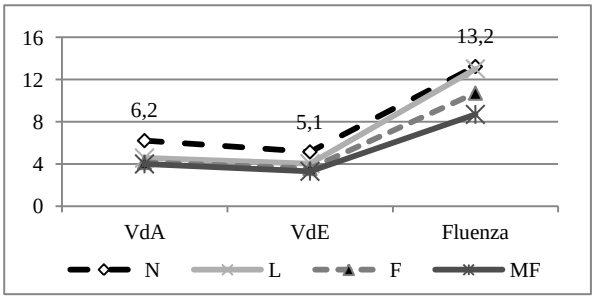


Figura 4: Valori medi di VdA, VdE e fluenza dei parlanti nativi e dei cinesi per grado di accento straniero.

Nativo	*Lieve*	*Forte*	*Molto forte*
9,5	8,2	7,7	5,8

Tabella 1: Valori medi del range tonale dei parlanti nativi e dei cinesi per grado di accento straniero.

Se si procede invece a ritroso, ossia si considerano i parametri che tendono a differenziare il parlato connotato da accento straniero molto forte, da quello considerato forte e lieve (tab. 2) è possibile annoverare da un lato un aumento un progressivo del range tonale e della fluenza; dall'altro una diminuzione della quantità di pause vuote.

	Fluenza	*Range Tonale*	*Nr. silenzi*
Molto forte	8,7	5,8	12
Forte	10,7	7,7	10
Lieve	13	8,2	9

Tabella 2: Valori medi della fluenza, del range tonale e quantità di silenzi per grado di accento straniero.

Per quanto riguarda la composizione del parlato, nel passaggio da un grado di accento straniero all'altro, si assiste ad un incremento graduale della percentuale del tempo fonatorio a scapito dei silenzi e delle disfluenze. Va notato pure che il parlato degli speaker giudicati lievi è composto solo da tempo fonatorio e silenzi, così come accade anche per i nativi (fig. 5).

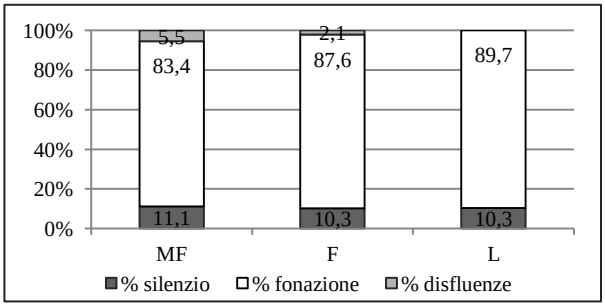


Figura 5: Composizione del parlato per grado di accento straniero.

Nessuna variazione significativa, invece, si attesta nei valori della velocità di articolazione e di eloquio e nella durata media dei silenzi (tab. 3).

	VdA (sill/s)	*VdE (sill/s)*	*Durata media silenzi (s.)*
Molto forte	4	3,3	0,3
Forte	4,1	3,6	0,3
Lieve	4,6	4	0,3

Tabella 3: Valori medi di VdA, VdE e durata dei silenzi per grado di accento straniero

A partire da questi dati, in linea generale, si può affermare che i parametri più rilevanti a differenziare i tre gradi di accento straniero sono: la percentuale di parlato, la quantità dei silenzi, la fluenza e il range tonale. Meno significative le differenze nei valori della velocità di articolazione e di eloquio. La durata media delle pause silenti resta invece invariata.

4. SECONDA SEZIONE: TEMPI DI REAZIONE DEGLI ASCOLTATORI ALL'ACCENTO STRANIERO

Nella seconda fase della ricerca l'obiettivo è stato duplice: innanzitutto stabilire il tempo necessario al nativo per definire la provenienza del proprio interlocutore. In seconda battuta, si è puntata ad accertare l'esistenza di un legame tra l'intensità dell'accento straniero e i secondi impiegati dagli ascoltatori a formulare il giudizio di forestierismo.

Per raggiungere tali scopi, è stato somministrato un secondo test percettivo a 20 ascoltatori campani, di età compresa tra i 25 e i 40 anni, tutti privi di esposizione all'italiano pronunciato dai cinesi. Il materiale del test consisteva nelle sequenze lette dai nativi e dai sinofoni che nel test precedente avevano conseguito un giudizio netto (1, 2, 6,7,9,10,11,13) Al primo ascolto, il somministratore chiedeva di selezionare la provenienza dello speaker tra due alternative, nativo/ non nativo. Le performance dei parlanti giudicati stranieri venivano fatte riascoltare una seconda volta per definire il grado di accento straniero. In entrambe le fasi del test, il somministratore appuntava sulla scheda appositamente predisposta i tempi impiegati dagli ascoltatori a dare le due risposte.

Dai dati di questa seconda indagine è emerso che sono sufficienti pochi secondi, in media cinque, per determinare la provenienza del parlante. Occorre invece più tempo per definire l'intensità dell'accento straniero (tab.4).

	Tempi di reazione
Lieve	13,7s
Forte	11,3s
Molto forte	14,5s

Tabella 4: Tempi di reazione al grado di accento straniero.

Tra i tre gradi, in media si giudicano più velocemente, dopo circa 11 secondi, i parlanti con accento straniero forte. Tempi più lunghi sono necessari per formulare i giudizi più estremi: circa 13,7 secondi per i "lievi" e 14,5 secondi per il "molto forte".

Sono state effettuate le analisi spettro-acustiche dei primi 5 secondi di parlato dei nativi e dei cinesi. Per i parlanti stranieri sono state esaminate anche le porzioni corrispondenti ai tre giudizi (12 secondi per i forti, 14 per i lievi e 15 per il molto forte). Per la discussione dei dati relativi ai primi 5 secondi, si rimanda alle conclusioni. Per quanto riguarda invece i dati spettro-acustici del corpus relativo ai tempi di reazione al grado di accento straniero, nessuna variazione significativa è stata riscontrata con i valori degli indici ritmico-prosodici emersi nella prima parte dello studio.

5. CONCLUSIONI

La ricerca si è posta un duplice obiettivo: valutare il ruolo dei correlati soprasegmentali nella percezione dell'accento straniero e definire il tempo necessario al nativo per stabi-

lire la provenienza del proprio interlocutore. Dai risultati dell'indagine è emerso che per l'ascoltatore madrelingua riconoscere l'origine del proprio interlocutore è un compito cognitivamente meno impegnativo rispetto che valutare qualitativamente le sue performance. Sono sufficienti infatti dai 4 ai 6 secondi per determinare l'appartenenza di un parlante alla propria comunità linguistica; ne servono oltre 10 invece per pronunciarsi sul grado di accento straniero.

Per quanto riguarda i parametri prosodici, il confronto fra i dati delle analisi spettroacustiche e i risultati dei test percettivi ha consentito di ricavare due tipi di informazioni, l'una più generale che informa dell'importanza degli indici ritmico-prosodici nella percezione dell'accento straniero; l'altra più specifica che circoscrive con maggiore perizia il ruolo dei singoli parametri nel determinare la provenienza e/o il grado di accento straniero. Tra gli indici analizzati, infatti, alcuni sono risultati discriminanti per ricavare entrambe le informazioni; altri invece sono stati significativi o solo per la provenienza o solo per il grado di devianza.

Il range tonale ad esempio è risultato influente per differenziare tanto l'appartenenza di un parlante alla comunità linguistica dei nativi, quanto il grado di accento straniero. Come mostrano i dati nelle figg. 6-7, sia dopo i primi cinque secondi, sia per l'intera durata della task di parlato letto il range tonale dei nativi era sempre più ampio di quello degli stranieri. Tra i cinesi inoltre i valori decrescevano all'intensificarsi dell'accento straniero.

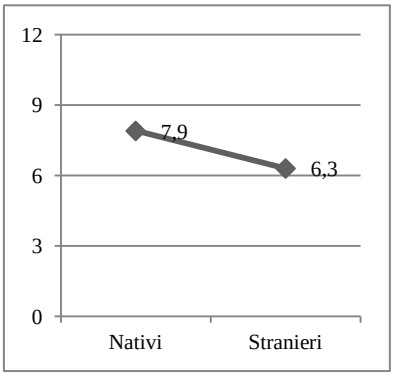


Figura 6: Valori medi del Range Tonale dopo 5''

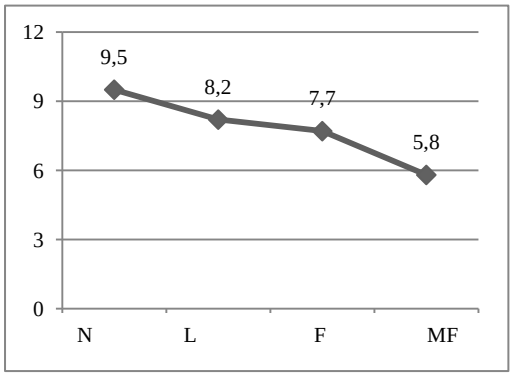


Figura 7: Valori medi complessivi del Range Tonale

La velocità di articolazione e di eloquio invece sono state distintive ai fini della provenienza, cioè hanno consentito all'ascoltatore di contraddistinguere il nativo dal cinese dopo i primi cinque secondi. Non hanno avuto effetti invece sulla percezione del grado di accento straniero: all'intensificarsi dell'intensità della devianza infatti non è corrisposta una netta diminuzione nei valori dei due parametri in esame (figg.8-9).

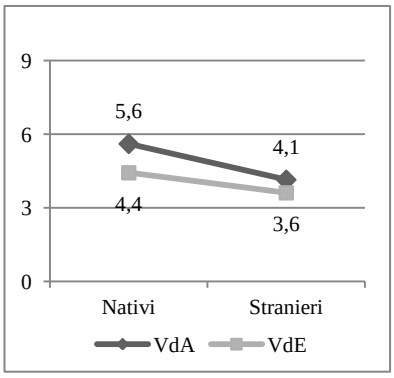


Figura 8: Valori medi di Vda e VdE dopo 5''

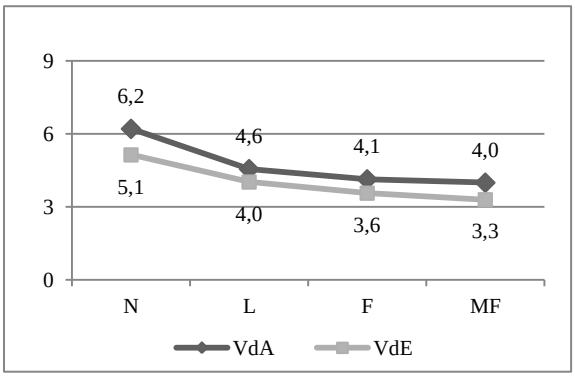


Figura 9: Valori medi complessivi di VdA e VdE

A differenza della VdA e della VdE, la fluenza è discriminante per il giudizio di accento straniero, ma non per la provenienza (figg. 10-11).

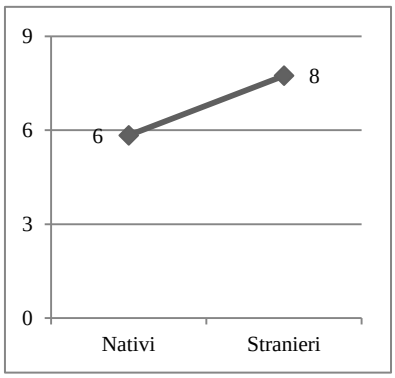


Figura 10: Valori medi di fluenza dopo 5''

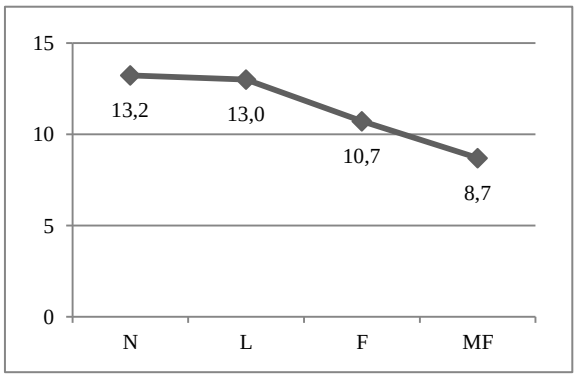


Figura 11:Valori medi complessivi di fluenza

Nei primi cinque secondi, infatti, gli italiani sono stati meno fluenti dei cinesi; ciononostante la loro appartenenza alla comunità linguistica dei nativi non è stata messa in discussione.

I dati ottenuti da questo studio avvalorano l'importanza degli indici ritmico-prosodici nella percezione dell'accento straniero. Essi spingono inoltre a proseguire le ricerche su apprendenti con lingue materne distanti dall'italiano, per verificare se i correlati soprasegmentali dell'accento straniero siano specifici per L1 o tipici del parlato in una L2.

Un ulteriore step della ricerca sarà l'analisi segmentale del corpus per determinare l'effettivo peso giocato dalla componente segmentale e da quella soprasegmentale nella percezione dell'accento straniero.

BIBLIOGRAFIA

Best, C.T. (1995), A direct realistic view of cross-language speech perception, in Speech perception and linguistic experience (W. Strange, editor), Baltimore (MD): York Press, 171-206.

Birdsong, D. (2006), Age and second language acquisition and processing: A selective overview, Language Learning, 56, 9-49.

Boula de Mareüil, P., Marotta, G., Adda-Decker, M. (2004), Contribution of prosody to the perception of Spanish/Italian accents, in Speech Prosody (B. Bel & I. Marlien, editors), Nara (Giappone), 681-684.

Boula de Mareüil, P., Vieru-Dimulescu, B. (2006), "The Contribution of Prosody to the Perception of Foreign Accent", Phonetica, Int. Journal of Phonetic Science, Vol.63, no 4, December, 247-267.

De Meo, A., Pettorino, M. (2011), L'acquisizione della competenza prosodica in italiano L2 da parte di studenti sinofoni, in Atti del XV seminario AICLU. La didattica dell'italiano a studenti cinesi e il progetto Marco Polo (E. Bonvino, S. Rastelli, editors) Pavia: Pavia University Press, 67-78.

De Meo, A., Pettorino, M. (2012), Prosodia e italiano L2: cinesi, giapponesi e vietnamiti a confronto", in Apprendere l'italiano da lingue lontane: prospettiva linguistica, pragmatica, educativa (R. Bozzone Costa, L. Fumagalli, A. Valentini, editors), Perugia: Guerra, 59-72.

De Meo, A., Pettorino, M., Vitale, M (2012), Non ti credo: i correlati acustici della credibilità in italiano L2, in Atti dell'XI Congresso dell'Associazione Italiana di Linguistica Applicata. Competenze e formazione linguistiche. In memoria di Monica Berretta (G. Bernini, C. Lavinio, A. Valentini, M. Voghera, editors), Perugia: Guerra Edizioni, 229-248.

Flege, J.E. (1991), Age of learning affects the authenticity of voice-onset time (VOT) in stop consonants produced in a second language, Journal of the Acoustical Society of America, Vol. 89, no.1, 395–411.

Flege, J.E., Munro, M.J., MacKay, I.R.A. (1995), Effects of age of second-language learning on the production of English consonants, Speech Communication, 16, 1–26.

Flege, J.E., Bohn, O.S., Jang, S. (1997), Effects of experience on non-native speaker's production and perception of English vowels, Journal of Phonetics, 25, 437–470.

Grover, C., Jamieson, D.G., Dobrovolsky, M.B. (1987), Intonation in English, French and German perception and production, Language and Speech, 30, 277–296.

Horgues, C. (2005), Contribution à l'étude de l'accent français en anglais. Quelques caractéristiques prosodiques de l'anglais parlé par des apprenants francophones et leur évaluation perceptive par des juges natifs, in Actes des VIIIèmes RJC ED268 'Langage et langues', Paris III, 21 mai 2005, 79–83.

Jilka, M. (2000), The contribution of intonation to the perception of foreign accent; PhD Thesis, University of Stuttgart, Stuttgart.

Kuhl, P.K. (1991), Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not, Perception & Psychophysics, 50, 93–107.

Magen, H.S. (1998), The perception of foreign-accented speech, Journal of Phonetics, 26, 381–400.

Major, R.C. (2001), Foreign Accent: The Ontogeny and Philogeny of Second Language Phonology, Mahwah, NJ: Erlbaum.

Marotta, G. (2008), Sulla percezione dell'accento straniero, in Diachronica et synchronica. Studi in onore di Anna Giacalone Ramat (R. Lazzeroni, E. Banfi, G. Bernini, M. Chini, G. Marotta, editors), Pisa: ETS, 327-347.

Marotta, G., Boula de Mareuil, P. (2010), Persistenza dell'accento straniero. Uno studio percettivo dell'italiano L2, in Atti della V conferenza dell'associazione italiana di scienze della voce, La dimensione temporale del parlato (S. Schmid, M. Schwarzenbach, D. Studer), 475-494.

Munro, M.J. (1995), Nonsegmental factors in foreign accent: Ratings of filtered speech, Studies in Second Language Acquisition, 17, 17–34 .

Pallier, C., Bosch, L., Sebastían-Gallés, N. (1997), A limit on behavioral plasticity in speech perception, Cognition, 64, 9-17.

Perry, G.M.J., Harris, C.L. (2002), Linguistically Distinct Sensitive Periods for Second Language Acquisition, in Proceedings of the 26th annual Boston University Conference on Language Development (B. Skarabela, S. Fish, A.H.J. Do, editors) Sommerville, MA: Cascadilla Press, 557-566.

Pettorino, M., De Meo, A., Pellegrino, E., Salvati, L., Vitale, M. (2011), Accento straniero e affidabilità del messaggio. Un'indagine acustico-percettiva, in Atti dell'XI Convegno AISV, Contesto comunicativo e variabilità nella produzione e percezione della lingua, (B. Gili Fivela, A. Stella, L. Garrapa, M. Grimaldi, editors) Lecce, Gennaio 2011, Roma: Bulzoni editore.

Piat, M., Fohr, D., Illina, I. (2008), Foreign accent identification based on prosodic parameters", INTERSPEECH- 2008, 759-762.

Piske, T., MacKay, I.R.A., Flege, J.E (2001), Factors affecting degree of foreign accent in an L2: a review, Journal of Phonetics, 29, 191–215.

Ramus, F., Mehler, J. (1999), Language Identification With Suprasegmental Cues: A Study Based on Speech Resynthesis, Vol. 105, no. 1, 512–521.

Tsukada, K. (2005), Cross-language speech perception of final stops by Australian-English, Japanese and Thai Listeners, in Proceedings of the ISCA Workshop on Plasticity in Speech Perception, London, 244–247.

Tsukada, K., Birdsong, D., Mack, M., Sung, H., Bialystok, E., Flege, J.E. (2004), Release bursts in English word-final voiceless stops produced by native English and Korean adults and children, Phonetica, 61, 67–83.

Walley, A.C., Flege, J.E. (1998), Effect of lexical status on children's and adults' perception of native and non-native vowels, Journal of Phonetics, 27, 307–332.

ANALISI PERCETTIVA, MUSICALE E "AUTOMATICA" DELL'ITALIANO L1 E L2

[*]Luciano Romito, **Renata Savy, **Andrea Tarasi, *Rosita Lio
*Università della Calabria, **Università degli Studi di Salerno
luciano.romito@unical.it, rsavy@unisa.it, and.tarasi@gmail.com, lio.rosita@libero.it

1. L'ITALIANO COME LINGUA SECONDA

La seconda lingua è una lingua appresa nel paese dove essa viene abitualmente parlata, come l'italiano appreso in Italia dagli immigrati. In Italia le varietà non native che vengono acquisite naturalmente in un contesto di L2 occupano un posto di interesse. All'interno di un processo di acquisizione spontanea il contatto con la seconda lingua avviene esclusivamente in situazioni comunicative naturali. La persona che apprende la lingua ha un controllo migliore sulla pragmatica della lingua, mentre la competenza linguistica può essere meno estesa e in alcuni casi ferma ad un livello elementare.

Rispetto a quello che succede nel percorso di acquisizione guidata di una L2 e in modo più prorompente per coloro che apprendono una lingua straniera, chi acquisisce naturalmente la L2 è esposto continuamente a numerosi stimoli di tipo differente che comportano una selezione autonoma delle informazioni linguistiche utili alla comunicazione. La scoperta dell'autonomia del filtraggio delle informazioni linguistiche, da parte del parlante, e l'interazione con i parlanti nativi permettono di testare ipotesi sul funzionamento della L2, in quanto rappresentano la base di un uso attivo e creativo della lingua.

Nella varietà parlata da immigrati, la comunicazione interculturale è caratterizzata da una distanza linguistica e culturale che porta gli stessi ad assumere due strategie differenti: la riduzione di queste distanze attraverso strategie di convergenza oppure accentuare come espressione di rifiuto delle norme culturali dell'altro in difesa della propria identità etnica la distanza tra le due lingue. In questo caso si parla di enclosure, cioè di chiusura e distanza sociale rispetto ai gruppi di parlanti nativi e da questo fattore dipende, probabilmente, la persistenza di forme pidginizzate nella produzione di parlanti immigrati che vivono in un paese da diversi anni.

Sull'acquisizione dell'italiano come L2 esistono numerosi studi che hanno permesso di individuare dei percorsi di acquisizione comuni. Le aree di maggiore interesse sono quelle della temporalità, della modalità e del genere (Pallotti, 2003).

La temporalità gioca un ruolo fondamentale nell'apprensione della lingua seconda e per l'italiano sono state individuate le seguenti sequenze:

1. nella prima fase non esistono mezzi morfologici usati in maniera produttiva per esprimere la temporalità. Gli apprendenti usano una forma unica del verbo, che in genere corrisponde alla radice verbale e solo in alcuni casi compare l'infinito. Questa condizione si verifica in situazioni molto particolari, con apprendenti di madrelingua isolante, in quanto sono esposti ad un input ridotto e/o distorto dal foreign talker. Negli altri casi, all'interno del sistema l'infinito è usato in un ambito d'uso ristretto ad esempio, è utilizzato per esprimere nozioni non fattuali (ipotetiche, futuri);

2. nella seconda fase la prima opposizione morfologica che compare è quella tra azioni concluse ed azioni presenti. In genere, la marca usata per esprimere le azoni passate è il suffisso –to del participio passato. È stato notato che il participio passato non è usato solo per indicare una azione avvenuta prima del momento di enunciazione ma anche per indicare un evento concluso;
3. ad un determinato momento gli apprendisti avvertono la necessità di esprimere morfologicamente e in modo puntuale la distinzione tra eventi passati ed eventi a carattere durativo. Questa funzione viene assorta dall'imperfetto. L'uso dell'imperfetto è tipicamente associato ad enunciati che esprimono lo sfondo di una narrazione;
4. compare il futuro ed il condizionale seguiti dal congiuntivo. La principale distinzione introdotta a questo livello è quella tra fattualità e non fattualità. Questo livello viene raggiunto da parlanti colti e che hanno avuto qualche tipo di istruzione esplicita. In genere, l'uso del futuro è accompagnato da esitazioni, autocorrezioni che vengono riscontrate anche nei tentativi di uso del condizionale e del congiuntivo.

L'uso del condizionale e del congiuntivo comportano lo studio della modalità, in quanto non sono dei tempi verbali ma dei modi. Essi evidenziano il modo in cui il parlante si rapporta a ciò che dice secondo una modalità di desiderio, di dubbio ecc. La modalità non riguarda solo il modo verbale ma include anche il modo in cui il parlante si pone verso alcune proposizioni, modalità che riguardano obblighi, libertà e volontà. Nel processo di acquisizione di una lingua esse vengono apprese attraverso sequenze regolari in cui si possono individuare tre stadi: modalità implicita, modalità lessicale e modalità grammaticale.

Nella modalità implicita, durante le prime fasi di apprendimento non esistono mezzi per esprimere in maniera esplicita le nozioni modali. Gli apprendenti ricorrono a segnali non verbali come gesti, intonazione, espressione del volto. L'intonazione segnala spesso esitazioni, dubbi apparendo quindi all'interlocutore come una richiesta di aiuto. In questo caso, la ricostruzione dei contenuti modali viene affidata alle capacità interpretative del parlante nativo che crea delle ipotesi basandosi sul contesto generale della osservazione.

Nella modalità lessicale gli apprendenti acquisiscono i primi mezzi linguistici per codificare la modalità. In questa fase viene usata la strategia dell'interlingua di base, che consiste nell'utilizzare elementi lessicali piuttosto che grammaticali. Nella modalità epistemica compaiono le forme verbali fisse (credo, penso) e successivamente vengono usati degli avverbi (magari, forse). Questi modificatori esprimono una modalità di incertezza, di dubbio mentre i modalizzatori che codificano la certezza compaiono in una fase successiva. Nella modalità deontica e dinamica, invece, sono usati verbi appositi (dovere, potere, volere).

Nella modalità grammaticale, i mezzi grammaticali usati per codificare la modalità appaiono molto tardi, infatti, le forme specifiche rappresentate dal condizionale e dal congiuntivo sono acquisite in maniera produttiva solo dagli apprendenti avanzati. Alcuni condizionali emergono presto ma sono forme apprese in modo non analizzato (sarebbe, vorrei). Il congiuntivo viene padroneggiato per ultimo e questo avviene per due ragioni: da un lato implica un paradigma verbale complesso e

dall'altro comprende due funzioni difficili da distinguere nell'input, quella di esprimere una dipendenza sintattica e quella di indicare la modalità ipotetica.

Per quanto riguarda l'imperativo, in quasi tutte le lingue del mondo, viene codificato da una forma molto basica del verbo che spesso equivale alla semplice radice verbale. Dal punto di vista comunicativo, esso svolge un ruolo fondamentale poiché consiste nel richiedere all'interlocutore di fare qualcosa.

Il genere rappresenta l'ultima area di indagine sulle sequenze evolutive in italiano. Esso corrisponde ad una forma di classificazione dei nomi in categorie (maschile e femminile). Gli apprendenti assimilano questa caratteristica in modo graduale. All'inizio gli apprendenti non notano questa categoria in quanto acquisiscono gli item lessicali come unità non analizzate. Solitamente, la terminazione appresa è quella giusta infatti, raramente compaiono errori come uomo→uoma. L'opposizione lui/lei viene assimilata velocemente dagli apprendisti. Superata questa fase, essi iniziano a combinare le parole tra di loro in sintagmi ed enunciati, iniziando a porsi il problema dell'accordo. All'inizio questa strategia si fonda sull'assonanza che spesso da risultati corretti (la scuola), ma a volte queste strategie portano alla formazione di sintagmi fuorvianti come 'la cinema'. Successivamente, l'accordo si estende e coinvolge l'aggettivo attributivo come 'acqua calda'. Accanto a queste forme possono comparirne altre come 'grande case' che sono orientate verso la lingua di arrivo. Dopo di ché l'accordo viene esteso anche agli aggettivi predicativi (la casa è piccola). Infine, solo negli apprendisti avanzati, viene acquisito l'accordo tra nome e participio passato (siamo andati, siamo arrivati).

2. MATERIALI E METODI

In questo lavoro presentiamo le analisi condotte sul parlato letto prodotto da parlanti italiano L1 e L2. Questi segnali vocali fanno parte del corpus Pangea che contiene codici linguistici di italiano L1 e L2 ed è composto da 25 parlanti con sesso, grado scolastico ed età differente. Inoltre, tutti vivono in Calabria almeno da due anni. Le lingue considerate in questo lavoro sono: albanese (3M e 2F età 20-30), cinese (3M e 1F età 20-30) italiano (4M età 35-70), polacco (2M e 3F età 20-55) e romeno (5F età 20-50). Le lingue L1 sono tipologicamente diverse tra di loro e i parlanti hanno letto una lista di frasi e prodotto anche almeno tre minuti di parlato spontaneo (Romito et alii, 2012, in press) in cui il topic è stato scelto dagli stessi parlanti così da ottenere un parlato più spontaneo possibile.

Nelle prove percettive il campione di ascolto è composto da 58 soggetti, tutti calabresi, con età compresa tra i 20 e i 65 anni con grado di istruzione e sesso differente. Al campione è stato somministrato un test percettivo composto da sei frasi lette da parlanti italiano L1 (italiano calabrese e italiano piemontese) e parlanti stranieri italiano L2 (albanese, cinese, polacco e rumeno). Agli ascoltatori è stato chiesto di discriminare l'italiano L1 dall'italiano L2 e, per non influenzare la risposta, le domande poste sono generiche (*la lingua che ascolta è italiana oppure straniera? Perché è straniera/italiana?)*. Nella fase di risposta una continua interazione tra ascoltatore e tester ha permesso di individuare in modo specifico le caratteristiche utilizzate dagli ascoltatori per discriminare le lingue. In un lavoro precedente (Romito et alii, 2011, in press) queste frasi sono state analizzate attraverso il paradigma della fono-

logia metrica allo scopo di individuarne le differenze sintattiche e intonative presenti negli enunciati. Il test è stato condotto all'interno della Camera Silente 2*2 Amplifon presso il Laboratorio di Fonetica dell'Università della Calabria utilizzando un portatile Toshiba con scheda audio Realtek HD e una cuffia Sennheiser. I tratti caratterizzanti individuati dagli ascoltatori sono stati considerati sia sull'intero campione di frasi, così da ottenere una descrizione generale dell'influenza dei livelli sulla percezione del linguaggio, che su ogni singola frase allo scopo di conseguire risultati più specifici. I dati sono stati organizzati in livello segmentale, livello soprasegmentale, livello segmentale/prosodico[1], livello morfologico, livello fonologico/morfologico e livello morfologico/prosodico.

Per le analisi musicali è stato necessario portare i segnali vocali su un pentagramma[2] e attraverso questa procedura sono state analizzate le relazioni esistenti tra musica e lingua.

Infine, è stata testata l'abilità di un sistema di Speech Recognition nel riconoscimento dell'italiano L1/L2 così da capire quali informazioni (livelli e tratti) tale sistema utilizza per discriminare le lingue. Il sistema di SR testato è Trascrivi.it sviluppato da Cedat85[3].

3. RISULTATI PROVE PERCETTIVE

I dati relativi alle prove percettive prodotte sul parlato letto richiedono un'attenta riflessione:

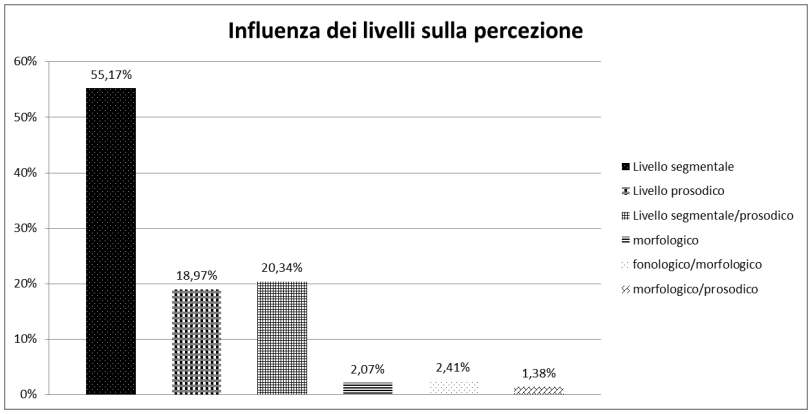


Figura 1.

Questa Figura mostra i risultati generali che descrivono il grado di influenza dei livelli linguistici sulla percezione delle differenze individuate nei segnali vocali da parte degli ascoltatori. Il livello segmentale, da questi dati, risulta il più rilevante così gli ascoltatori sono attenti alle variazioni subite dai foni e ai processi fonologici

[1] '/' indica l'interazione tra due livelli, una frase è stata discriminata in base a caratteristiche che appartengono a due livelli differenti.

[2] L'analisi musicale è stata prodotta dal musicista Leonzio Gobbi.

[3] È un sistema di trascrizione automatica commerciale con riconoscimento vocale che si basa su modelli generici per l'italiano standard senza specifico addestramento.

che intervengono all'interno delle frasi. Inoltre, l'importanza del livello segmentale è dimostrata anche dal fatto che questo livello interagisce spesso insieme ad altri nella distinzione tra italiano L1 e italiano L2 (segmentale/prosodico, fonologico/morfologico). I tratti segmentali individuati dagli ascoltatori per differenziare le lingue sono RFS (italiano meridionale), scempiamento ed errori nella produzione di alcuni foni e data l'omogeneità delle risposte non sono stati realizzati grafici che riassumessero tali differenze, mentre una maggiore variabilità si rileva nel livello prosodico:

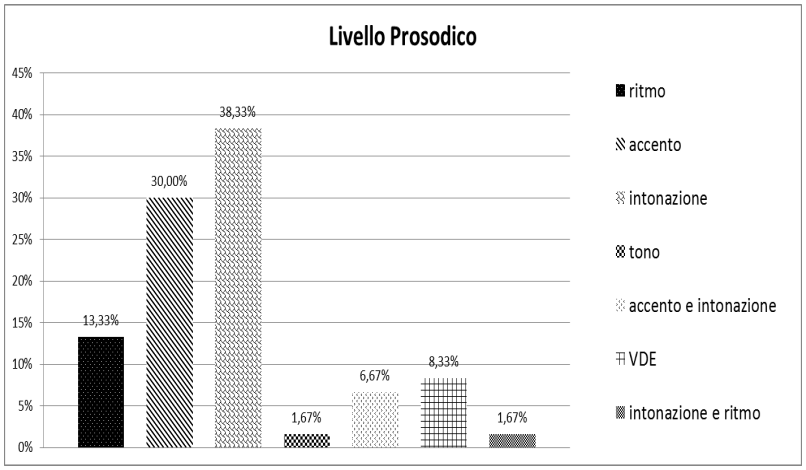


Figura 2.

Come si può notare dalla Figura, l'intonazione e l'accento sono i tratti soprasegmentali prominenti che consentono agli ascoltatori di discriminare l'italiano L1 dall'italiano L2. Le variazioni intonative e di accento sono immediatamente individuate dagli ascoltatori ma un dato rilevante riguarda soprattutto l'intonazione. Spesso gli ascoltatori hanno sottolineato differenze nell'andamento melodico delle frasi come se avessero una competenza fonologica del modello intonativo dell'italiano L1 riconoscendo la melodia prodotta da alcuni parlanti come diversa rispetto a quella che caratterizza un parlante italiano nativo, in questo caso calabrese.
Per interpretare in maniera più specifica i dati e capire l'effettiva influenza dei livelli linguistici sulla percezione sono state analizzate le risposte del campione su ogni singola frase prodotta dai parlanti delle diverse lingue considerate in questa ricerca:

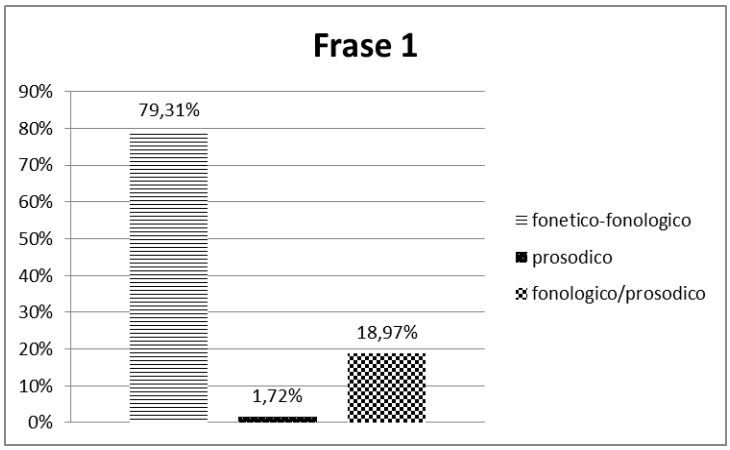


Figura 3: albanese L1.

In questo caso la frase prodotta dal parlante albanese viene riconosciuta come italiano L2 in modo netto attraverso il livello fonetico-fonologico. Gli ascoltatori individuano lo scempiamento delle consonanti geminate e la produzione di alcuni foni come [r] e [l] diversi dall'italiano L1. Questi fenomeni denotano il trasferimento di alcune caratteristiche dalla propria lingua prima all'italiano L2. La mancata produzione delle consonanti geminate è giustificata dal fatto che sono assenti nell'inventario fonologico dell'albanese, mentre il fono [l] è pronunciato come laterale alveolare e il fono [r] come monovibrante post-alveolare cioè con luoghi di articolazione propri dell'albanese L1.
Un discorso diverso riguarda invece il riconoscimento del cinese italiano L2 di cui riportiamo i dati:

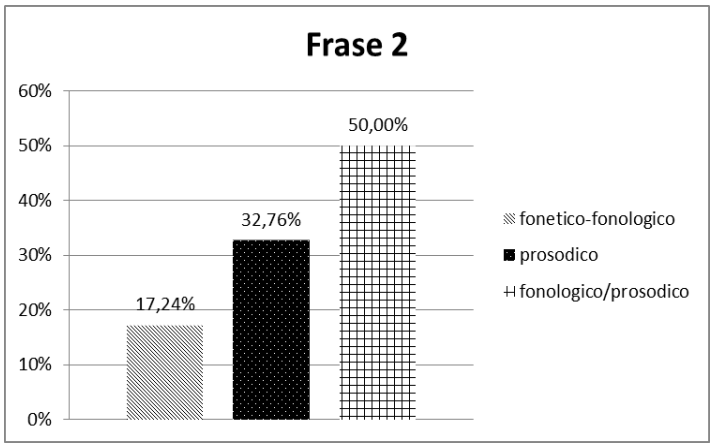


Figura 4: cinese L1.

In questo caso l'italiano L2 prodotto dal cinese viene riconosciuto ricorrendo sia a fenomeni fonetico-fonologici come la degeminazione delle consonanti geminate e la desonorizzazione delle consonanti sonore che comporta la realizzazione diversa di alcuni foni come [b] (es. *pase*, 'base'), che al livello prosodico data la diversa disposizione degli accenti quindi il diverso andamento melodico. Tutto ciò dipende dal fatto che il cinese è una lingua tonale ed è tipologicamente distante rispetto alle altre

lingue considerate nel corpus. Anche in questo caso si può parlare di transfer linguistico in quanto sia le geminate che il fono [b] sono assenti nel loro inventario fonologico di partenza. Così, sulle geminate interviene il fenomeno della degeminazione (es. *silape* 'sillabe') mentre sulla occlusiva bilabiale sonora [b] la desonorizzazione che comporta la sostituzione di questo fono con la corrispondente occlusiva bilabiale sorda [p] presente nell'inventario fonologico del cinese. Anche l'andamento melodico risulta differente rispetto alle altre lingue considerate nella ricerca. Percettivamente l'andamento intonativo risulta 'piatto' e solo su alcune sillabe gli ascoltatori percepiscono un innalzamenti del pitch. Queste variazioni caratterizzano le lingue tonali poiché nella loro lingua prima consentono di distinguere significati.

Ora, consideriamo i dati dell'italiano meridionale (calabrese) e dell'italiano settentrionale (piemontese):

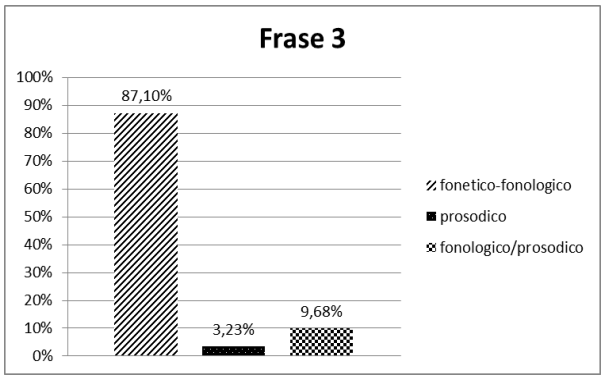


Figura 5: italiano meridionale.

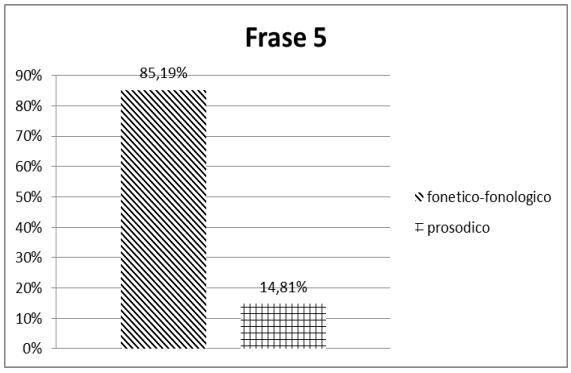


Fig. 6: italiano settentrionale.

Come possiamo notare dalla Figura 5 l'italiano meridionale è riconosciuto sostanzialmente attraverso il livello fonetico-fonologico e i tratti caratterizzanti sono il raddoppiamento fonosintattico caratteristico dei dialetti e dell'italiano meridionale e il forte VOT con il quale sono prodotte le consonanti occlusive sorde e sonore. Anche l'italiano settentrionale è riconosciuto prevalentemente ricorrendo al livello fonetico-fonologico ma si può notare un incremento del livello prosodico e la scomparsa dell'interazione tra livello fonologico e prosodico. Il tratto prosodico utilizzato

in questo caso per discriminare i due tipi di italiano è l'intonazione, gli ascoltatori hanno sottolineato che esiste una intonazione tipica del settentrione diversa da quella meridionale come se avessero anche in questo caso un modello intonativo nella loro competenza che gli consente di distinguere i due andamenti melodici. All'interno del livello fonetico-fonologico compare la sonorizzazione del fono [s] in posizione intervocalica (es. *baze* 'base') tipica dell'italiano settentrionale.

Passiamo all'analisi dei dati ottenuti sul polacco italiano L2:

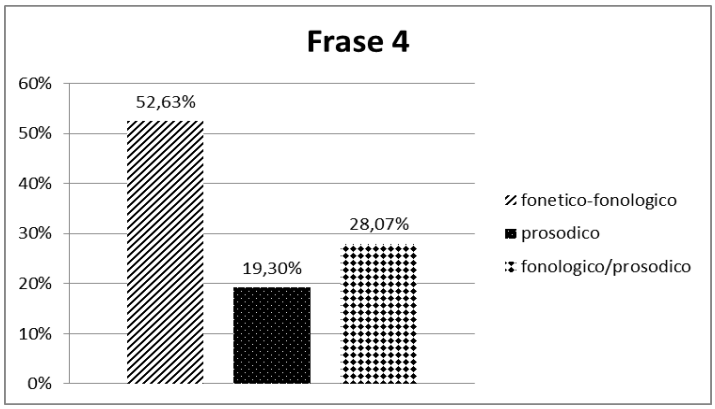


Figura 7: polacco L1

Ancora una volta il livello fonetico-fonologico risulta importante per l'identificazione della lingua. Il fenomeno fonologico riconosciuto è la degeminazione (es. *ragrupamento* 'raggruppamento') mentre dal punto di vista fonetico il nesso consonantalo 'str' è prodotto in modo difficoltoso. Questa difficoltà è dovuta all'assenza di tale nesso nell'inventario fonologico della loro L1 e lo stesso discorso vale per le consonanti geminate le quali sono assenti nell'inventario fonologico della lingua di partenza. Il tratto prosodico utilizzato nel riconoscimento da parte degli ascoltatori è ancora una volta l'intonazione che viene descritta come diversa rispetto a quella dell'italiano L1. Anche in questo caso si evince l'influenza della lingua prima sulla lingua seconda che avviene sia sul livello fonetico-fonologico sia su quello prosodico.

Infine, riportiamo i dati del rumeno italiano L2:

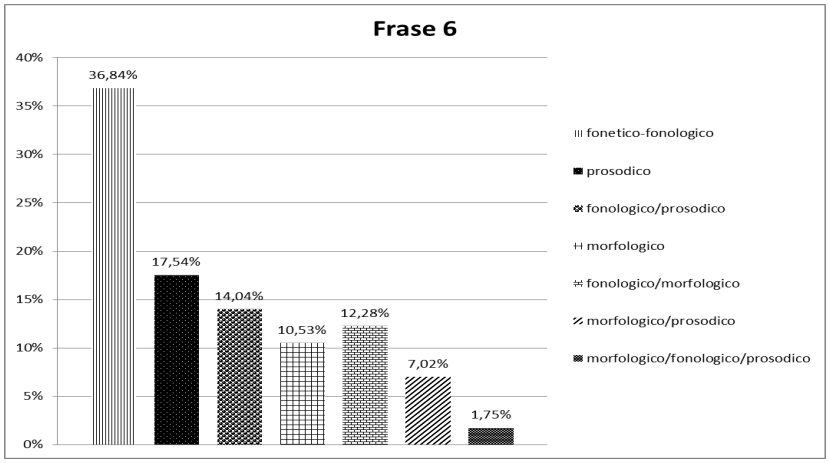


Figura 8: rumeno L1.

Come si nota dal grafico, il riconoscimento del rumeno italiano L2 da parte degli ascoltatori ha una situazione variegata e diversa rispetto a quanto visto in precedenza. Ancora una volta il livello fonetico-fonologico risulta prevalere ma una competenza diversa dell'italiano L2 rispetto agli altri parlanti ha consentito agli ascoltatori di individuare nuovi tratti che caratterizzano questo italiano L2. Un livello mai individuato in precedenza è quello morfologico che in questo caso gioca un ruolo rilevante data anche la sua interazione con altri livelli (Fig.7). Infatti, il mancato accordo tra il nome e l'aggettivo (es. *strutture più elementare*) è stato individuato dalla maggior parte degli ascoltatori. Anche in questo caso viene individuata la degeminazione delle consonanti geminate che è correlata alla loro assenza nell'inventario fonologico della lingua prima. Diversamente da quanto visto per le lingue precedenti viene individuata la cancellazione delle vocali atone finali e una pronuncia diversa delle vocali che dipende dal sistema vocalico della lingua di partenza. Il rumeno ha un sistema eptavocalico diverso da quello calabrese (pentavocalico), simile per numero di elementi a quello settentrionale ma diverso per la presenza dei fonemi /ɨ/ e /ɮ/.

4. RISULTATI DEL SISTEMA DI SR

Il sistema di trascrizione automatica con riconoscimento vocale usato in questa ricerca è Trascrivi.it di Cedat85. Di seguito riportiamo la tabella riassuntiva dei risultati ottenuti attraverso questa sperimentazione:

Frase	Le sillabe sono le strutture più elementari che stanno come base di ciascun raggruppamento di fonemi
Alb	Le sillabe sono le strutture più elementari che stanno come base di chi ascolta raggruppamento di fonemi
Cin	La sera per sono le strutture più elementari castano come passati ciascuno accorpamento che fonte fonemi
Ita mer	Le sillabe sono le strutture più elementari che stanno come base [di][4] ciascuno commento di fonemi
Pol	Nessuna persona le strutture più elementare che stanno

[4] Le parentesi quadre indicano l'assenza della proposizione semplice.

	come base di ciascun era componente del fallimento
Ita sett	La figlia ha detto alle strutture più elementari che hanno come base di ciascun raggruppamento di fonemi
Rum	Nel sillabe sono le strutture più elementari che stanno come base di ciascuna raggruppamento dei fonemi

Figura 9: trascrizione restituita dal sistema di SR.

Come possiamo notare dai risultati riportati nella Figura 8, il sistema di SR restituisce dati molto validi sul riconoscimento dell'italiano meridionale e settentrionale, dell'albanese e del rumeno italiano L2 mentre una trascrizione meno accurata è resa per il cinese e il polacco italiano L2.

Il parlante albanese italiano L2 realizza il termine 'ciascun' come '*chiascun*'[5] e questo errore induce il sistema a trascrivere 'chi ascolta'. In questo caso, il sistema di SR, cerca nella propria banca dati il termine più aderente alla produzione realizzata e lo inserisce nella risposta. Per quanto riguarda invece il rumeno italiano L2 si osservano errori poco rilevanti che dipendono dalla velocità di eloquio adottata dal parlante. Nella trascrizione dell'italiano meridionale non è riportata la preposizione semplice 'di'[6], mentre 'raggruppamento' è reso come '*commento*'. Nel primo caso l'alta VDE del parlante impedisce al sistema di individuare la preposizione, nel secondo caso il parlante calabrese realizza un forte VOT sulle consonanti occlusive che porta il sistema a commettere l'errore di trascrizione. Il riconoscimento dell'italiano settentrionale mostra problematiche simili a quelle dell'italiano meridionale e nella prima parte della frase il sistema trascrive '*la figlia ha detto alle*'. In questo caso interviene il fenomeno fonologico della sonorizzazione del fono [s] in posizione intervocalica caratteristico dell'italiano settentrionale che induce il sistema all'errore, mentre l'alta VDE porta il sistema a trascrivere '*hanno*' al posto di stanno.

La frase prodotta dal parlante cinese risulta molto problematica per il riconoscimento e la trascrizione da parte del sistema. L'output restituito presenta numerosi errori che dipendono da fenomeni fonetico-fonologi intervenuti nell'eloquio e il primo errore restituito è '*la sera per sono*'. In questo caso il cinese italiano L2 è interessato da fenomeni di degeminazione e desonorizzazione che portano il sistema a produrre una trascrizione errata. La forma realmente resa dal parlante è '*le silape sono*' con desonorizzazione del fono [b] assente nell'inventario fonologico del cinese e lo scempiamento delle geminate. Lo stesso discorso vale per gli errori '*castano come passati*' (che stanno come base) e '*ciascuno accorpamento che fonte*' (ciascun raggruppamento di fonemi). Ancora una volta in questi casi intervengono fenomeni di degeminazione e desonorizzazione che inducono il sistema all'errore. Risultati discordanti riguardano anche quelli ottenuti sul polacco italiano L2. In questo caso la

[5] L'inventario fonologico dell'albanese ha la consonante occlusiva palatale sorda /c/.

VDE molto alta e lo scempiamento delle consonanti geminate giocano un ruolo centrale nella trascrizione errata della frase.

5. RISULTATI DELLE PROVE MUSICALI

La continua ricerca di caratteristiche che consentano una migliore descrizione delle preferenze ritmiche dell'italiano L1 e L2 ha portato gli autori di questo articolo a indagare il ritmo musicale così da verificare eventuali relazioni con quello linguistico.

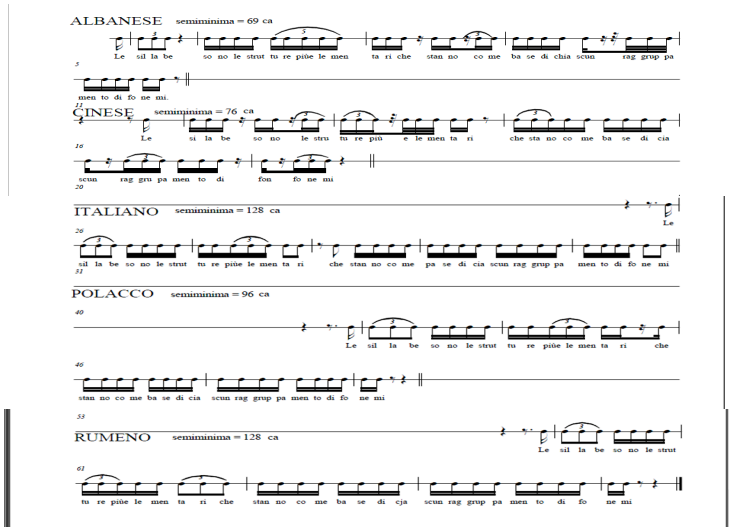


Figura 10: risultati prove musicali.

L'andamento ritmico dell'enunciato è determinato dal tempo di esecuzione che in questo caso è 2/4 e il numeratore del tempo specifica anche la tipologia degli accenti. Questi segnali vocali sono caratterizzati dalla successione tra un accento debole e uno forte che si alternano in ogni gruppo ritmico individuato ed è possibile anche conseguire due accenti forti adiacenti (Fig. 9). La nota che porta l'accento è prodotta con una maggiore intensità e la presenza di una pausa comporta lo spostamento dell'accento sulla nota successiva.

Dalle analisi musicali si evince la presenza di diversi gruppi irregolari che caratterizzano i segnali vocali e tali gruppi portano l'accento forte sempre sulla prima nota. Durante la produzione di tali gruppi si registra un incremento della velocità di produzione poiché all'interno del tempo di esecuzione deve rientrare necessariamente un numero maggiore di note.

La prima relazione che possiamo riscontrare tra lingua e musica riguarda proprio la velocità di produzione che dal punto di vista linguistico corrisponde, in modo certo, all'incremento della velocità di elocuzione. Questo parametro in alcune metriche ritmiche comporta problematiche nei risultati infatti tale parametro, in alcuni casi, viene normalizzato (nPVI, VarcoΔC). Un'altra relazione individuata, sempre tra questi due ambiti, riguarda i gruppi ritmici specificati nelle analisi musicali. Dal punto di vista linguistico questi gruppi ritmici corrispondono certamente alle unità

fonetiche realizzate che in questo studio sono intesi come run. Una differenza sostanziale è stata riscontrata invece nei parametri correlati all'accento, in musica la nota accentata è caratterizzata da una maggiore intensità rispetto alla nota non accentata, mentre nelle analisi linguistiche la sillaba accentata o tonica presenta anche una maggiore durata rispetto a quella atona. Quindi, nelle analisi musicali la nota accentata è caratterizzata da una forte intensità ma non da una durata superiore, rispetto a quanto avviene nelle sillabe toniche.

Di seguito, per una maggiore chiarezza, riportiamo il modello ritmico estratto dalle analisi musicali:

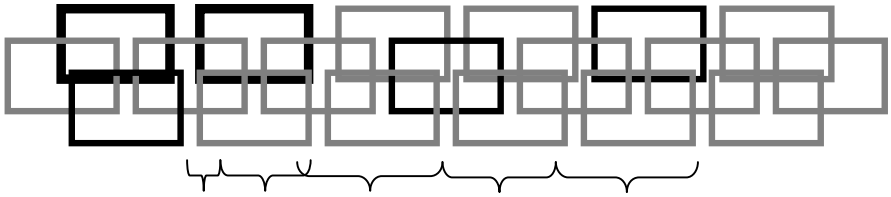


Figura 11: rappresentazione ritmica degli enunciati.

Queste considerazioni effettuate sul rapporto tra musica e lingua insieme alla Figura sopra esposta suggeriscono il ricorso ad altre caratteristiche insite nella lingua, come ad esempio i parametri connessi alla prominenza, che permetterebbero una maggiore caratterizzazione del ritmo delle lingue naturali.

6. CONCLUSIONI

In conclusione possiamo affermare che l'influenza della lingua prima sulla seconda è evidente e si manifesta in tutti i livelli linguistici.

Le prove percettive e i risultati ottenuti attraverso il sistema di Speech Recognition mostrano l'importanza che il livello fonetico-fonologico assume nel riconoscimento del parlato. Sia gli ascoltatori nel test percettivo che il sistema di SR usano il livello fonetico-fonologico per discriminare l'italiano L1 da quello L2. Una considerazione importante riguarda l'intonazione. Durante le prove percettive, gli ascoltatori hanno distinto spesso gli andamenti intonativi dell'italiano L1 da quelli dell'italiano L2 e, nello stesso tempo, l'andamento prodotto dall'italiano settentrionale rispetto a quello dell'italiano meridionale. Questo dato implica che gli ascoltatori hanno all'interno della loro competenza anche un modello intonativo e tale modello consente di distinguere i parlanti nativi da quelli non nativi ma anche il parlante del nord Italia e quello del sud.

Anche le prove musicali avvalorano questa osservazione, infatti, si notano disposizioni accentuali che determinano andamenti ritmici molto differenti tra di loro. Ciò che emerge con chiarezza è l'importanza che riveste nell'andamento ritmico degli enunciati la distribuzione delle prominenze. Proprio su questo ultimo parametro si concentreranno future ricerche volte a investigare dal punto di vista acustico le caratteristiche insite nella prominenza.

BIBLIOGRAFIA

Arvaniti, A. (2009), Rhythm, Timing and the Timing of Rhythm, Phonetica, 66, 46-63.

Bertini, C. & Bertinetto, P. M. (2008), On modeling the rhythm of natural languages, in Proceedings of Speech Prosody 2008, Campinas, Brazil, May 6-9, 427-430.

Mendicino, A. & Romito, L. (1991), Isocronia e base di articolazione: uno studio su alcune varietà meridionali, in Quaderni del Dipartimento di Lingue dell'Università della Calabria, S.L. 3, 49-65.

Nespor, M. (1993), Fonologia, Bologna, Bologna: il Mulino.

Pallotti, G. (2003), La seconda lingua, Milano, Milano: Strumenti Bompiani.

Romito, L., Tarasi, A. & Lio, R. (2011), Italian Index: Rhythmical-prosodic analysis of Italian L2 produced by Albanian, Chinese, Polish and Romanian speakers, V CFE, in press.

Romito, L., Tarasi, A. & Lio, R. (2012, in press), A rhythmical-prosodic analysis of Italian L1 and L2, in Prosodic and Rhythmic Aspects of L2 Acquisition. The case of Italian, Cambridge, Cambridge Scholar Publishing, in press.

Russo, M. (2010), Prosodic Universals, Roma, Roma: Aracne editrice.

CONDITIONAL RANDOM FIELDS COME STRUMENTO DI INDAGINE PER LA RILEVAZIONE AUTOMATICA DI PROMINENZE SILLABICHE

Enrico Leone[1], Antonio Origlia[1], Bogdan Ludusan[2]
[1]LUSI-Lab, Dip. Fisica, Università di Napoli "Federico II"
[2]CNRS-IRISA, Rennes, France

erik.leone82@gmail.com, antonio.origlia@unina.it, bogdan.ludusan@irisa.fr

1. INTRODUZIONE

Il fenomeno della prominenza sillabica, secondo il quale a specifiche unità segmentali di un enunciato viene conferita una particolare enfasi, è stato ampiamente studiato sia da un punto di vista linguistico [12, 3, 5] che da un punto di vista tecnologico [10, 11, 1, 2, 7]. L'elevato interesse nei confronti della prominenza nasce soprattutto dal fatto che l'occorrere di queste particolari unità segmentali rappresenta un importante punto di contatto con la componente soprasegmentale dell'enunciato. Si tratta, in effetti, di punti nei quali si realizza maggiormente la sincronizzazione tra eventi occorrenti sui due diversi livelli. La distribuzione di tali unità e il modo in cui ognuna di esse è realizzata all'interno dell'enunciato arricchisce il contenuto dell'enunciato stesso e, in molti casi, rimuove ambiguità. Vista l'importanza linguistica del fenomeno, è evidente quanto elevato sia l'interesse tecnologico che la sfida legata alla rilevazione automatica della prominenza rivesta. Le sillabe prominenti sono eccellenti punti di riferimento per chi si propone, come spesso è accaduto negli ultimi anni, di trattare automaticamente il parlato coinvolgendone la componente prosodica, rimasta tagliata fuori dalla prima ondata di applicazioni informatiche basate sulla voce. Applicazioni della rilevazione automatica delle prominenze si trovano nel riconoscimento del parlato, tramite gli approcci *island-driven*, nello studio del ritmo, considerandone la frequenza di occorrenza, e del parlato emotivo, consentendo l'estrazione di features acustiche particolarmente affidabili e significative.

Allo scopo di rilevare automaticamente l'occorrere di sillabe prominenti, in passato sono stati proposti sia modelli supervisionati, tramite l'impiego di *machine learning*, che non supervisionati, implementati sulla base di studi linguistici al riguardo. In generale, questi approcci considerano la prominenza sillabica come un fenomeno esclusivamente locale, secondo il quale evidenti differenze nel contenuto energetico, nella durata del nucleo sillabico e nei movimenti di pitch sincronizzati tra una sillaba e le due sillabe adiacenti, sono sufficienti a definire una funzione di prominenza. In questo lavoro presentiamo una investigazione del problema supportata dall'uso di particolari strumenti di classificazione automatica, i Conditional Random Fields (CRF) [6] e una loro variante, i Latent-Dynamics Conditional Random Fields (LDCRF) [8]. Tale indagine mira a verificare se vi siano vantaggi nell'estendere il contesto all'interno del quale valutare la prominenza come fenomeno locale, se una valutazione probabilistica riguardo la distribuzione di prominenze su tutto l'enunciato sia vantaggiosa e se vi siano altre features oltre a quelle classiche che debbano essere prese in considerazione quando si svolge questo tipo di indagine. La nostra speranza è che questo approccio, pur essendo puramente statistico e limitato nel fornire indicazioni precise riguardo al modo in cui le features proposte vengono utilizzate per risolvere il problema, possa fornire utili indicazioni al miglioramento dei sistemi non supervisionati, e quindi alla realizzazione di un modello predittivo dell'occorrere delle prominenze più approfondito di

quelli attualmente proposti. L'uso che facciamo dello strumento al fine di raggiungere questo scopo è quello di ottenere una stima delle prestazioni che è possibile raggiungere variando l'estensione del contesto, il set di features utilizzate e confrontando un modello che assume un'unica dinamica nella sequenza delle unità proposte con uno che, al contrario, assume che esistano delle sottodinamiche interne alla sequenza principale.

2. I MODELLI SUPERVISIONATI

Il problema di etichettare sequenze di dati è stato affrontato più volte in passato. L'esempio di maggior successo per questo genere di compito sono senz'altro gli Hidden Markov Models (HMM), un modello generativo che definisce una distribuzione di probabilità congiunta P(X,Y) dove X e Y sono variabili aleatorie che descrivono rispettivamente una sequenza di osservazioni e la corrispondente sequenza di etichette. I modelli generativi, per definire tale distribuzione, devono basarsi sull'enumerazione di tutte le possibili sequenze di osservazioni, rendendo necessaria l'introduzione di un importante vincolo sulle osservazioni stesse al fine di rendere trattabile il problema: l'indipendenza di ogni elemento della sequenza di osservazioni da tutti gli altri elementi. Sebbene le HMM abbiano fornito buone prestazioni in compiti di etichettatura legati all'analisi del parlato, si veda per esempio il riconoscimento automatico, è evidente che modelli capaci di tenere conto dell'interazione tra le features e di dipendenze a lungo termine tra le osservazioni siano ben più adatti a questo genere di compito. I CRF rispondono a questa esigenza rimuovendo l'assunzione di indipendenza e modellando la probabilità condizionata P(Y|x), evitando di modellare la distribuzione di probabilità delle sequenze di osservazioni, tipicamente poco interessante, e rendendo trattabile il problema dell'inferenza per problemi in cui la dipendenza tra le osservazioni non può essere trascurata, come avviene nel parlato. La sequenza di etichette restituita dal CRF viene calcolata in base alla seguente equazione

$$Y^* = \arg\max_Y P(Y|X)P(X) = \arg\max_Y P(Y|X) \qquad (1)$$

Un CRF descrive la dinamica delle etichette e la loro relazione con le osservazioni in termini di *feature functions*. Nella formulazione generale, queste particolari funzioni descrivono sia la relazione tra le osservazioni e le etichette che le relazioni tra coppie consecutive di etichette. E' possibile costruire *feature functions* anche per modellare relazioni più lunghe. In fase di training i dati di addestramento vengono utilizzati per stimare un insieme di pesi che descrivono quanto un certo valore sia collegato ad una determinata etichetta e quanto sia frequente trovare il valore della variabile y seguito dalla valore y'. Calcolare la probabilità di osservare una certa sequenza di etichette data la sequenza di osservazioni ed i pesi α_y e $\alpha_{y,y'}$ stimati in fase di addestramento si riduce a calcolare la seguente equazione

$$p(Y|X,\alpha) = \frac{1}{Z(X)} exp\left(\begin{array}{c} \sum_{t=1}^{N} \sum_{y \in Y} \alpha_y f_{y,t}(y_t, x_t) + \\ \sum_{t=1}^{N-1} \sum_{(y,y') \in Y^2} \alpha_{y,y'} f_{y,y'}(y_t, y_{t+1}) \end{array} \right) \qquad (2)$$

dove Z(X) è una costante di normalizzazione, N è il numero di osservazioni, $f_{y,t}$ sono le *feature functions* che modellano le relazioni tra osservazioni ed etichette ed $f_{y,y'}$ sono le *fea-*

ture functions che modellalo la dinamica tra etichette consecutive. In Figura 1 riportiamo uno schema della struttura di un CRF.

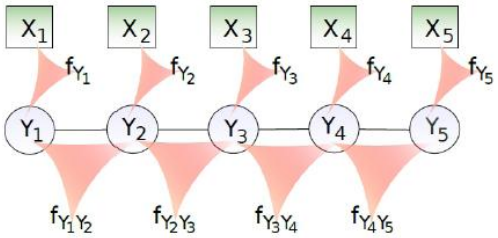


Figura 1: Rappresentazione grafica di un CRF (versione a catena lineare).

Uno dei limiti dei CRF è la mancanza della possibilità di modellare variabili latenti. Una estensione dei CRF che introduce le variabili latenti per etichettare un'intera sequenza con un'unica etichetta sono gli Hidden Conditional Random Fields (HCRF) [13], dei quali riportiamo uno schema in Figura 2. Impiegando un CRF, considerando la sequenza di etichette come variabili latenti ed etichettando ogni sequenza di etichette come appartenente ad una certa classe C, il framework precedente può essere interamente riutilizzato intruducendo la classe C tra i parametri delle *feature functions*. In Figura 2 riportiamo uno schema semplificato degli HCRF nel quale, per motivi di leggibilità, non vengono riportati i collegamenti tra osservazioni e classe relativamente alle funzioni f_{yiC}. Evidentemente gli HCRF non sono adatti al problema in questione in quanto il nostro scopo è etichettare sequenze di sillabe mentre gli HCRF sono progettati per assegnare un'unica etichetta ad una sequenza di osservazioni. L'introduzione delle variabili latenti, tuttavia, è utile per valutare se esistano due tipi di dinamica dei quali tenere conto per quanto riguarda lo studio del fenomeno della prominenza: una dinamica generica, legata alla distribuzione delle prominenze a livello di enunciato secondo vincoli ritmici, ed una dinamica locale, legata più strettamente ai rapporti tra sillabe entro un contesto più o meno esteso. Gli LDCRF sono una ulteriore estensione dei CRF che unisce la possibilità di assegnare un'etichetta ad ogni osservazione e la possibilità di modellare variabili latenti degli HCRF. In Figura 3 riportiamo uno schema della struttura di un LDCRF.

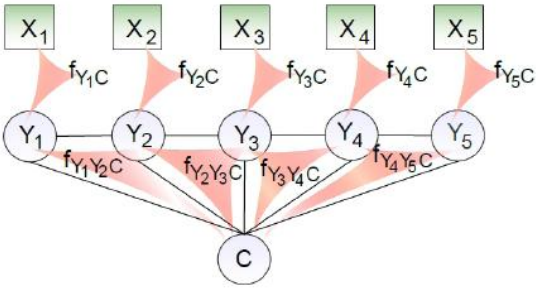


Figura 2: Rappresentazione grafica di un HCRF

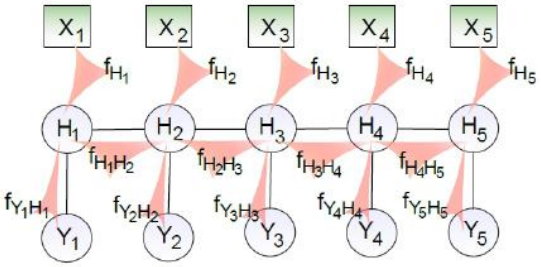


Figura 3: Rappresentazione grafica di un LDCRF

3.TEST

Durante i test abbiamo utilizzato un corpus di parlato italiano contenente numeri (sottoinsieme di SPEECON [9]) ed un corpus inglese di parlato letto (sottoinsieme di TIMIT [4]). Entrambi i corpora sono stati precedentemente annotati manualmente da un esperto linguista per indicare le sillabe prominenti ed impiegati per addestrare e testare il sistema automatico. Il sottoinsieme di SPEECON utilizzato è costituito da 288 files contenenti un minimo di 5 sillabe (15 in media, 4265 totali) per 15 minuti di parlato. Questo materiale è stato utilizzato anche per gli esperimenti presentati in [1]. Il sottoinsieme di TIMIT utilizzato è costituito da 382 files contenenti un minimo di 4 sillabe (12.51 in media, 4780 totali) per 17 minuti di parlato. Questo materiale è stato utilizzato anche in [11].

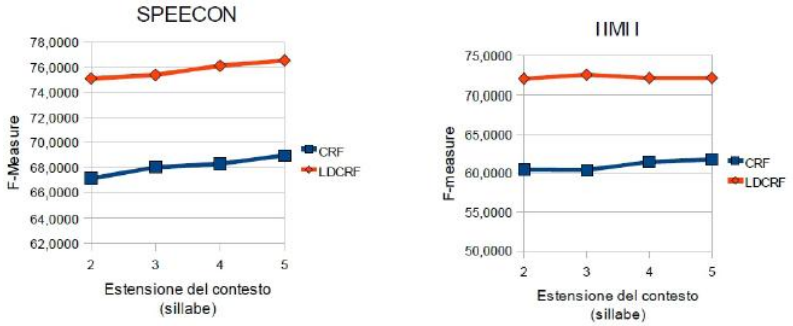


Figura 4: Confronto tra CRF ed LDCRF (Set base)

Come riferimento, utilizziamo un sistema non supervisionato di annotazione automatica di prominenze basato sulla combinazione di movimenti del pitch all'interno del nucleo sillabico e di energia media e durata di questo. Per la prima serie di esperimenti, costruiamo, per ogni enunciato, una serie di vettori, uno per ogni sillaba, contenenti la durata del nucleo sillabico, stimata come la banda -3db del massimo dell'energia interno all'intervallo, l'energia media all'interno del nucleo sillabico ed un parametro di valutazione del movimento del pitch calcolato come segue

$$m = \begin{cases} \frac{p_{max} - p_{mix}}{max(p_{max} - p_{min})} & if\ t_{max} > t_{min} \\ \\ \frac{1}{max(p_{max} - p_{min})} & altrimenti \end{cases} \qquad (3)$$

dove p_{max} è il valore massimo del pitch all'interno del nucleo, p_{min} è il minimo, t_{max} è il tempo di occorrenza del massimo e t_{min} è il valore di occorrenza del minimo. Questo parametro è lo stesso utilizzato dall'approccio di riferimento per tenere conto dei movimenti di pitch per il calcolo della prominenza. La sequenza di sillabe, così rappresentata, viene data in input ai classificatori automatici, che producono l'etichettatura automatica. Il protocollo di test impiegato è la 10-fold cross validation. Il contesto, espresso in numero di sillabe adiacenti da considerare nella costruzione delle *feature functions* è stato fatto variare tra 2 e 5. Per la seconda serie di esperimenti, volta a valutare l'impatto di un nuovo parametro nel calcolo della prominenza, introduciamo una nuova feature nel vettore di ogni sillaba: la lunghezza dell'intero segmento. Per valutare gli approcci proposti viene calcolata la F-measure (Classe true: prominente). In questo modo si intende mettere a confronto la capacità dei sistemi proposti di individuare la posizione delle sillabe effettivamente prominenti (veri positivi) evitando, allo stesso tempo, di introdurre nell'insieme delle sillabe etichettate come prominenti un numero eccessivo di sillabe non prominenti (falsi positivi).

L'andamento della F-measure al variare dell'estensione del contesto nella prima serie di esperimenti, eseguiti utilizzando il set di features base, è riportato in Figura 4. Per quanto riguarda la seconda serie di esperimenti, nella quale viene introdotta la lunghezza di ogni singola sillaba nei vettori di features, l'andamento della F-measure al variare dell'estensione del contesto è rappresentato in Figura 5. La Tabella 1 mostra il prospetto dei risultati ottenuti.

Tabella 1: Confronto tra il metodo di riferimento ed il metodo supervisionato. Per CRF ed LDCRF viene utilizzato il risultato migliore. Le misure riportate indicano la F-measure (Classe true: prominente)

	SPEECON	TIMIT
Baseline	73.31%	58.55%
CRF Set base	69.29%	61.75%
LDCRF Set base	76.54%	72.51%
CRF Set esteso	83.06%	72.16%
LDCRF Set esteso	84.48%	78.43%

CONCLUSIONI

Abbiamo proposto due modelli di etichettatura automatica di sequenze di osservazioni, i CRF e gli LDCRF, come strumento di indagine per studiare il fenomeno della prominenza sillabica. I test presentati mostrano che, facendo variare l'estensione del contesto sillabico considerato, il sistema ottiene le prestazioni migliori impiegando 3-4 sillabe di contesto, in evidente contrasto con i metodi tradizionali che si limitano alle due sillabe adiacenti. La sistematica superiorità del modello LDCRF rispetto al modello CRF suggerisce che sia utile affiancare alla tradizionale visione della prominenza quale fenomeno locale, modellata negli LDCRF dalle dinamiche interne, una valutazione sulla distribuzione delle prominenze a livello globale, modellata dalla dinamica generale. Questa linea di studio potrebbe essere utile, in particolare, nel formalizzare il collegamento tra la frequenza di occorrenza delle sillabe prominenti con la classificazione ritmica delle lingue. Infine, abbiamo osservato un importante aumento delle prestazioni includendo nel set di features l'informazione relativa alla durata di ogni sillaba oltre che del suo nucleo. Questo parametro sembra quindi dover essere tenuto in maggiore considerazione nello sviluppo di un sistema di annotazione di prominenze non supervisionato.

BIBLIOGRAFIA

[1] G. Abete, C. Cutugno, B. Ludusan, and A. Origlia. *Pitch behavior detection for automatic prominence recognition*. In Proc. of Speech Prosody [Online], 2010.

[2] M. Avanzi, A. Lacheret-Dujour, and B. Victorri. *A corpus based learning method for prominence detection inspontaneous speech*. In Proc. of Speech Prosody [Online], 2010.

[3] A. Eriksson, E. Grabe, and H. Traunmueller. *Perception of syllable prominence by listeners with and without competence in the tested language*. In Proc. of Speech Prosody, pages 275-278, 2002.

[4] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. *DARPA TIMIT acoustic phonetic continuous speech corpus CDROM*. 1993.

[5] C. Jensen. *Perception of prominence in standard british english*. In Proc. of ICPhS, 2003.

[6] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. In Proc. of ICML, pages 282-289, 2001.

[7] B. Ludusan, A. Origlia, and F. Cutugno. *On the use of the rhythmogram for automatic syllabic prominence annotation*. In Proc. of Interspeech, pages 2413-2416, 2011.

[8] A. Quattoni, Wang S., L.P. Morency, M. Collins, T. Darrell, and M. Csail. *Hidden-state conditional random fields*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29:1848-1852, 2007.

[9] R. Siemund, H. H□oge, S. Kunzmann, and K. Marasek. *Speecon-speech data for consumer devices*. In Proc. of LREC, pages 883-886, 2000.

[10] R. Silipo and S. Greenberg. *Automatic transcription of prosodic stress for spontaneous English discourse*. In Proc. of ICPhS, pages 2351-2354, 1999.

[11] F. Tamburini. *Reliable prominence identification in english spontaneous speech*. In Proc. of Speech Prosody [Online], 2006.

[12] R. Vanderslice and P. Ladefoged. *Binary suprasegmental features and transformational word-accentuation rules*. Language, 48:819-836, 1972.

[13] S. B. Wang, A. Quattoni, L. P. Morency, and D. Demirdjian. *Hidden conditional random fields for gesture recognition*. In Proc. of CVPR, pages 1521-1527, 2006.

PERCEZIONE II

COMUNICARE IN UNA LINGUA SECONDA. IL RUOLO DELL'INTONAZIONE NELLA PERCEZIONE DELL'INTERLINGUA DI APPRENDENTI CINESI DI ITALIANO

Anna De Meo, Massimo Pettorino, Marilisa Vitale
Università degli studi di Napoli "L'Orientale"
ademeo@unior.it, mpettorino@unior.it, vitalem@unior.it

1. INTRODUZIONE

Gli studi sull'acquisizione delle lingue seconde mostrano una crescente attenzione nei confronti della competenza prosodica, elemento ormai considerato determinante sia nella comprensione del messaggio sia nell'identificazione del carattere non nativo di una pronuncia (Anderson-Hsieh et alii 1992, Munro & Derwing 2001, Trofimovich & Baker 2006, Kang 2010, Marotta & Boula de Mareüil 2010). Il presente lavoro si inserisce in una ricerca più ampia, di approccio pragmatico acquisizionale, che ha come obiettivo l'identificazione e la valutazione del ruolo svolto da alcuni parametri prosodico-intonativi (intonazione, velocità, pause) e dal livello segmentale nella realizzazione di una comunicazione efficace, con una particolare attenzione al rapporto tra grado di accento straniero percepito dal nativo e giudizio di efficacia comunicativa attribuito all'enunciato prodotto dal locutore non nativo (De Meo & Pettorino 2011, De Meo et alii 2011, Pettorino et alii 2011, De Meo & Pettorino 2012, De Meo et alii 2012).

Oggetto di questo studio è l'intonazione dell'interlingua di sinofoni apprendenti di italiano L2, impegnati nella realizzazione di sei specifici atti linguistici: asserzione, domanda, ordine, concessione, promessa e minaccia.

2. TEST PERCETTIVO: PARLATO NATURALE

Nella costruzione del corpus sono stati coinvolti 4 italiani madrelingua di diversa provenienza geografica,[1] e 12 cinesi apprendenti di italiano L2.[2] I non nativi sono stati selezionati sulla base di alcune caratteristiche, tutte considerate rilevanti negli studi sull'acquisizione delle lingue seconde (Schumann 1986, Piske et alii 2001, Flege 2009):

1. età di apprendimento (prima esposizione alla L2)
2. durata dell'apprendimento guidato in contesto LS e/o L2
3. tempo di permanenza in Italia
4. quantità di input (esposizione totale o parziale alla L2)
5. livello di competenza (*Quadro Comune Europeo di Riferimento per le lingue - QCER).*

I parlanti, suddivisi in 8 coppie con L1 comune per annullare la variabile interlinguistica (nativo-nativo e straniero-straniero), hanno letto e registrato un breve dialogo (ca. 1 minuto), così strutturato: parlante A: asserzione e promessa; parlante B: domanda e ordine; parlante A: concessione e minaccia.

[1] Tre voci femminili (Sardegna IT1, Piemonte IT2, Campania IT3) e una voce maschile (Lazio IT4).
[2] Otto voci femminili e quattro maschili (vedi tabella 1).

soggetto	sesso	età	età di apprendimento	durata apprendimento guidato in contesto LS	durata apprendimento guidato in contesto L2	tempo di permanenza in Italia	quantità di input	livello di competenza
CIN1	F	21	18	2 anni	0	1 mese	esposizione parziale	B1
CIN2	F	20	18	2 anni	0	1 mese	esposizione parziale	B1
CIN3	M	21	18	2 anni	0	1 mese	esposizione parziale	B1
CIN4	M	21	18	2 anni	0	1 mese	esposizione parziale	B1
CIN5	F	31	18	4 anni	3 anni	3 anni	esposizione parziale	C1
CIN6	F	19	18	2 anni	0	1 mese	esposizione parziale	B1
CIN7	F	47	30	1 anno	0	17 anni	esposizione totale	B2
CIN8	M	24	18	4 anni	1 anno	1 anno	esposizione parziale	C1
CIN9	F	26	18	4 anni	2 anni	2 anni	esposizione parziale	C1
CIN10	F	21	18	2 anni	0	1 mese	esposizione parziale	B2
CIN11	F	27	12	0	12 anni	15 anni	esposizione parziale	C2
CIN12	M	23	12	0	10 anni	11 anni	esposizione totale	C2

Tabella 1: Riepilogo dati parlanti non nativi.

Al fine di avere a disposizione tutti i tipi di enunciato per ogni parlante, ciascuna coppia ha registrato due versioni del dialogo invertendo i ruoli degli interlocutori. I 16 dialoghi risultanti sono stati segmentati, al fine di isolare gli enunciati corrispondenti ai sei atti linguistici oggetto dell'analisi. I 96 file audio così ottenuti (24 in italiano L1 e 72 in italiano L2 di cinesi) sono stati montati in maniera randomizzata in 50 diverse sequenze e somministrati a 50 ascoltatori italiani madrelingua di area campana, tutti docenti esperti nell'ambito della didattica dell'italiano a stranieri.

Ogni ascoltatore ha avuto il compito di valutare per ciascun enunciato il grado di accento straniero percepito (forte, lieve, nativo), l'efficacia comunicativa (nulla, sufficiente, buona) e di indicare quale parametro fosse risultato più rilevante nell'attribuzione del giudizio (qualità articolatoria, intonazione, velocità, silenzi).

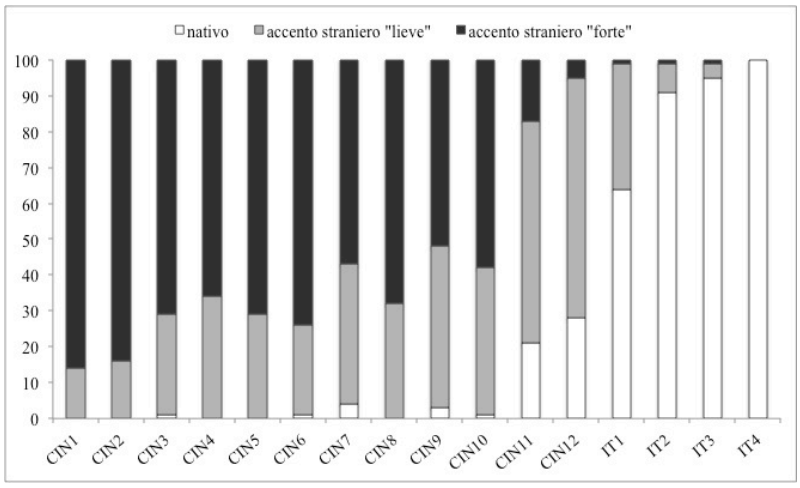


Figura 1: Test percettivo, parlanti cinesi (CIN) e italiani (IT). Giudizio di accento straniero ("nativo", straniero "lieve" e straniero "forte").

I dati riportati nella figura 1 mostrano una netta separazione tra i nativi e gli stranieri. Gli ascoltatori hanno attribuito un giudizio di accento nativo molto netto ai locutori italiani, sebbene la parlante sarda (IT1) venga valutata come nativa solo nel 64% dei casi. Decisamente distanziati CIN11 e CIN12, i quali ricevono un giudizio di accento nativo rispettivamente pari al 21% e al 28% e un giudizio di accento straniero prevalentemente "lieve" (62% e 67%). Tutti gli altri locutori sono chiaramente identificati come parlanti con accento straniero "forte" da una percentuale di ascoltatori che va dal 55% al 90% circa.

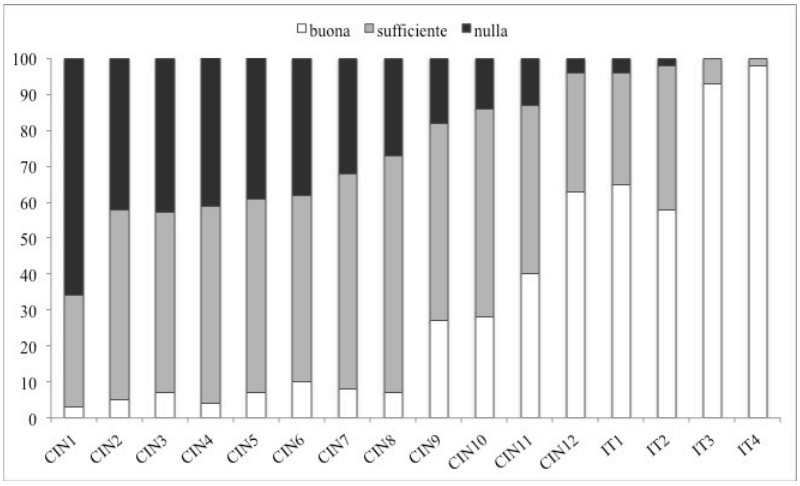


Figura 2: Test percettivo, parlanti cinesi (CIN) e italiani (IT). Giudizio di efficacia comunicativa: buona, sufficiente, nulla.

Rispetto all'efficacia comunicativa, come mostrato nella figura 2, gli unici locutori a non aver ricevuto alcun giudizio negativo sono i due madrelingua IT3 e IT4, che hanno ottenuto una valutazione di "buono" dalla quasi totalità degli ascoltatori (93% e 98%). Gli altri due locutori madrelingua, IT1 e IT2, così come il parlante cinese CIN12, ottengono un

risultato positivo superiore al 95%, ma con un 30% di ascoltatori che valuta solo "sufficiente" l'efficacia comunicativa dei loro enunciati.

La diversa valutazione ottenuta dai 4 parlanti nativi può essere messa in relazione alla diversa area geografica di provenienza: gli ascoltatori hanno mostrato un atteggiamento più favorevole verso le varianti diatopiche più simili alla propria. L'ottimo risultato ottenuto dal parlante cinese CIN12 può essere ricondotto alla precoce età di apprendimento (arrivo in Italia: 12 anni), combinato con un'esposizione alla L2 pressoché totale (ha un patrigno italiano, frequenta quasi esclusivamente giovani italiani, studia all'università).

Risultati piuttosto omogenei sono stati ottenuti dalle locutrici CIN9/10/11, con giudizi negativi compresi tra 13 e 18%. Da notare che la parlante CIN11, nonostante la precoce età di apprendimento (età di arrivo in Italia: 12 anni), non raggiunge i risultati ottenuti da CIN12, probabilmente a causa del grado di integrazione molto basso nella comunità italofona, con conseguente esposizione parziale all'italiano (vive con i genitori nella comunità cinese, lavora in un ristorante cinese, studia lingua e letteratura cinese all'Università). Le altre due parlanti straniere, CIN9 e CIN10, aventi la stessa età di prima esposizione (18 anni), ma percorsi di formazione guidata in contesto LS ed L2, tempo di permanenza in Italia e livello di competenza molto diversi (vedi Tabella 1), ottengono risultati simili sul piano dell'efficacia comunicativa.

Un altro gruppo di parlanti ad aver ottenuto risultati omogenei è quello formato da CIN7 e CIN8, accomunati esclusivamente da un'esposizione tardiva alla lingua italiana (età di apprendimento: 30 e 18 anni). La prima ha appreso l'italiano in contesto L2 ed ha un grado di esposizione all'input molto alto, poiché vive da 17 anni in Italia, è sposata con un italiano e insegna cinese a italiani; il secondo ha studiato l'italiano in Cina, dunque in contesto LS, vive solo da un anno in Italia, ma ha un'esposizione all'input molto alta, poiché vive e studia con italiani. I due parlanti hanno ottenuto un prevalente giudizio di "sufficiente" (57% e 64%) e un 30% circa di giudizi negativi.

I restanti locutori non nativi ottengono risultati più modesti (giudizio negativo superiore al 40% e, in un caso, quasi pari al 70%) e i giudizi positivi non vanno oltre il "sufficiente". Tutti i parlanti di questo gruppo, ad esclusione di CIN5, hanno la stessa condizione di apprendimento di CIN10, sebbene il loro livello di competenza sia leggermente inferiore (B1 del QCER). Colpisce il giudizio piuttosto negativo attribuito alla parlante CIN5 che ha un livello di competenza C1 del QCER e condizioni di apprendimento identiche a quelle di CIN9, il che sottolinea comunque la rilevanza dei fattori individuali rispetto a quelli sociolinguistici.

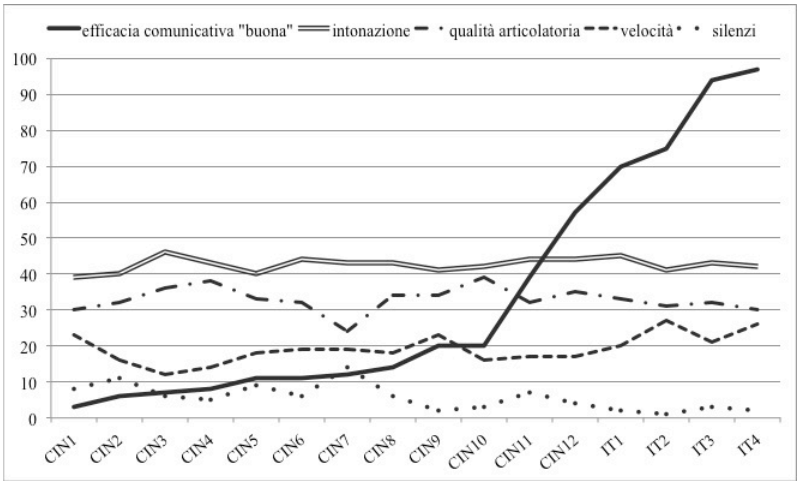


Figura 3: Test percettivo, parlanti cinesi (CIN) e italiani (IT). Fattori giudicati rilevanti nell'attribuzione del giudizio di efficacia comunicativa "buona" (valori percentuali).

La figura 3 mostra i risultati del test percettivo relativi ai singoli parametri acustici, per gli enunciati giudicati di efficacia comunicativa "buona" dell'intero corpus. Gli ascoltatori hanno attribuito il ruolo più rilevante all'intonazione e quello meno importante ai silenzi.

Relativamente al grado di accento straniero percepito, appare evidente (Figura 4) una relazione inversa con il grado di efficacia comunicativa. Gli enunciati giudicati pienamente soddisfacenti sul piano comunicativo vengono percepiti come prodotti da parlanti nativi (cfr. IT3 e IT4); a una diminuzione del grado di efficacia comunicativa si associa una crescente presenza di accento straniero, inizialmente lieve (cfr. IT1 e IT2), poi sempre più forte. È interessante notare come i due parlanti non nativi con precoce età di apprendimento della L2, CIN11 e CIN12, rappresentino una sorta di stadio intermedio tra nativi e stranieri: nonostante abbiano trascorso in Italia la metà della loro vita e abbiano raggiunto un livello di competenza C2 del QCER, continuano a essere identificati come stranieri da un'alta percentuale di ascoltatori (70%-80%). Tutti gli altri parlanti cinesi sono giudicati stranieri, con accento "lieve" o "forte", dalla quasi totalità degli ascoltatori.

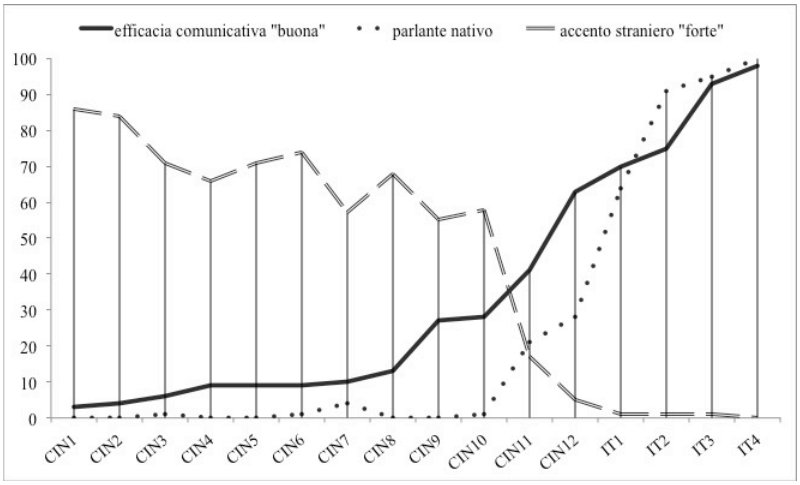


Figura 4: Test percettivo, parlanti cinesi (CIN) e italiani (IT). Relazione tra valutazione dell'accento straniero e giudizio di efficacia comunicativa (valori percentuali).

In figura 5 sono riportati i valori medi dei giudizi di efficacia comunicativa "buona" per i parlanti nativi e non nativi, per ciascuno dei sei atti linguistici presi in considerazione. Il grafico, come atteso, mostra che i nativi realizzano in maniera adeguata tutti i compiti linguistici, ottenendo giudizi positivi compresi tra il 76% e 90%. Per quanto riguarda i parlanti non nativi, i valori sono complessivamente bassi (tra il 10% e il 28%). Gli atti linguistici prodotti in maniera peggiore sono l'ordine e l'asserzione; i migliori risultano la concessione e la domanda; in posizione intermedia si collocano la promessa e la minaccia.

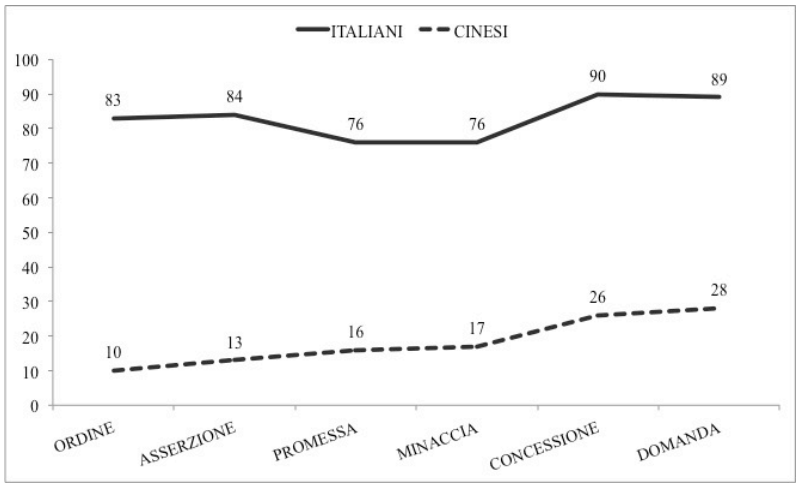


Figura 5: Test percettivo, parlanti cinesi e italiani. Relazione tra atto linguistico e giudizio di efficacia comunicativa.

3. TEST PERCETTIVO: PARLATO SINTETIZZATO

I risultati del test percettivo condotto sul parlato naturale hanno messo in evidenza il ruolo predominante dell'intonazione nella valutazione dell'efficacia comunicativa di un

enunciato da parte dell'ascoltatore nativo. Per trovare una conferma a questo dato e, al tempo stesso, per valutare il peso effettivo dell'intonazione nella percezione dell'accento straniero/nativo, è stato condotto un secondo test percettivo su file audio modificati artificialmente. Il nuovo corpus è stato costruito sulla base degli enunciati prodotti dai due parlanti nativi che hanno ottenuto i risultati migliori nel primo test (IT3 e IT4) e dai due parlanti cinesi considerati, al contrario, i peggiori a livello di efficacia comunicativa (CIN1 e CIN2). L'attenzione si è concentrata solo su 3 dei 6 atti linguistici considerati, i quali sono stati selezionati sulla base dei giudizi di efficacia comunicativa ottenuti dai parlanti cinesi nel primo test: ordine (valore minimo), minaccia (valore intermedio) e domanda (valore massimo).

Utilizzando il software Praat si è proceduto a una modifica dei segnali audio originali, trapiantando l'intonazione e, fatta eccezione per l'ultima fase della ricerca, lasciando invariati intensità, durate e silenzi[3]. Sono state effettuate le seguenti modifiche:[4]

DOMANDA: da Nm a NNm, da Nf a NNf, da NNm a Nm

MINACCIA: da Nm a NNm, da Nf a NNf, da NNf a Nf

ORDINE: da Nm a NNm, da Nf a NNf, da NNm a Nm, da NNf a Nf.

I 10 enunciati sintetizzati, tutti risultati di buona qualità uditiva, sono stati organizzati in maniera randomizzata in un file di ascolto, somministrato a 40 ascoltatori italiani madrelingua, tutti docenti di italiano a stranieri.

Relativamente al giudizio di accento straniero, il trapianto intonativo determina un peggioramento della valutazione degli enunciati prodotti dai parlanti italiani, che vengono riconosciuti come stranieri dal 56% degli ascoltatori (accento "lieve" 47%; accento "forte" 9%). I due cinesi mostrano invece un miglioramento, poiché il giudizio di accento straniero, inizialmente etichettato come "forte" dal 74% degli ascoltatori, si riduce al 51%, a vantaggio di un giudizio di accento "lieve", che passa dal 24% al 46% (Figura 6).

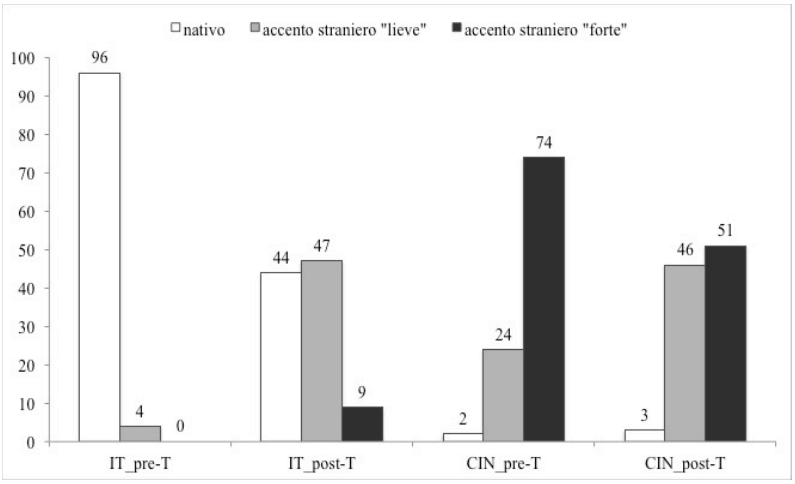


Figura 6: Test percettivo su parlato naturale e sintetizzato, parlanti italiani (IT) e cinesi (CIN). Giudizio di accento straniero (valori percentuali medi, calcolati su tutti gli enunciati).

[3] Sul procedimento del trapianto intonativo, si rimanda a Yoon (2007).

[4] N = nativo, NN = non nativo, f = femmina, m = maschio.

I risultati del test confermano l'importanza dell'andamento intonativo anche per l'efficacia comunicativa, che, infatti, per i due locutori italiani peggiora in maniera significativa (dal 96% al 35% di "buono") quando l'andamento intonativo originale viene sostituito da quello dei due parlanti stranieri. Al contrario, l'efficacia comunicativa dei locutori cinesi migliora notevolmente (dal 5% al 34%) quando sui loro enunciati viene trapiantato l'andamento intonativo dei nativi (Figura 7).

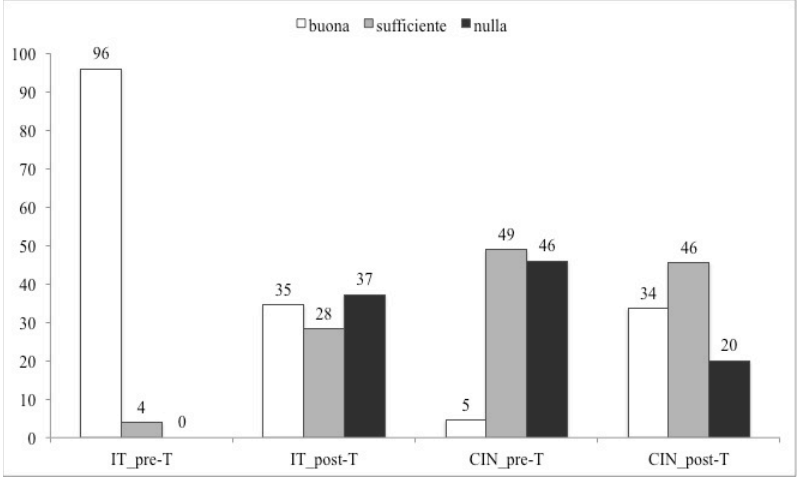


Figura 7: Test percettivo su parlato naturale e sintetizzato, parlanti italiani (IT) e cinesi (CIN). Giudizio di efficacia comunicativa (valori percentuali medi, calcolati su tutti gli enunciati).

Un'analisi dei dati scorporati per atti linguistici mostra un generale miglioramento dei locutori cinesi, sia per la domanda (Figura 8) sia per la minaccia (Figura 9), e un corrispettivo peggioramento dei locutori italiani. Per quanto riguarda la domanda, la voce nativa passa da un 100% di "buono" a un 62%, nel caso della minaccia dal 95% al 24% di "buono". Al contrario, per le due voci straniere si osserva un miglioramento pari a circa il 35%.

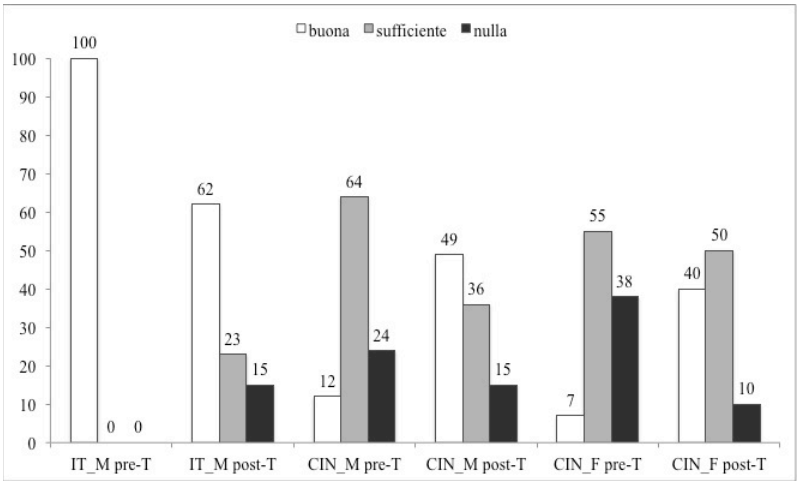


Figura 8: Test percettivo su parlato naturale (pre-T) e sintetizzato (post-T) di parlanti cinesi (CIN) e italiani (IT). Giudizio di efficacia comunicativa, atto linguistico DOMANDA (valore percentuale).

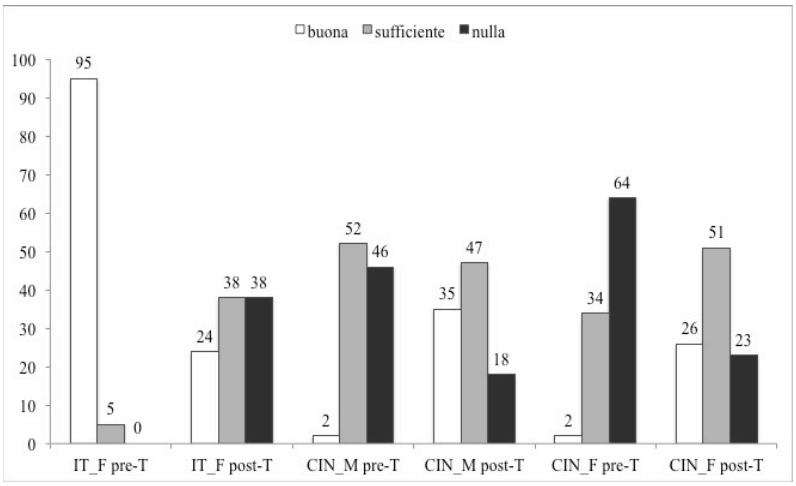


Figura 9: Test percettivo su parlato naturale (pre-T) e sintetizzato (post-T) di parlanti italiani (IT) e cinesi (CIN). Giudizio di efficacia comunicativa, atto linguistico MINACCIA (valore percentuale).

Per quanto riguarda l'atto linguistico dell'ordine, il trapianto dell'andamento intonativo determina un netto miglioramento dell'efficacia comunicativa dell'enunciato del parlante cinese maschio, ma non di quello della voce femminile straniera (Figura 10). I trapianti inversi, dagli stranieri ai nativi, confermano tale tendenza: l'efficacia comunicativa "buona" passa da 100% a 52% per la voce maschile italiana e crolla da 89% a 0% per la voce femminile. Le differenze appena evidenziate, relative ai diversi risultati ottenuti dalle voci femminili e da quelle maschili, potrebbero essere in parte dipese dalle difficoltà di tipo tecnico riscontrate per il trapianto delle voci femminili. Le pause silenti presenti nell'enunciato della parlante non-nativa, non corrispondendo a quelle prodotte dalla

parlante nativa, hanno richiesto un intervento di manipolazione più invasivo, che ha alterato la qualità del segnale acustico risultante.

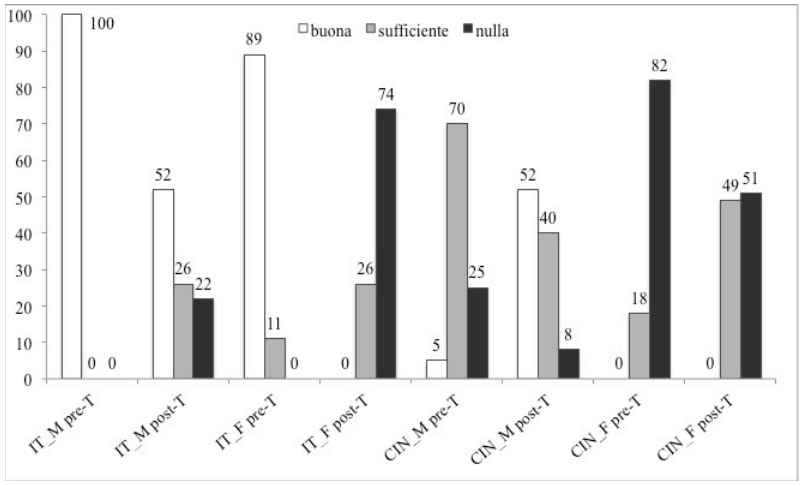


Figura 10: Test percettivo su parlato naturale (pre-T) e sintetizzato (post-T) di parlanti italiani (IT) e cinesi (CIN). Giudizio di efficacia comunicativa, atto linguistico ORDINE (valore percentuale).

Infine, poiché per l'enunciato prodotto dalla parlante cinese i dati del primo test percettivo avevano evidenziato un giudizio negativo degli ascoltatori sia relativamente all'intonazione (39%) sia relativamente alla velocità di articolazione (28%), si è proceduto a un secondo trapianto in cui entrambi i parametri soprasegmentali sono stati trasferiti dalla parlante italiana a quella cinese. Il risultato del test percettivo relativo al secondo trapianto ha mostrato un incremento del 40% del giudizio di efficacia comunicativa "buona" (Figura 11), confermando l'importanza di entrambi i parametri ai fini di una comunicazione efficace.

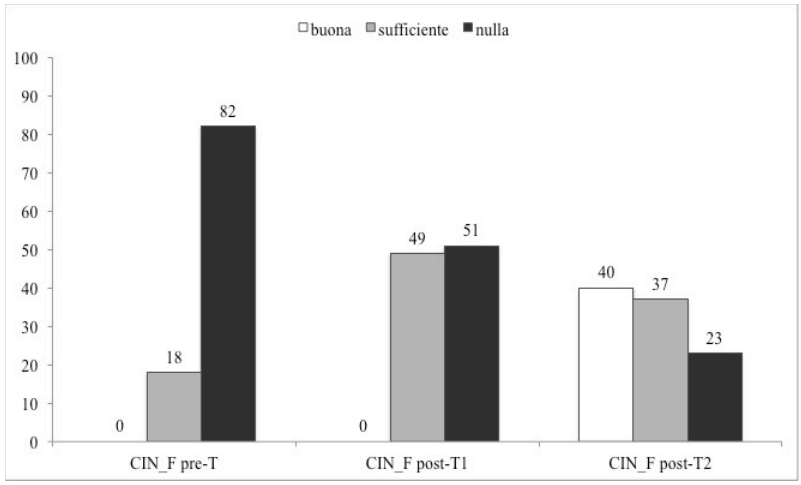


Figura 11: Test percettivo su parlato naturale (pre-T) e sintetizzato (post-T; post-T2 =

trapianto intonazione e velocità di eloquio) di parlanti cinesi (CIN). Giudizio di efficacia comunicativa, atto linguistico ORDINE (valore percentuale).

3. CONCLUSIONI

I risultati di questo studio confermano l'importanza dell'intonazione sia nello sviluppo della competenza comunicativa nella L2 sia nell'acquisizione di un accento simile a quello nativo, sebbene i dati suggeriscano l'opportunità di un approfondimento del ruolo svolto anche da altri parametri soprasegmentali, tra i quali la velocità di articolazione. La ricerca ha confermato la presenza di gradi di difficoltà variabili nella realizzazione prosodicamente efficace dei diversi atti linguistici. I sinofoni coinvolti nello studio, pur avendo ricevuto una valutazione generalmente scarsa dal punto di vista dell'efficacia comunicativa, hanno mostrato difficoltà più marcate nel caso dell'ordine e dell'asserzione.

Per quanto riguarda l'interferenza sul processo di acquisizione dei vari fattori linguistici ed extralinguistici considerati nello studio (età di apprendimento, durata dell'apprendimento guidato in contesto LS e/o L2, tempo di permanenza in Italia, quantità di input a cui si è esposti, livello di competenza), è opportuno tenere distinti l'efficacia comunicativa e il grado di accento straniero.

Un precoce età di apprendimento agisce positivamente solo se rafforzata da una esposizione totale alla L2; nel caso di input l'effetto "età" si riduce notevolmente: apprendenti con età di prima esposizione tarda, ma con alto livello di integrazione nella comunità del paese straniero, ottengono risultati analoghi ad apprendenti precoci con basso livello di integrazione (cfr. CIN12 e CIN7 *vs* CIN11).

In generale un apprendimento in contesto L2, con interazione tra un percorso guidato e uno spontaneo, produce effetti migliori del solo apprendimento guidato in contesto LS, sebbene anche quest'ultimo permetta di raggiungere livelli di competenza linguistica avanzati (cfr. CIN5, CIN8 e CIN9: C1 del QCER). Non possono essere sottovalutati casi specifici, in cui caratteristiche idiosincratiche di un individuo rendono particolarmente efficace l'apprendimento di una seconda lingua, permettendo di raggiungere anche in ambiente LS risultati paragonabili a quelli conseguibili in ambiente L2 (cfr. CIN10).

Relativamente all'accento straniero, solo un'età di apprendimento precoce sembra incidere in maniera significativa. Tutti i sinofoni coinvolti nello studio hanno un accento straniero "forte", i cui valori sono piuttosto elevati (dall'86% al 55%), tranne CIN11 e CIN12, i quali, pur avendo un accento straniero "lieve" (62% e 67%), riescono a essere riconosciuti come nativi da circa il 20% di italiani.

Il test percettivo condotto su parlato sintetizzato conferma l'impatto dell'intonazione sia sul giudizio di efficacia comunicativa sia su quello di accento. Il trapianto dell'andamento intonativo dagli enunciati dei nativi (IT3 e IT4) a quelli dei due cinesi giudicati peggiori nel primo test percettivo (CIN1 e CIN2), determina un miglioramento sia dell'efficacia comunicativa sia dell'accento straniero, che passa nettamente da "forte" a "lieve". Il trapianto in direzione inversa provoca marcati peggioramenti dei nativi, che vengono percepiti dagli ascoltatori come poco efficaci e con accento straniero "lieve".

I risultati del secondo test percettivo inducono a riflettere sulla possibilità di ottenere una riduzione dell'accento straniero e un miglioramento dell'efficacia comunicativa attraverso un percorso di insegnamento o autoapprendimento centrato sulla prosodia.

BIBLIOGRAFIA

Anderson-Hsieh, J., Johnson, R. & Koehler, K. (1992), The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody and syllable structure, Language Learning, 42, 529-555.

Boula de Mareuil, P. & Vieru-Dimulescu, B. (2006), The contribution of prosody to the perception of foreign accent, Phonetica, 63, 247-267.

De Meo, A. & Pettorino, M. (2012), Prosodia e italiano L2: cinesi, giapponesi e vietnamiti a confronto, in Apprendere l'italiano da lingue lontane: prospettiva linguistica, pragmatica, educativa (R. Bozzone Costa, L. Fumagalli & A. Valentini, editors), Perugia: Guerra Edizioni, 59-72.

De Meo, A., Pettorino, M. & Vitale, M. (2012), Non ti credo: i correlati acustici della credibilità in italiano L2, in Atti dell'XI Congresso dell'Associazione Italiana di Linguistica Applicata. Competenze e formazione linguistiche. In memoria di Monica Berretta (G. Bernini, C. Lavinio, A. Valentini & M. Voghera, editors), Perugia: Guerra Edizioni, 229-248.

De Meo, A., Pettorino, M. (2011), L'acquisizione della competenza prosodica in italiano L2 da parte di studenti sinofoni, in La didattica dell'italiano a studenti cinesi e il progetto Marco Polo. Atti del XV seminario AICLU (E. Bonvino & S. Rastelli, editors), Pavia: Pavia University Press, 67-78.

De Meo, A., Vitale, M., Pettorino, M. & Martin, P. (2011), Acoustic-perceptual credibility correlates of news reading by native and Chinese speakers of Italian, in Proceedings of the 17th International Congress of Phonetic Sciences (L. Wai-Sum & E. Zee, editors), City University of Hong Kong, Hong Kong, Cina, 1366-1369.

Flege, J. (2009), Give input a chance!, in Input Matters in SLA (T. Piske & M. Young-Scholten, editors), Bristol: Multilingual Matters, 175-190.

Kang, O. (2010), Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness, System, 38, 301-315.

Marotta, G. & Boula de Mareüil, P. (2010), Persistenza dell'accento straniero. Uno studio percettivo sull'italiano L2, in La dimensione temporale del parlato, Atti della V Conferenza dell'Associazione Italiana di Scienze della Voce (S. Schmid, M. Schwarzenbach & D. Studer, editors), Torriana (Italy): Universität Zürich, CD-rom Proceedings, 475-494.

Munro, M.J. & Derwing T.M. (2001), Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate, Studies of Second Language Acquisition, 23, 451-468.

Pettorino, M., De Meo, A., Pellegrino, E., Salvati, L. & Vitale, M. (2011), Accento straniero e credibilità del messaggio: un'analisi acustico-percettiva, in Contesto comunicativo e variabilità nella produzione e percezione della lingua, Atti del 7° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce (B. Gili Fivela, A. Stella, L. Garrapa & M. Grimaldi, editors), Roma: Bulzoni editore.

Piske, T., MacKay, I.R.A. & Flege, J.L. (2001), Factors affecting degree of foreign accent in an L2: a review, Journal of Phonetics, 29, 191-215.

Schumann, J. H. (1986), Research on the acculturation model for second language acquisition, Journal of Multilingual and Multicultural Development, 7, 379-392.

Trofimovich, P. & Baker, W. (2006), Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech, Studies in Second Language Acquisition, 28, 1-30.

Yoon, K. 2007. Imposing native speakers' prosody on non-native speakers' utterances: Thetechnique of cloning prosody. Journal of the Modern British & American Language & Literature 25(4). pp.197-215.

LA COMPRENSIONE ORALE IN ITALIANO L2. PER UN PROFILO SOPRASEGMENTALE DELLE INTERAZIONI BIDIREZIONALI A DUE VOCI DESTINATE AD APPRENDENTI DI LIVELLO A2

Giuseppina Vitale[1], Luisa Salvati[2], Elisa Pellegrino[3]
Università degli Studi di Napoli "L'Orientale"
[1]*vitalepina@gmail.com,* [2]*lsalvati@unior.it,* [3]*epellegrino@unior.it*

0. SOMMARIO

Il lavoro rappresenta il primo passo di un'indagine conoscitiva che si pone l'obiettivo di rilevare i modelli soprasegmentali di parlato utilizzati dagli autori dei manuali di italiano L2 al fine di favorire la comprensione orale degli apprendenti e di individuare gli indici più soggetti alla variazione. La ricerca nasce dalla constatazione che tanto negli studi sull'apprendimento/insegnamento dell'italiano L2, quanto nel Quadro Comune Europeo di Riferimento per le Lingue (QCER) mancano dati specifici relativi ai valori degli indici soprasegmentali che l'apprendente deve essere in grado di gestire per accertare il raggiungimento di un determinato livello di competenza linguistica. In conformità all'articolo 2 del Decreto Ministeriale del 4 giugno 2010 del Ministero dell'Interno in materia di disposizioni sulla conoscenza della lingua italiana per il rilascio del permesso di soggiorno, lo studio parte dall'analisi dei materiali audio presentati in quattro dei più aggiornati corsi di lingua italiana di livello A2.

Ai fini della ricerca, sono state quindi selezionate dieci interazioni bidirezionali simmetriche su due argomenti condivisi tra i quattro manuali: le vacanze e la casa. Il corpus di parlato raccolto è stato oggetto di analisi spettro-acustica. I dialoghi sono stati esaminati per singole catene foniche. Di ciascuna catena fonica è stato misurato il numero di sillabe realmente prodotte, la durata in secondi, la presenza di disfluenze, la durata delle pause vuote. Sulla base delle misure effettuate sono stati calcolati i seguenti parametri soprasegmentali: velocità di articolazione, velocità di eloquio, fluenza e valore della percentuale di silenzio e di tempo articolato.

Dai dati delle analisi spettro acustiche è stato riscontrato che a parità di approccio adottato, unità didattica, argomento, tipologia di parlato e livello linguistico, gli indici più stabili tra i quattro manuali sono la velocità di articolazione e la velocità di eloquio. Sebbene nel QCER le parole chiave utilizzate per definire lo stile di parlato da destinare agli apprendenti di livello A2 siano "chiarezza" e "lentezza" articolatoria, i valori riscontrati nei manuali esaminati sono più affini a quelli che in letteratura si considerano caratterizzanti di un parlato "normale". Parametri più soggetti alla variazione sono invece l'alternanza nel parlato di pause piene e vuote e la fluenza. Per quanto riguarda le disfluenze esse sono presenti solo in alcuni dialoghi di alcuni volumi; per quanto riguarda la fluenza, non solo fra i diversi manuali, ma anche fra i locutori dello stesso dialogo, ci sono parlanti che producono catene foniche molto lunghe, altri che invece si concedono pause dopo aver pronunciato una quantità inferiore di sillabe.

La metodologia di analisi adoperata in questo studio sarà successivamente estesa alle tracce audio destinate ad apprendenti di altri livelli di competenza linguistica, al fine di evidenziare analogie e differenze con i risultati finora conseguiti.

1. INTRODUZIONE

Lo sviluppo dell'abilità di comprensione orale rappresenta una prerogativa irrinunciabile per qualsiasi percorso di apprendimento linguistico: tanto per i bambini che acquisiscono la L1, quanto per una larga maggioranza di parlanti di L2, l'ascolto rappresenta il canale privilegiato attraverso cui entrare in contatto con l'ambiente linguistico circostante. Nonostante la centralità di tale abilità nello sviluppo della competenza linguistica, per molto tempo il *listening* ha rivestito un ruolo secondario nella riflessione scientifica di taglio didattico-acquisizionale, tanto da meritarsi la designazione di "Cinderella skill in second language learning" (Nunan, 1999:199).

L'importanza dell'ascolto nell'uso e nell'acquisizione delle lingue inizia, invece, a essere riconosciuta nella seconda metà del secolo scorso, con il rinvigorirsi dell'interesse verso lo sviluppo delle abilità orali (Nunan 2002: 238). In ambito psicolinguistico, la ricerca si orienta verso i meccanismi psico-cognitivi coinvolti nella decodifica del messaggio in L2. Il modello psicolinguistico dell'ascolto proposto da Levelt (1989) rappresenta una delle più note formalizzazioni di questo processo, in quanto vengono descritti nel dettaglio gli elaboratori (uditore, decodificatore e interprete), le procedure routinizzate e le conoscenze (lessicali e generali) cui l'utente attinge nel processo di decodifica dell'enunciato.

In ambito glottodidattico, l'attenzione è invece posta sulle tecniche da applicare in classe per favorire negli studenti il potenziamento delle abilità meta-cognitive necessarie per lo svolgimento dei compiti di ascolto (Vandergrift, 2002). A favore della rivalutazione di tale abilità, ha giocato anche l'affermarsi in anni recenti di una nuova concezione della comprensione orale: da una visione un po' restrittiva dell'ascolto come mera ricezione dell'*input* fonico, si è via via passati ad una definizione del *listening* più ampia e più attenta alle dinamiche messe in atto dall'ascoltatore per la decodifica degli enunciati. Rost (2002), infatti, qualifica la comprensione orale come un processo "ricettivo", "costruttivo", "collaborativo" e "interpretativo", rispetto al quale l'utente non si comporta più da spettatore passivo, così come previsto negli approcci didattici più datati; al contrario il parlante/ascoltatore prende parte attiva al processo ravvisando indizi contestuali, informazioni linguistiche ed enciclopediche utili per la comprensione del discorso (Vandergrift, 2003).

Un supporto alla decodifica dell'*input* di L2 è giocato anche dalle scelte lessicali, dal registro e dalle modalità espositive del parlante nativo nella comunicazione bilingue. Molti studi sono stati infatti dedicati alle strategie comunicative, alle modifiche fonologiche, morfosintattiche, lessicali e pragmatiche dell'*input* offerte dal parlante nativo (*Foreigner talk*) o dal docente di lingua (*Teacher talk*) con l'intenzione di rendere il messaggio più intelligibile all'interlocutore straniero (Ferguson, 1971; Freed, 1981; Long, 1981, 1983a, 1983b; Tarone, 1980).

Un filone di ricerca molto produttivo è stato pure quello incentratosi sul ruolo giocato dalla modulazione dei parametri soprasegmentali sull'abilità di ascolto. In letteratura, ci si è soffermati molto sul ruolo giocato dalla velocità di eloquio sulla comprensione orale. A partire dagli anni '90, infatti, nel mondo anglofono sono stati condotti molti studi sperimentali volti proprio a verificare il legame tra il livello di velocità con cui veniva prodotto un testo e il grado di comprensibilità dello stesso da parte di apprendenti o parlanti non nativi di inglese. Si puntava infatti ad identificare dei *range* di velocità di eloquio entro

i quali la comprensione di un testo veniva favorita (Conrad ,1989; Blau, 1990, Griffith, 1990, 1991, 1992).

Di recente è stato pure indagato il peso giocato da velocità ridotte o accelerate sulla capacità di decodifica testuale (Brindle & Salyter, 2002; Munro & Derwing, 2001) e sui vantaggi derivanti dalla possibilità di affidare agli ascoltatori il controllo della velocità dei propri interlocutori in *task* di parlato letto (Zhao,1997). In massima parte questi studi hanno riguardato l'inglese come lingua straniera. Studi specifici sulle caratteristiche ritmico-prosodiche che favoriscono la comprensione del parlato in apprendenti di Italiano L2 mancano. Generiche sono pure le indicazioni fornite nei descrittori del Quadro Comune Europeo di Riferimento per le lingue (QCER) relative ai valori degli indici soprasegmentali che l'apprendente deve essere in grado di gestire per accertare il raggiungimento di un determinato livello di competenza linguistica.

2. LO STUDIO

Le lacune presenti in letteratura e l'approssimazione dei descrittori del QCER hanno quindi portato ad avviare una ricerca mirante a:

1) rilevare e classificare i modelli soprasegmentali di parlato utilizzati dagli autori dei manuali di italiano L2 al fine di favorire la comprensione orale degli apprendenti e

2) individuare gli indici più soggetti alla variazione.

In linea con le disposizioni governative in materia di conoscenza linguistica per il rilascio del permesso di soggiorno, lo studio parte dalla disamina dei materiali audio presentati nei corsi di lingua di livello A2. L'articolo 2 del Decreto Ministeriale del 4 giugno 2010 del Ministero dell'Interno sancisce infatti che "per il rilascio del permesso di soggiorno Ce per soggiornanti di lungo periodo, lo straniero deve possedere un livello di conoscenza della lingua italiana che consente di comprendere frasi ed espressioni di uso frequente in ambiti correnti, in corrispondenza al livello A2 del Quadro comune di riferimento europeo per la conoscenza delle lingue approvato dal Consiglio d'Europa".

Nel QCER, l'A2 viene definito "Livello di sopravvivenza" (*Waystage*), poiché come si legge dalla 'Scala Globale' l'utente di tale livello è colui che:

"*Riesce a comprendere frasi isolate ed espressioni di uso frequente relative ad ambiti di immediata rilevanza (ad es. informazioni di base sulla persona e sulla famiglia, acquisti, geografia locale, lavoro). Riesce a comunicare in attività semplici e di routine che richiedono solo uno scambio di informazioni semplice e diretto su argomenti familiari e abituali. Riesce a descrivere in termini semplici aspetti del proprio vissuto e del proprio ambiente ed elementi che si riferiscono a bisogni immediati*" (QCER: 32).

Il livello A2 è pure quello in cui si ritrovano i maggiori descrittori relativi alle funzioni sociali. Si legge ad esempio che un apprendente deve essere in grado di "usare semplici espressioni convenzionali per salutare e rivolgere la parola a qualcuno, salutare le persone, chiedere come stanno, reagire alla risposta e portare a termine scambi comunicativi molto brevi" (QCER: 149). In sintesi, la competenza dell'apprendente di livello A2 gli consente di soddisfare le esigenze comunicative di base attraverso la gestione di brevi enunciati, strutturalmente non complessi.

Relativamente all'abilità che costituisce il focus della ricerca, nella 'Scala di Livello' specifica per la comprensione orale, si trovano informazioni molto meno dettagliate. Si leggono infatti solo delle generiche descrizioni sul tipo di testo e sulla modalità di esposizione - lenta e chiara- che lo studente è in grado di gestire:

"*È in grado di comprendere quanto basta per soddisfare bisogni di tipo concreto, purché si parli lentamente e chiaramente. È in grado di comprendere espressioni riferite ad aree di priorità immediata (ad es. informazioni veramente basilari sulla persona e sulla famiglia, acquisti, geografia locale e lavoro), purché si parli lentamente e chiaramente.* (QCER: 83).

Come accennato precedentemente, nel documento europeo, considerato attualmente un sistema di riferimento teorico-concettuale e politico-attuativo per la gestione del contatto linguistico e una guida per l'attuazione di un modello di glottodidattica dell'italiano L2 (Vedovelli, 2001), mancano descrittori specifici relativi ai valori degli indici soprasegmentali che l'apprendente deve essere in grado di gestire per accertare il raggiungimento di un determinato livello di competenza.

2.1. Materiali e metodi

Per delineare il profilo soprasegmentale del parlato rivolto ad apprendenti di livello A2 e, di conseguenza, per colmare le lacune del QCER, sono state passate al vaglio le sezioni dedicate alla comprensione orale di un composito campione di manuali di lingua e cultura italiana rivolti a studenti stranieri. Nel dettaglio, sono state selezionate quattro opere tra le proposte editoriali più innovative e coerenti con le ultime metodologie didattiche, pubblicate nell'ultimo decennio dalle case editrici specializzate nella didattica dell'italiano L2. Fanno parte del corpus alcuni file audio estratti dai CD-rom allegati ai seguenti testi:

1. Balboni, P. E. & Mezzadri, M. (2002), *Rete!* 1. Perugia: Guerra Edizioni.
2. Bozzone Costa, R., Ghezzi, C. & Piantoni, M. (2005), *Contatto*. 1B. Corso di italiano per stranieri. Torino: Loescher editore.
3. Costamagna, L., Falcinelli, M. & Servadio, B. (2008), *Io & l'italiano*. Corso di lingua per principianti assoluti. Livello A1- A2.1. Firenze: Le Monnier.
4. Trifone, M., Filippone, A. & Sgaglione, A. (2008), *Affresco italiano* A2. Corso di lingua per stranieri. Firenze: Le Monnier.

I quattro volumi del corpus, che si pongono come obiettivo principale quello di fornire agli studenti degli strumenti linguistici e culturali necessari per stabilire i primi contatti con la lingua italiana, sono stati selezionati in quanto rappresentano i manuali maggiormente utilizzati presso i Centri Linguistici d'Ateneo e gli enti formativi privati In secondo luogo, ritenendo inappropriato rispetto agli obiettivi della nostra ricerca valutare i testi audio di manuali basati su un metodo grammaticale-traduttivo, i suddetti volumi sono tutti accomunati da un approccio comunicativo e un metodo diretto,: essii puntano a sviluppare la padronanza di capacità linguistico-comunicative minime, quali ad esempio, comprendere frasi ed espressioni di uso frequente relative ad ambiti fondamentali o comunicare in attività non complesse che implichino scambi di informazioni elementari su argomenti familiari. La grammatica è presentata in maniera induttiva e l'accuratezza della forma acquista pari dignità rispetto alla competenza pragmatica e comunicativa. Anche le abilità ricettive e produttive vengono esercitate secondo i dettami dell'approccio comunicativo-funzionale, attraverso attività che consentono agli studenti di "process spoken discourse for functional purposes" (Flowerdew & Miller, 2005: 13).

Infine, i testi sono stati selezionati in quanto si rivolgono ad apprendenti stranieri di lingue materne diverse, anche tipologicamente distanti dall'italiano, aventi magari un sistema grafico non alfabetico. Nei manuali non si forniscono specifiche indicazioni sulla natura autentica o didattizzata dei testi presentati per lo sviluppo della capacità di ascolto. Tuttavia,

in nessuno dei volumi vengono offerti campioni di parlato autentico; si forniscono piuttosto tracce audio in cui si tentano di riprodurre situazioni comunicative concrete. Tale scelta metodologica può essere interpretata alla luce delle considerazioni sviluppate in letteratura sulla possibilità, nonché sulla convenienza a presentare nei corsi di lingua materiali audio che tendano all'autenticità, ma che allo stesso tempo vengano calibrati sulle competenze effettive degli apprendenti. Ur infatti scrive: "Students may learn best from listening to speech which, while not entirely authentic, is an approximation to real thing, and is planned to take into account the learners' level of ability and particular difficulties" (Ur, 2007: 23).

Per il nostro studio, la tipologia testuale indagata è stata il dialogo a due voci, una maschile e una femminile. Al fine di annullare la variabile argomento, sono state selezionate 10 interazioni bidirezionali simmetriche su due argomenti condivisi tra i quattro manuali: le vacanze e la casa. I testi hanno uno stile espressivo informale e registro linguistico colloquiale. Il corpus di parlato così raccolto, è stato esaminato per singole catene foniche (CF)[1]. Di ciascuna catena fonica è stato misurato il numero di sillabe realmente prodotte, la durata in secondi, la presenza di disfluenze, la durata delle pause vuote. Sulla base delle misure effettuate sono stati calcolati i seguenti parametri soprasegmentali. (Savastano et al., 1995; Pettorino & Giannini, 1997):

- velocità di articolazione (VdA), intesa come il rapporto fra il numero delle sillabe e il tempo in cui esse sono prodotte (sill/s);
- velocità di eloquio (VdE), calcolata come il rapporto fra il numero delle sillabe e il tempo totale comprensivo di pause vuote e piene (sill/s);
- fluenza (F), data dal rapporto fra il numero delle sillabe e il numero delle catene foniche (sill/CF);
- valore della percentuale di silenzio, disfluenza e tempo articolato.

Per indagare il profilo soprasegmentale delle interazioni bidirezionali a due voci proposte nei volumi di lingua italiana selezionati, i dati delle analisi spettro acustiche saranno presentati per argomento, prima le vacanze e poi la casa, e per singolo manuale.

2.2. Analisi dei dati

In "Affresco italiano", l'abilità di ascolto viene esercitata attraverso tre dialoghi della durata di circa 20 secondi ciascuno. Le *performance* dei sei locutori si compongono di tempo fonatorio e di silenzio; nessuno produce infatti disfluenze. Fatta eccezione per il secondo parlante del secondo dialogo, per il quale la percentuale di silenzio è del 33%, il tempo articolato delle altre voci non è mai inferiore all'80% (Figura 1).

[1] "Per catena fonica si intende la porzione di parlato compresa tra due pause silenti, mentre per enunciato si intende la durata complessiva dell"eloquio comprese la durata delle sequenze articolate, la durata delle pause non silenti e la durata delle pause silenti" (Giannini, 2010: 1232).

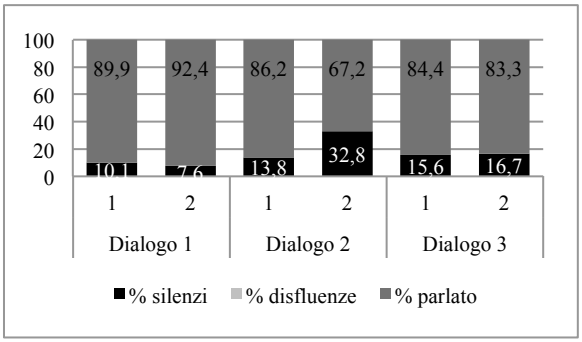


Figura 1: "Affresco italiano" - composizione del parlato per parlante.

Passando agli indici prosodici analizzati, in tutti e tre i dialoghi, è possibile notare una certa uniformità nei valori della velocità di articolazione e di eloquio. Nel caso del primo parametro i valori oscillano da 4,6 a 5,6 sill/s; per la velocità di eloquio invece da 3,7 a 4,9 sill/s (Figura 2).

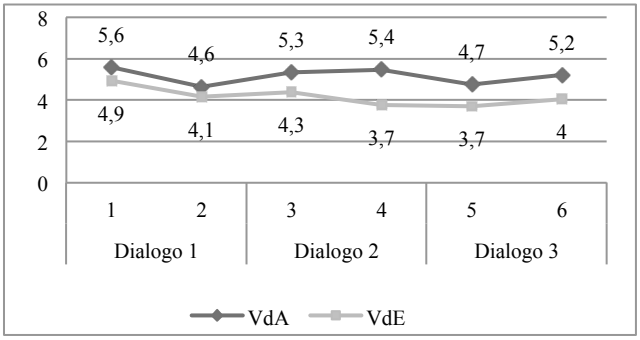


Figura 2: "Affresco italiano" - valori di VdA e di VdE per parlante.

Molto variabile appare invece la consistenza sillabica delle catene foniche prodotte dagli speaker dei tre dialoghi (Tabella 1).

	1° Speaker	**2° Speaker**
Dialogo nr. 1	5,7	11,5
Dialogo nr. 2	7,0	7,5
Dialogo nr. 3	9,8	6,6

Tabella 1: "Affresco italiano" - valori di fluenza per parlante.

Nel primo dialogo, addirittura, il secondo speaker produce catene foniche di lunghezza raddoppiata rispetto al primo. Più stabili i dati della fluenza tra gli speaker del 2° dialogo, che si concedono pause dopo circa 7 sillabe. All'interno dello stesso manuale, dunque, nell'ambito della stessa unità didattica, risulta che gli interlocutori producono quantità di pause vuote completamente diverse.

In "Rete!" la traccia audio dedicata alle vacanze è sotto forma di un dialogo della durata di circa 50 secondi, nei quali l'interazione tra i due speaker è scandita da percentuali diverse di tempo articolato e di silenzi (Figura 3).

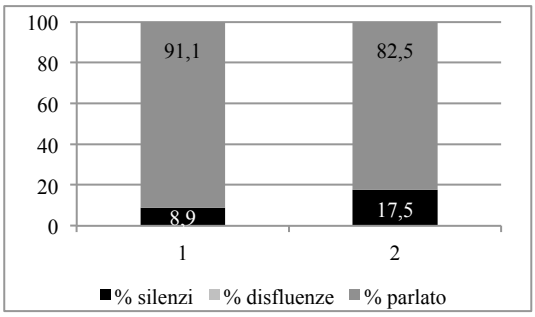


Figura 3: "Rete!" - composizione del parlato per locutore.

Nello specifico, con il secondo *speaker* la percentuale di silenzi sale dall'8,9 al 17,5%. In nessuno dei due casi, ricorrono pause piene all'interno delle catena fonica. Per quanto riguarda gli altri indici prosodici analizzati, analogamente al testo precedentemente esaminato, a valori simili di velocità di articolazione ed eloquio (rispettivamente 6 e 5 sill/s) si contrappongono valori di fluenza completamente differenti: il primo parlante, infatti, produce catene foniche di lunghezza dimezzata rispetto al primo (Tabella 2).

	VdA	VdE	Fluenza
Parlante 1	5,9	5,4	10,9
Parlante 2	6	5	6

Tabella 2: Valori di VdA, VdE e Fluenza per parlante.

La similarità dei valori della velocità di articolazione e di eloquio, sempre compresi in quel *range* oscillante fra 4,05 e 8,06 sill/s (Giannini, 2000) identificato per l'italiano standard e la disparità della consistenza sillabica delle catene foniche, la si ritrova anche negli altri due volumi "Io & L'Italiano" e "Contatto" (Tabelle 3 e 4).

.

	VdA (sill/s)	VdE (sill/s)	Fluenza (sill/CF)
Parlante 1	7	5,8	9,7
Parlante 2	6,1	4	4,8

Tabella 3: "Io & L'Italiano"- valori di VdA, VdE e Fluenza per parlante.

	VdA (sill/s)	VdE (sill/s)	Fluenza (sill/CF)
Parlante 1	6,3	5	9,1
Parlante 2	6,1	5,7	18,4

Tabella 4: "Contatto"- valori di VdA, VdE e Fluenza per parlante.

Tuttavia, i dialoghi di "Io & L'Italiano" e di "Contatto" differiscono da quelli di "Affresco italiano" e di "Rete!" per l'occorrenza sia di pause vuote sia di pause piene (Figure 4 e 5).

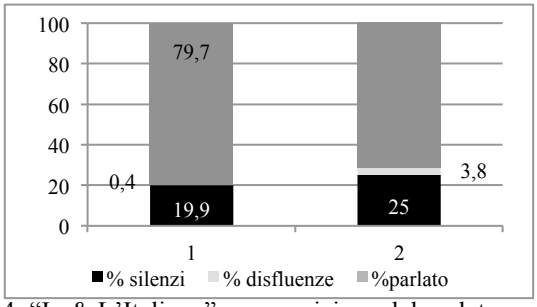


Figura 4: "Io & L'Italiano": composizione del parlato per parlante.

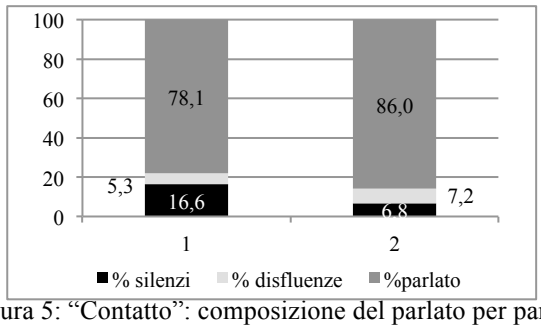


Figura 5: "Contatto": composizione del parlato per parlante.

Non trattandosi di parlato spontaneo, bensì di dialoghi recitati a partire da un testo ben definito e calibrato sui bisogni linguistico-comunicativi previsti dal livello A2, la presenza di pause piene tipiche invece del parlato spontaneo (Pettorino & Giannini, 2005), può essere letta come la scelta degli autori dei testi di presentare all'apprendente forme di interazioni quanto più vicine a quelle della comunicazione reale. Ricorrono infatti correzioni, false partenze, nasalizzazioni che segnalano all'ascoltatore la messa in atto da parte dell'interlocutore di processi di pianificazione linguistica *on line*.

Passando ai dati delle analisi spettro acustiche dei dialoghi aventi per argomento la casa, nei quattro volumi si confermano i valori della velocità di articolazione e di eloquio ricavate nei dialoghi sulle vacanze (Figura 6).

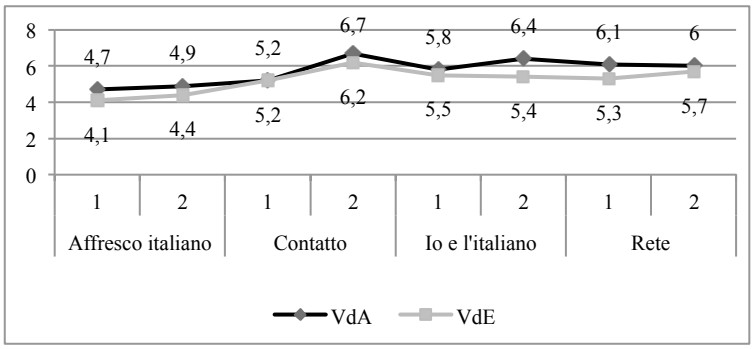


Figura 6: Valori di VdA e VdE per parlante e per volume.

In “Rete!” e in “Contatto”, inoltre, si ripresentano in maniera piuttosto accentuata le differenze nei valori della fluenza (Figura 7). In “Contatto” il primo speaker produce catene foniche di circa 15 sillabe, quindi di lunghezza raddoppiata rispetto a quelle del suo interlocutore, che si concede invece una pausa ogni 8 sillabe. In “Rete!”, le differenze nei valori della fluenza si attenuano, e si attestano intorno alle due sillabe per catena fonica (8,7 sill/CF rispetto a 6,7 sill/CF). Tale parametro invece non differenzia le prestazioni dei partecipanti ai dialoghi di “Affresco italiano” e di “Io & L’Italiano”: nei due volumi gli speaker producono in media rispettivamente una pausa silente ogni 6 e 7 sillabe.

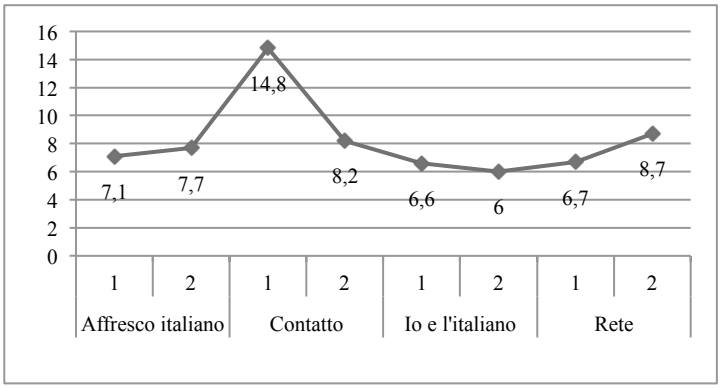


Figura 7: Valori di Fluenza per parlante e per volume.

Anche i dati relativi alla distribuzione del tempo fonatorio concorrono a differenziare “Io & L’Italiano” e “Affresco italiano” da un lato e “Rete!” e “Contatto” dall’altro (Figura 8).

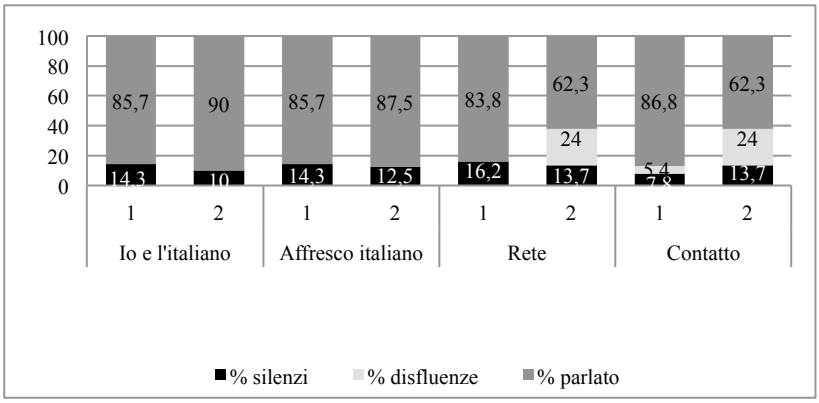


Figura 8: Composizione del parlato per parlante e per volume.

Nei primi due volumi, la percentuale di tempo articolato supera l’85% del totale. La restante parte è occupata da silenzi. In “Rete!” e in “Contatto”, invece, le percentuali di tempo fonatorio si riducono per la presenza delle disfluenze.

3. CONCLUSIONI

Il lavoro fa seguito alla constatazione che, tanto in letteratura quanto nel Quadro Comune Europeo di Riferimento per le Lingue, mancano descrittori specifici relativi ai valori degli indici soprasegmentali che l'apprendente deve essere in grado di gestire per accertare il raggiungimento di un determinato livello di competenza linguistica. Tali lacune hanno indotto quindi ad avviare un'indagine sui modelli soprasegmentali di parlato utilizzati dagli autori dei manuali di italiano L2 al fine di favorire la comprensione orale degli apprendenti e a verificare gli indici più soggetti alla variazione.

Dai dati delle analisi spettro-acustiche condotte su un corpus di 10 interazioni bidirezionali a due voci, sul tema delle vacanze e della casa, tratte da 4 dei più aggiornati manuali di Italiano L2, è stato riscontrato che a parità di approccio adottato, unità didattica, argomento, tipologia di parlato e livello linguistico, gli indici più stabili sono la velocità di articolazione e la velocità di eloquio.

A una lettura più approfondita dei dati si evince, inoltre, quanto i valori di tali indici siano distanti dalle disposizioni del QCER relative al parlato da produrre con apprendenti di livello *Waystage*. Nei descrittori dell'abilità di ascolto le parole chiave sono "chiarezza" e "lentezza" articolatoria; al contrario i valori riscontrati nei manuali (5-6 sill/s) sono più affini a quelli che in letteratura si considerano caratterizzanti di un parlato "normale" (Giannini, 2000). Parametri soggetti alla variazione inter e intra-testuale sono invece l'alternanza nel parlato di pause piene e vuote e la fluenza. Per quanto riguarda le disfluenze esse sono presenti solo in alcuni dialoghi di alcuni volumi, eppure l'intento pedagogico di tutti i manuali esaminati, è quello di presentare un parlato quanto più possibile vicino ai contesti reali della vita quotidiana.

Per quanto riguarda la fluenza, non solo fra i diversi manuali, ma anche fra i locutori dello stesso dialogo, ci sono parlanti che producono catene foniche molto lunghe (15 sill/CF), altri che invece si concedono pause dopo aver pronunciato una quantità inferiore di sillabe. Tale variabilità è plausibile in un parlato spontaneo, ma è lecito chiedersi se sia legittimo in un parlato preparato *ad hoc* per determinate tipologie di apprendenti.

Questi primi risultati inducono a sviluppare la ricerca su diversi piani. Innanzitutto si lavorerà all'elaborazione e alla somministrazione di test di comprensione orale al fine di verificare gli effetti della variabilità degli indici prosodici sul grado di comprensibilità dei dialoghi stessi. Si auspica, inoltre, di estendere la metodologia di analisi adoperata in questo studio ai materiali audio offerti per lo sviluppo dell'abilità di ascolto per apprendenti di altri livelli di competenza linguistica, al fine di evidenziare analogie e differenze con i risultati finora conseguiti. Lo scopo ultimo sarà quello di stilare un profilo soprasegmentale multilivello dell'apprendente di italiano L2 rispetto alla capacità di comprensione orale.

BIBLIOGRAFIA

A.A.V.V. (2002), Quadro comune europeo di riferimento per le lingue: apprendimento, insegnamento, valutazione, Firenze: La Nuova Italia.

Balboni, P.E. & Mezzadri, M. (2002), Rete!1. Perugia: Guerra Edizioni.

Blau, E.K. (1990), The Effect of Syntax, Speed and Pauses on Listening Comprehension, TESOL Quarterly, Vol. 24, no. 4, 746-753.

Bozzone Costa, R., Ghezzi, C. & Piantoni, M. (2005), Contatto. 1B. Corso di italiano per stranieri. Torino: Loescher editore

Brindley, G. & Slatyer, H. (2002), Exploring Task Difficulty in ESL Listening Assessment, Language Testing, Vol. 19, no.4, 369-394.

Conrad, L. (1989), The effects of time-compressed speech on native and EFL listen-ingcomprehension, Studies in Second Language Acquisition, no. 11, 1-16.

Costamagna, L., Falcinelli, M. & Servadio, B. (2008), Io & l'italiano. Corso di lingua per principianti assoluti. Livello A1- A2.1. Firenze: Le Monnier.

D.M. 4 giugno 2010 del Ministero dell'Interno. http://www.interno.it/mininterno/export/sites/default/it/sezioni/servizi/legislazione/immigrazione/0954_2010_06_16_DM_04062010.html.

Ferguson, C.A. (1971), Absence of copula and the notion of simplicity, in Pidginization and creolization of language (D. Hymes, editor). London: Cambridge University Press.

Flowerdew, J. & Miller, L. (2005), Second Language Listening. Theory and Practice. Cambridge: Cambridge University Press.

Freed, B. (1981), Foreigner talk, baby talk, native talk, International Journal of the Sociology of Language, 28, 19-39.

Giannini, A. (2000), Range di variabilità della velocità di articolazione in italiano, in Atti del XXVIII Convegno Nazionale AIA, Trani, Italia, Giugno 10-13, 253-256.

Giannini, A. (2010), Uno sguardo al ritmo e alla prosodia, in Oriente, Occidente e Dintorni. Scritti in onore di Adolfo Tamburello (F. Mazzei, P. Carioti, editors). Napoli: Il Torcoliere, Vol III, 1227-1239.

Griffiths, R. (1990), Speech Rate and NNS Comprehension: A Preliminary Study in Time-Benefit Analysis, Language Learning, Vol. 40, no. 3, 311-336.

Griffiths, R. (1991), Language Classroom Speech Rates: A Descriptive Study, TESOL Quarterly, Vol. 25, no. 1, 189-194.

Griffiths, R. (1992), Speech Rate and Listening Comprehension: Further Evidence of the Relationship, TESOL Quarterly, Vol. 26, no. 2, 385-390.

Levelt, W.J.M. (1989), Speaking. From intention to articulation, Cambridge, MA: MIT Press.

Long, M.H. (1981), Input, interaction and second-language acquisition, in Native language and foreign language acquisition (H. Winitz, editor). New York: Annals of the New York Academy of Sciences.

Long, M.H. (1983a), Native speaker/non-native speaker conversation and the negotiation of comprehensible input, Applied linguistics, Vol. 4, no.2, 126-141.

Long, M.H. (1983b), Native speaker/non-native speaker conversation in the second language classroom, in On TESOL '82. Pacific perspectives on second language learning and teaching (M.A. Clarke & J. Handscombe, editors),Washington, D.C.: TESOL.

Munro, J.M. & Derwing, T. (2001), What Speaking Rates Do Non-Native Listeners Prefer?, Applied Linguistics, Vol. 22, no. 3, 324-337.

Nunan, D. (1999), Second language teaching & learning, Boston: Heinle & Heinle Publishers.

Nunan, D. (2002), Listening in language learning, in Methodology in Language Teaching. An Anthology of Current Practice (J. Richards & W.A. Renandya, editors), Cambridge: Cambridge University Press, 238-241.

Pettorino, M. & Giannini, A. (2005), Analisi delle disfluenze e del ritmo del dialogo romano, in Italiano parlato. Analisi di un dialogo (F. Albano Leoni & R. Giordano, editors), Napoli: Liguori editore, 89-104.

Pettorino, M. & Giannini, A. (1997), Il discorso politico: una questione di stile, in Atti delle VI Giornate di Studio del GFS, Napoli.

Rost, M. (2002), Teaching and Researching Listening, London, UK: Longman.

Rost, M. (2006), Areas of research that influence L2 listening instruction, in Current Trends in the Development and Teaching of the four language skills (E. Usó-Juan & A. Martínez-Flor, editors) Walter de Gruyter: Berlin, 47-74.

Savastano, E., Giannini, A. & Pettorino, M. (1995), Aspetti prosodici del parlato dei politici, in AIA Atti del XXIII Convegno Nazionale, Bologna, 171-176,

Tarone, E. (1980), Communication strategies, foreigner talk, and repair in interlanguage, Language Learning, 30, 417-431.

Trifone M., Filippone, A., Sgaglione, A. (2008), Affresco italiano A2. Corso di lingua per stranieri. Firenze: Le Monnier.

Ur, P. (2007), Teaching Listening Comprehension, Cambridge: Cambridge University Press.

Vandergrift, L. (2002), 'It was nice to see that our predictions were right': Developing Metacognition in L2 Listening Comprehension, Canadian Modern Language Review, no 58, 555-575.

Vandergrift, L. (2003), Listening: theory and practice in modern foreign language competence. Retrieved January 4, 2009, da http://www.llasc.asc.uk/resources/gpg/67.

Vedovelli, M. (2001), Guida all'italiano per stranieri. Dal Quadro comune europeo per le lingue alla Sfida salutare, Roma: Carocci.

Zhao, Y. (1997), The Effects of Listeners' Control of Speech Rate on Second Language Comprehension, Applied Linguistics, Vol. 18, no. 1, 49-68.

LA COMPETENZA PERCETTIVA NELL'APPRENDIMENTO DELL'ITALIANO L2: UNO STUDIO SU APPRENDENTI SINOFONI

Luisa Salvati
Università degli Studi di Napoli "L'Orientale"
lsalvati@unior.it

1. SOMMARIO

La comprensione orale di una lingua straniera è un obiettivo fondamentale per gli apprendenti di una seconda lingua: si tratta di un'abilità complessa che si sviluppa progressivamente fin dalle prime fasi dell'apprendimento, contestualmente allo sviluppo della percezione uditiva che è parte integrante della comprensione orale. La maggior parte degli studi sulla percezione in L2 è incentrata sugli aspetti segmentali del parlato, trascurando l'analisi delle informazioni prosodiche veicolate dal segnale acustico, sebbene nella L1 sia stato riconosciuto un ruolo importante alla dimensione prosodica. Infatti, è stato dimostrato che nel discorso spontaneo, il processo percettivo si avvale anche delle informazioni date dai tratti soprasegmentali: l'intonazione, le pause, il ritmo, la quantità, le variazioni del tono e della velocità di eloquio, che indicano le intenzioni del parlante e segnano i confini interni e i punti di enfasi all'interno di un enunciato.

La presente ricerca ha l'obiettivo di analizzare la percezione dei tratti ritmico-prosodici in apprendenti d'italiano L2, in un'ottica comparativa con l'italiano L1. In particolare, è stato scelto il parlato argomentativo spontaneo, che implica l'atto perlocutorio di convincere: lo scopo è quello di indagare la relazione fra il grado di persuasività raggiunto dal parlante e i correlati ritmico-prosodici del suo parlato, rispetto alla competenza percettiva dell'apprendente non nativo.

In termini di percezione del parlato, i risultati ottenuti sono significativi in quanto dimostrano che non solo esiste una relazione tra persuasività e tratti ritmico-prosodici, ma che tale relazione è influenzata dalla competenza percettiva dell'ascoltatore. In particolare, i dati emersi dall'analisi del *corpus* in L1/L2, confrontati con i dati dei *corpora* rispettivamente in L1 e in L2, rivelano che mentre i parlanti nativi percepiscono una netta relazione fra persuasività e correlati ritmico-prosodici, la competenza percettiva degli apprendenti cinesi sembra non rilevare una relazione significativa fra prosodia e capacità di persuasione. Ne è una riprova la differenza fra le valutazioni attribuite dagli stranieri e quelle assegnate dagli italiani: alla persuasività raggiunta da ciascun parlante, gli ascoltatori non nativi attribuiscono valutazioni sempre elevate, siano essi nativi o non nativi, a differenza dei giudizi eterogenei dati dagli ascoltatori italiani. Rispetto a tali risultati, sono state avanzate due ipotesi, l'una di natura cognitiva, l'altra di natura socio-linguistica e socio-pragmatica, che gettano una nuova luce sugli studi riguardanti la competenza percettiva da un punto di vista acquisizionale e, dunque, sull'abilità di comprensione orale rispetto ad alcune tipologie testuali; inoltre, sul piano socio-fonetico e comunicativo, i dati ricavati risultano essere interessanti se si considera che oggi i media ospitano anche stranieri che parlano l'italiano come seconda lingua, in linea con i cambiamenti in atto nell'odierna società, e che dall'altra parte dello schermo, una grande fetta di pubblico è costituita proprio dagli stranieri.

2. INTRODUZIONE

2.1. La competenza percettiva in apprendenti sinofoni di italiano L2

La competenza percettiva in una seconda lingua è il risultato di una complessa costruzione di variabili, relativi all'apprendente, alla quantità e alla qualità dell'esposizione alla seconda lingua, all'uso della L1 e della L2 nel tempo, alla tipologia di insegnamento e, infine, alle differenze individuali relative alla motivazione, all'attitudine e al filtro affettivo che si innesca. A tali variabili si aggiungono altresì gli effetti dei modelli prosodici della L1 sulla percezione della L2.

Rispetto agli studi del passato, che concepivano il transfer come un mero passaggio di qualcosa da un lingua all'altra e lo associavano solo alle teorie comportamentiste, di recente l'attenzione è stata focalizzata proprio su come la differente organizzazione ritmico-temporale nella produzione delle varie lingue dipenda dalla segmentazione degli elementi acustici e da come quest'ultimi vengano valutati in termini fonetici e fonologici. Nel caso di lingue in contatto con un alto livello di similarità, l'incidenza della L1 sulla L2 può essere positiva per quanto riguarda l'apprendimento di fonologia, morfosintassi, lessico e pragmatica. In altre parole, l'influenza che un sistema linguistico esercita su un altro può facilitare il processo di apprendimento nel caso di lingue imparentate e non tipologicamente distanti; mentre, in caso contrario, può renderlo difficoltoso (Giacalone Ramat, 1994; Chini, 2010). L'incidenza negativa della L1 può risultare molto evidente in ambito fonologico, tanto che la pronuncia degli apprendenti nella L2 indica come i suoni e i tratti prosodici subiscano l'influenza delle caratteristiche della L1 o di altre lingue note.

Flege (1987) osserva che tale influenza opera anche sulla competenza percettiva nella L2, poiché gli apprendenti riscontrano enormi difficoltà nella discriminazione, ad esempio, dei suoni della L2, in particolar modo se sono simili a quelli della L1.[1] Tuttavia, è opportuno sottolineare che, pur esistendo molti altri fattori che condizionano il processo di acquisizione di una L2, accelerando o bloccando lo sviluppo dell'interlingua (come la marcatezza degli elementi linguistici, il tipo di *input* e di apprendimento, lo stile cognitivo, la personalità e la motivazione), l'età rimane uno dei fattori che maggiormente influenzano lo sviluppo della competenza prosodica in L2. Il periodo in cui un individuo può apprendere le medesime competenze del parlante nativo è limitato ai primi anni di vita. Dopo questa fase, è molto difficile che i non nativi riescano ad acquisire padronanza dell'assetto prosodico di una seconda lingua, a un livello paragonabile a quello di un nativo (Birdsong, 1999).

Nel caso particolare degli apprendenti sinofoni, il cinese è una lingua tonale ed isolante, tipologicamente distante dall'italiano. Pertanto, quando gli apprendenti cinesi si accostano all'italiano incontrano numerosi ostacoli e impiegano molto tempo nel cercare di entrare in sintonia con una lingua che presenta strutture completamente diverse dalla propria L1 (a meno che non abbiano in precedenza appreso altre lingue straniere, soprattutto se tipologicamente vicine all'italiano). In specifico, dal punto di vista della comprensione orale, Costamagna (2011: 51) afferma che "gli apprendenti sinofoni, la cui lingua materna ha strutture ritmico-intonative distanti dall'italiano, accedono alla comprensione del parlato con grande difficoltà non riuscendo a percepire e segmentare la catena parlata in modo efficace". Lo sviluppo della comprensione è influenzato anche dall'organizzazione morfologica dell'italiano, lingua flessiva-fusiva. Nei primi stadi, i cinesi tentano di afferrare gli elementi prominenti che li possano aiutare a orientarsi nell'ascolto, anche se, non riconoscendo ele-

[1] Cfr. Bunham & Mattok, 2007; Aoyama & Guion, 2007.

menti che possano fungere da punti di riferimento, percepiscono il messaggio linguistico in L2 come una massa sonora in cui non riescono a distinguere gli elementi discriminanti. In stadi più avanzati di conoscenza della L2, essi sviluppano una maggiore consapevolezza della distanza esistente fra le due lingue, anche e soprattutto per quanto riguarda la struttura prosodica che risulta tale da non permettere pratiche di *transfer* dalla L1. Anche l'acquisizione della capacità di gestione delle variazioni intonative ai fini di una comunicazione che sia pragmaticamente efficace si può riscontrare soltanto nei livelli più avanzati, poiché ai primi stadi di apprendimento, ci si ferma in genere alla distinzione fra valenza interrogativa o esclamativa di una frase (Costamagna, 2011).

Dunque, ciò che caratterizza la competenza percettiva in apprendenti cinesi di italiano L2 è la progressione molto lieve fra uno stadio e l'altro dell'interlingua, come dimostrato da De Meo e Pettorino (2011: 76) in uno studio sul rapporto tra competenza linguistica e competenza prosodica: "[…] un cinese può raggiungere un livello di competenza linguistica C1 e non sviluppare adeguatamente l'abilità di comunicare efficacemente con parlanti madrelingua italiani attraverso la gestione della prosodia e dell'intonazione". La comprensione orale, inoltre, è ritardata anche dalla diversità, tra cinese e italiano, dei modelli pragmatico-comunicativi, che spesso rendono difficoltosa la partecipazione dei cinesi all'interazione orale in italiano L2.[2]

3. MATERIALI E METODI

3.1. Obiettivi della ricerca

La presente ricerca ha l'obiettivo di analizzare la percezione dei tratti ritmico-prosodici in apprendenti d'italiano L2, in un'ottica comparativa con l'italiano L1, nell'ambito di un medesimo contesto, il *debating*, relativo a tre situazioni differenti: a) il *debating* in L1; b) il *debating* in L2; c) il *debating* in cui nativi e non nativi si confrontano rispettivamente in italiano L1 e L2. Lo scopo specifico è quello di indagare la relazione fra il grado di persuasività raggiunto dal parlante e i correlati ritmico-prosodici del suo parlato, rispetto alla competenza percettiva dell'apprendente non nativo.

È opportuno chiarire che la ricerca è stata condotta con la consapevolezza che la persuasività di un'argomentazione e del parlante che la esprime è il risultato di una sommatoria di elementi: il contenuto del testo, il modo con cui il parlante esprime le proprie opinioni, le modalità in cui attraverso il corpo egli accompagna le proprie parole. Nell'ambito di queste variabili, è stata isolata la componente prosodica per verificare il peso che quest'ultima può assumere sulla capacità di convincere un uditorio, non solo perché, attraverso l'uso della voce, l'oratore può suscitare commozione nel pubblico e, dunque, persuaderlo, ma anche perché i risultati di un'analisi spettro-acustica consentono di avere risultati misurabili e, pertanto, confrontabili.

3.2. La struttura del debating

Quale strumento di indagine per la raccolta dei corpora analizzati nel presente lavoro, è stato scelto il *debating*, un confronto fra due o più persone, oppure fra due o più squadre, in cui ciascuna persona, o ciascuna squadra – attraverso i propri componenti – è chiamato a esprimere e motivare i propri pro o i propri contro circa un determinato argomento, con lo

[2] Per un approfondimento degli studi sulla conversazione in L2, si rimanda a Chun, 2002; Gut, 2009.

scopo principale di persuadere un pubblico delle proprie argomentazioni. Nel presente lavoro, i dibattiti a squadre consistevano in due fasi. Nella prima, presieduta da un moderatore, ogni componente di ciascuna squadra aveva fino a due minuti di tempo per esprimere le proprie argomentazioni, per poi cedere la parola al componente del gruppo avversario. Nella seconda, entrambi i gruppi avevano fino a sei minuti di tempo per discutere liberamente fra loro, senza l'intervento di alcun moderatore, per cercare di convincere il pubblico, chiamato a valutare ciascun parlante su una serie di parametri – fra cui il grado di persuasività – in termini di "positivo" o "negativo".

3.3. I corpora

La ricerca si basa su tre diversi corpora: un corpus in italiano L1, un corpus in italiano L2, un corpus in italiano L1/L2. La raccolta dei dati per la costituzione dei tre corpora è durata tre mesi, da novembre 2010 a gennaio 2011. I tre *debating* sono stati audioregistrati mediante il software *Goldwave* versione 5.58 e videoregistrati con una videocamera Sony HDR-SR8E. Infine, è stata eseguita una trascrizione ortografica annotata dei tre *corpora* sulla base delle norme del progetto CLIPS – *Corpora e Lessici di Italiano Parlato e Scritto* (Albano Leoni & Giordano, 2005). I *corpora* raccolti sono stati sottoposti ad analisi spettro-acustica mediante il software *Wavesurfer* versione 1.8.1. Per quanto concerne l'analisi spettro-acustica del parlato L1, per ciascun parlante sono stati misurati: il numero delle catene foniche; il numero delle sillabe di ciascuna catena fonica; la durata delle singole catene foniche; la durata delle pause silenti; la durata delle pause non silenti denominate anche disfluenze; il valore massimo e minimo della *f*o di ciascuna catena fonica. A partire dalle misure rilevate, sono stati calcolati i seguenti indici prosodici per ciascun parlante: la velocità di articolazione, intesa come rapporto tra numero delle sillabe e durata delle catene foniche (sill/s); la velocità di eloquio, intesa come rapporto tra numero delle sillabe e tempo dell'enunciato (sill/s); la fluenza, intesa come rapporto fra il numero delle sillabe e il numero delle catene foniche (sill/CF); la durata media delle pause silenti (s); la durata delle disfluenze[3] in valore percentuale; il *range* tonale, inteso come la differenza tra il valore massimo e minimo della *f*o dell'enunciato, calcolato in semitoni (st), in quanto tale unità di misura consente di rendere confrontabili i dati relativi a parlanti diversi.

3.4. I partecipanti nativi e non nativi

Dall'analisi delle schede socio-biografiche somministrate ai partecipanti al progetto di ricerca, risulta che: gli italiani sono studenti universitari, prevalentemente di sesso femminile (45 donne e 6 uomini), appartenenti a una fascia di età compresa tra i 20 e i 25 anni e di provenienza campana; i cinesi sono studenti della Facoltà di Lingua e Letteratura Italiana presso l'Università di Tianjiin in Cina, prevalentemente di sesso femminile (13 donne e 4 uomini) e tutti appartenenti a una fascia di età compresa tra i 20 e i 25 anni. Al momento della ricerca, risiedevano da quattro mesi a Napoli e avevano un livello di italiano pari al B2 del Quadro Comune di Riferimento per le Lingue (QCER) .

[3] Le disfluenze sono nella maggior parte dei casi pause di natura non intenzionale e sono considerate o come momenti di pianificazione – in cui il parlante deve operare delle scelte – oppure come un segnale che il parlante lancia all"interlocutore allo scopo di mantenere il turno conversazionale. Le disfluenze sono altresì spia di insicurezza o titubanza da parte del parlante, che colma il tempo fra una parola e l"altra mediante pause non silenti.

Sia per gli italiani, sia per i cinesi, è stato necessario, in momenti separati ma paralleli, svolgere una fase di pre-test che preparasse i parlanti a svolgere i *debating* ufficiali nel miglior modo possibile, dal momento che la competenza argomentativa, in L1 e in L2, è un'abilità linguistico-comunicativa, che necessita di una preparazione preliminare. Pertanto, prima dei *debating* veri e propri, sono state spiegate a entrambi i gruppi le regole che caratterizzano i *debating*, sono stati forniti suggerimenti su come prepararsi ai dibattiti lavorando sia individualmente, sia in gruppo. Inoltre, una buona parte di questa fase introduttiva del lavoro è stata dedicata a commentare i parametri su cui i parlanti sarebbero stati giudicati: persuasività, volume, velocità, pause, intonazione, postura e sguardo, gestualità, competenza e uso della lingua. Successivamente, sono state effettuate diverse simulazioni di *debating*. Nel caso degli apprendenti sinofoni, ci si è avvalsi del supporto di un manuale mirato allo sviluppo delle competenze argomentative in italiano L2 (Barki & Diadori, 1997).

4. LA PERCEZIONE DELLA PERSUASIVITÁ NEL *CORPUS* IN L1

Il corpus in italiano L1 è costituito da un *debating* fra parlanti nativi (PN) di fronte a un pubblico di ascoltatori nativi (AN). Il confronto si è svolto fra due squadre che hanno argomentato, l'una a favore, l'altra contro, il seguente tema: "È bello vivere in una grande città". La squadra pro era composta da 1 uomo e 3 donne, la squadra contro da 4 donne. Il pubblico, misto, costituito da 19 ascoltatori nativi, è stato inviato a giudicare la persuasività dei singoli oratori in termini di "positivo" o "negativo".

Per indagare il piano percettivo, il grado di persuasività di ciascun parlante è stato messo in relazione con gli indici prosodici realizzati nel proprio parlato, al fine non solo di verificare l'esistenza di un legame fra grado di persuasività e prosodia, ma anche di reperire una costante in tale relazione. Le relazioni più significative sono state riscontrate fra persuasività e velocità di articolazione, fluenza, durata media delle pause silenti, disfluenze.

Le figure 1, 2 e 3 mostrano le linee di tendenza illustranti il rapporto tra il grado di persuasività raggiunto da ciascun parlante e la velocità di articolazione (VdA), la fluenza e la durata media delle pause silenti:

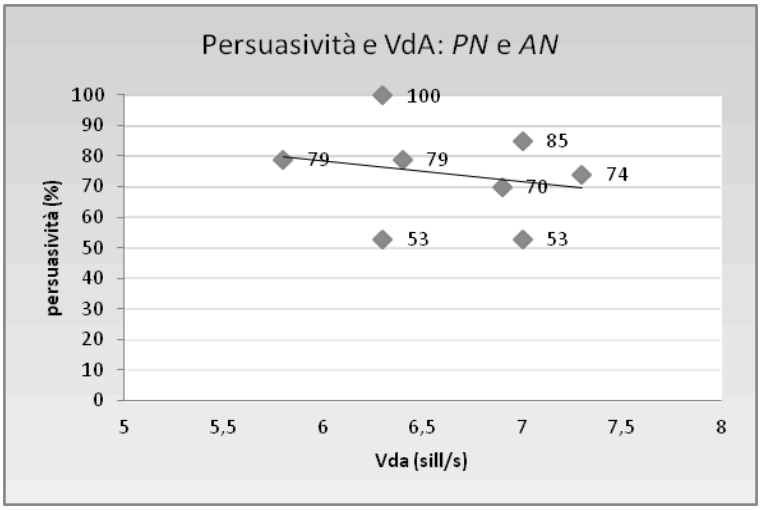


Figura 1: VdA

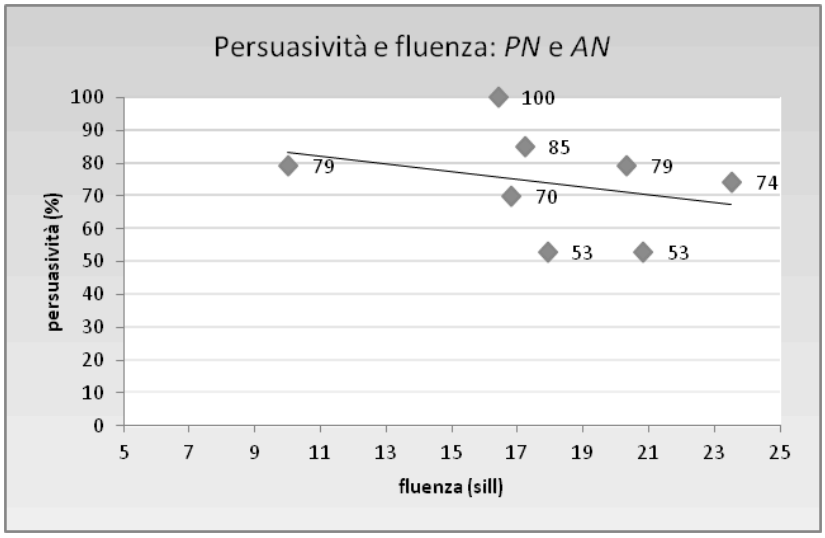


Figura 2: Fluenza

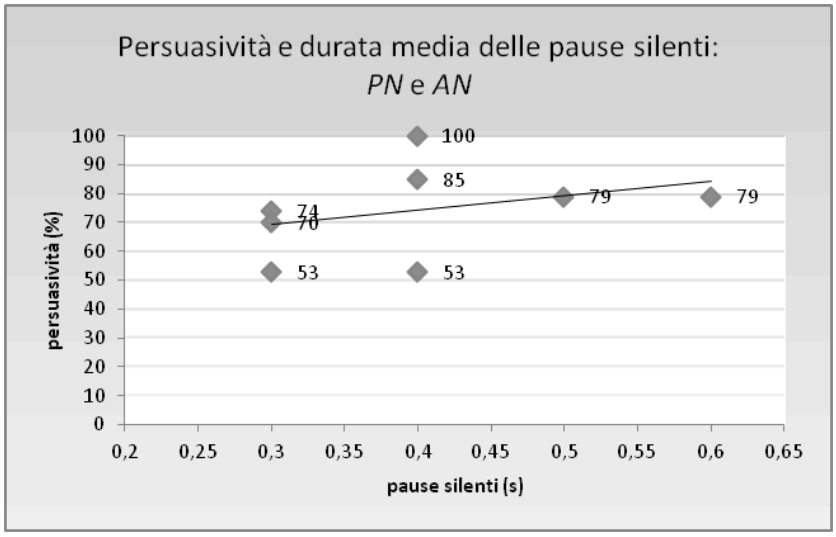


Figura 3: Durata media delle pause silenti

I tre grafici mostrano che, rispetto al discorso argomentativo dei parlanti italiani, l'ascoltatore italiano è più propenso ad accettare tesi pronunciate con maggiore accuratezza articolatoria e in cui siano presenti molte catene foniche di breve durata e pause silenti di media-lunga durata, che possano dare all'ascoltatore il tempo di riflettere su quanto detto: più il parlante italiano pro-duce silenzi di lunga durata, più l'ascoltatore nativo lo percepisce come maggiormente persuasivo.[4] La figura 4, infine, illustra come all'aumentare delle disfluenze diminuisca la capacità persuasiva del parlante nativo, come se l'ascoltatore nati-

[4] Sul senso di autorevolezza del silenzio, si veda Bazzanella, 2002.

vo percepisse quali elementi di disturbo le pause non silenti utilizzate per riempire gli spazi durante l'eloquio:

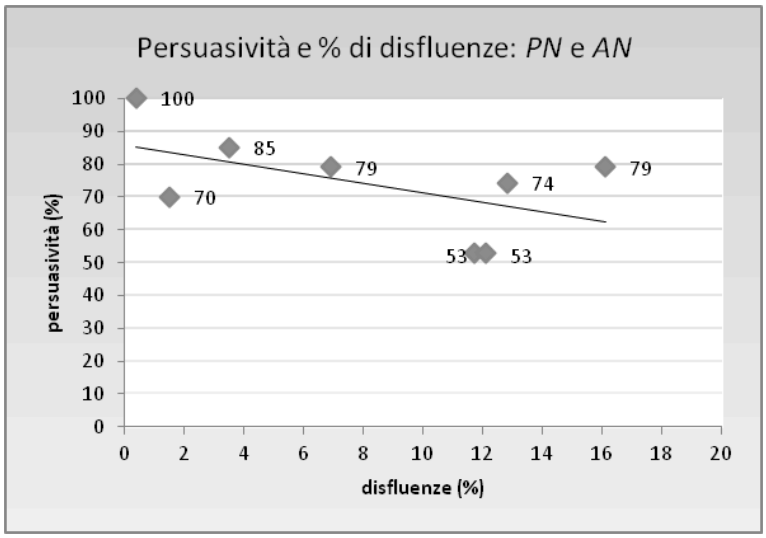


Figura 4: Percentuale di disfluenze

Non sono state invece riscontrate relazioni significative fra persuasività, velocità di eloquio e range tonale.

6. LA PERCEZIONE DELLA PERSUASIVITÁ NEL *CORPUS* IN L2

Il corpus in italiano L2 è composto dal *debating* fra parlanti non nativi (PNN) davanti ad ascoltatori non nativi (ANN), tutti di nazionalità cinese. Secondo la procedura stabilita, hanno preso parte al *debating* in L2 due squadre: la squadra pro era costituita da 5 cinesi – di cui 1 uomo e 4 donne – mentre la squadra contro era formata da 3 uomini e 2 donne. Al fine di annullare, per quanto possibile, la variabile semantica, l'argomento assegnato è stato il medesimo di quello su cui hanno discusso i parlanti nativi nel *debating* in L1, ossia "È bello vivere in una grande città".

In questo paragrafo, si procederà a indagare la relazione tra gli indici prosodici realizzati da locutori non nativi e la capacità di questi ultimi di persuadere degli ascoltatori non nativi. Emergerà la medesima relazione riscontrata nel *debating* in L1? Come avranno percepito gli ascoltatori non nativi i loro pari, mentre parlavano in una lingua straniera? Come avranno colto gli elementi prosodici della velocità, della fluenza, delle pause e dell'intonazione?

Gli ascoltatori non nativi hanno attribuito valutazioni estremamente positive a tutti i locutori (in media, 84% di giudizi positivi). Sarebbe, pertanto, inutile, soffermarsi sulle relazioni esistenti fra la persuasività e i singoli indici prosodici, poiché la linea di tendenza – ascendente, discendente o piatta - riguarderebbe sempre valutazioni molto alte che non sono, dunque, influenzate, dalla prosodia. Vale la pena, invece, soffermarsi su come la relazione fra persuasività e prosodia sia influenzata dalla capacità percettiva dell'ascoltatore. Infatti, se paragoniamo le valutazioni che gli ascoltatori italiani hanno dato ai locutori nativi nel *corpus* in L1, con le valutazioni che i non nativi hanno dato ai parlanti non nativi nel

corpus in L2, i dati emersi rivelano che, mentre i parlanti nativi percepiscono una netta relazione fra persuasività e correlati prosodici, la competenza percettiva degli apprendenti cinesi non sembra rilevare una relazione significativa fra prosodia e capacità di persuasione. Rispetto a questa differenza, siamo stati indotti a supporre l'esistenza di due cause concomitanti. Da un lato, l'attribuzione di un giudizio sulla persuasività implica un processo costituito da quattro fasi: l'ascolto del discorso, la comprensione acustica, ovvero la decodifica del messaggio, il confronto del messaggio con le proprie opinioni personali e, infine, il giudizio. Rispetto a questo processo, l'apprendente non nativo sarebbe così concentrato a decodificare e comprendere le parole del messaggio ascoltato che presterebbe più attenzione alle parole ascoltate piuttosto che all'argomentazione in sé, a differenza del nativo, che ha gli strumenti tali da poter passare alle fasi successive. La percezione del parlato in L2, infatti, è fortemente influenzata dalla struttura prosodica della lingua materna, che può influenzare in modo incisivo la capacità dell'apprendente di orientarsi sull'ascolto. In tal senso, gli apprendenti cinesi, che hanno una lingua materna caratterizzata da strutture ritmico-intonative distanti dall'italiano, accederebbero alla percezione del parlato con enorme difficoltà, non riuscendo a percepire e segmentare la sequenza parlata in modo efficace.

Dall'altro lato, si innescano meccanismi di natura sociolinguistica e culturale idiosincratici dei parlanti non nativi: da questa prospettiva, i cinesi valuterebbero positivamente i propri compagni per premiare lo sforzo e l'impegno nell'aver affrontato un *task* in un'altra lingua. Lo sviluppo della capacità percettiva in L2 sarebbe, dunque, rallentato a causa dei diversi modelli di comportamento pragmatico-comunicativo degli apprendenti, che sono distanti da quelli della cultura della lingua *target*.

La combinazione di questi due elementi – l'uno, di natura cognitiva, l'altro di natura socio-linguistica – determinerebbe la difficoltà, a livello percettivo, di discriminare le componenti soprasegmentali di un discorso e la loro valenza pragmatica, da parte di non nativi. [5]Ciò risulta ancora più interessante, se consideriamo che il QCER[6], rispetto alla capacità di comprensione orale dell'apprendente B2, indica che "è in grado di comprendere le idee fondamentali di testi complessi su argomenti sia concreti sia astratti, comprese le discussioni tecniche nel proprio settore di specializzazione".

Dal nostro punto di vista, considerando i risultati ottenuti dalla nostra ricerca, è possibile aggiungere che un apprendente di italiano L2, in possesso del cosiddetto livello di autonomia di conoscenza della lingua, è in grado di percepire e decodificare messaggi complessi, ma è meno capace di valutarli in termini di efficacia persuasiva. I dati ricavati gettano nuova luce sugli studi riguardanti la competenza percettiva da un punto di vista acquisizionale e sull'abilità di comprensione orale rispetto ad alcune tipologie testuali; inoltre, rivelano una certa carenza di attenzione verso la dimensione prosodica della comunicazione in L2 sia nel campo dell'acquisizione, sia nel campo della didattica, sia, infine, nell'ambito della valutazione, vista l'assenza di riferimenti agli aspetti soprasegmentali della lingua nei descrittori del QCER.

[5] Si potrebbe ipotizzare anche che sia la prosodia prodotta dai non nativi in L2 a non essere particolarmente informativa per gli ascoltatori non nativi.
[6] Cfr. http://www.coe.int/t/dg4/linguistic/CADRE_EN.asp

7. LA PERCEZIONE DELLA PERSUASIVITÁ NEL *CORPUS* IN L1/L2

Il *corpus* in L1/L2 è costituito da un *debating* fra una squadra pro, costituita da 4 italiani (1 uomo e 3 donne) e una squadra contro, formata da 4 cinesi (2 uomini e 2 donne). Il *debating* si è svolto in italiano, che ha rappresentato, pertanto, la L1 per i partecipanti nativi e la L2 per quelli non nativi, davanti a un pubblico, misto, costituito da 33 ascoltatori, di cui 20 italiani e 13 cinesi. L'argomento del *debating* è stato: "È lecito usare parole e immagini provocatorie nelle pubblicità".

Rispetto a quanto fatto precedentemente, in questo paragrafo si esporranno i risultati relativi a più piani percettivi: a) parlanti nativi valutati da ascoltatori nativi; b) parlanti non nativi valutati da ascoltatori nativi; c) parlanti nativi giudicati da ascoltatori non nativi; d) parlanti non nativi giudicati da ascoltatori non nativi.

La valutazione media attribuita dai cinesi agli italiani è del 100%, quella attribuita dai cinesi ai locutori cinesi è del 96 %. Tali risultati avvalorano le considerazioni formulate nel precedente paragrafo: in questo *debating*, inoltre, emerge come l'ascoltatore cinese premi non solo il parlante cinese per lo sforzo compiuto nell'argomentare in una lingua straniera, ma anche, da un punto di vista socio-linguistico, il modello del *native speaker*. Pertanto, per le motivazioni sopra esposte, appare superfluo soffermarci sull'analisi della relazione fra persuasività e prosodia dal punto di vista dell'ascoltatore straniero. Procederemo, dunque, a condurre un'analisi relativa alle coppie: a) parlante nativo/ascoltatore nativo; b) parlante non nativo/ascoltatore nativo.

Per quanto riguarda i giudizi di persuasività assegnati dagli ascoltatori nativi ai parlanti nativi, si può osservare che all'aumentare della velocità di articolazione (figura 5), aumenta anche la percentuale di persuasività. Rispetto al *debating* in L1, si ha un capovolgimento dei risultati: nel dibattito fra soli italiani, infatti, gli ascoltatori italiani avevano ritenuto più convincente un parlato più accurato da un punto di vista articolatorio. Qui, invece, gli italiani risultano maggiormente persuasi da un parlato tendente all'iperarticolato:

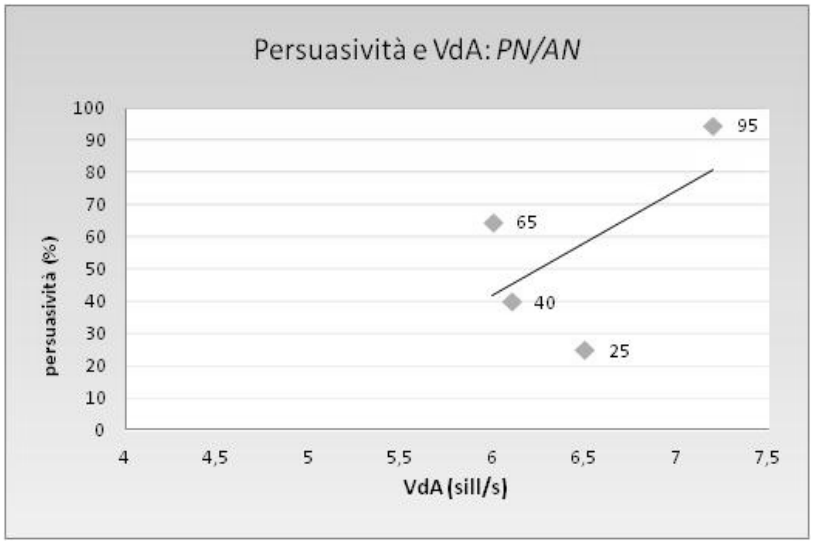


Figura 5: VdA

La disparità tra la percezione dei nativi nel *debating* in L1 e quella nel dibattito interculturale può essere spiegata probabilmente con il fatto che i nativi, percependo il parlato degli stranieri esasperatamente iperarticolato (figura 10), trasferiscono tale percezione negativa sui parlanti nativi, da cui, di conseguenza, ci si aspetta un parlato meno articolato e più veloce. Per quanto concerne la correlazione tra correlazione tra fluenza, durata delle pause silenti e persuasività, nel *debating* interculturale, i parlanti nativi valutano più persuasivi gli italiani che producono catene foniche più lunghe (figura 6) e pause silenti più brevi (figura 7), contrariamente a quanto avviene nel corpus in L1 (figure 2 e 3).

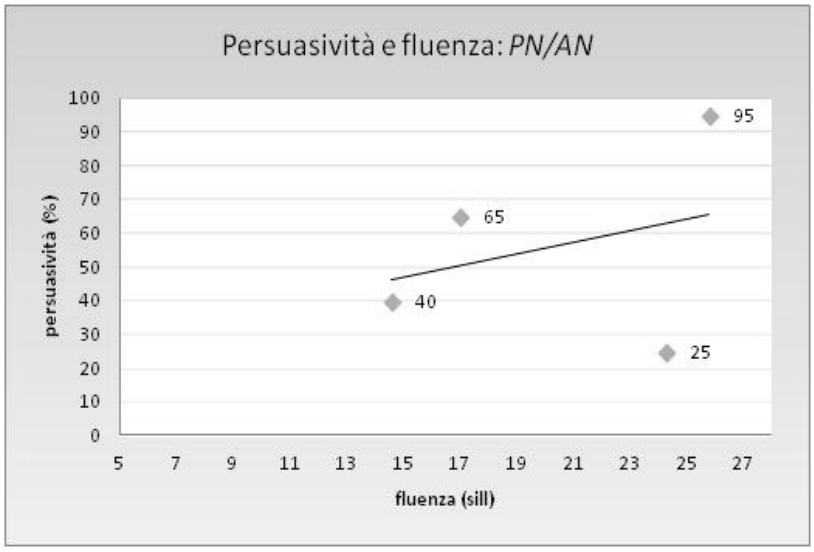


Figura 6: Fluenza

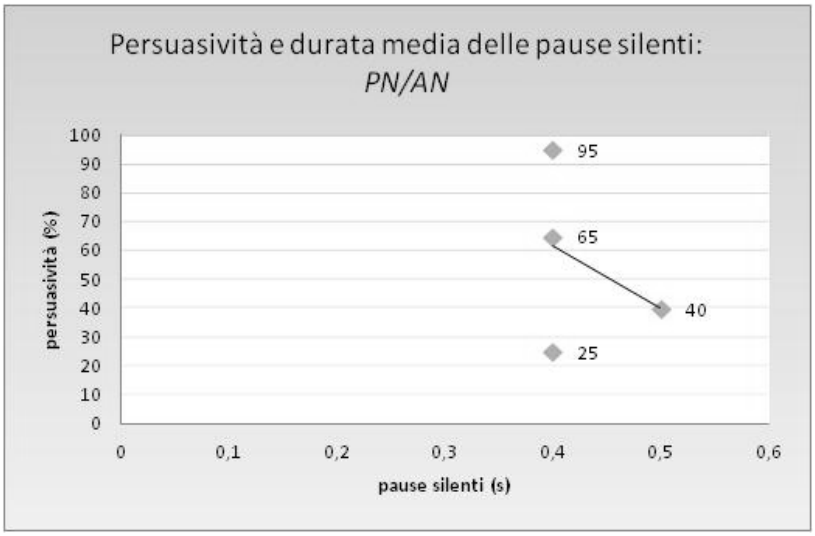


Figura 7: Durata media delle pause silenti

Per quanto riguarda le disfluenze, esse risultano essere inversamente proporzionali alla persuasività anche nel *debating* in L1/L2, confermandosi come elementi di disturbo, usati non intenzionalmente dal locutore, ma percepiti come fastidiosi all'orecchio dell'ascoltatore (figura 8):

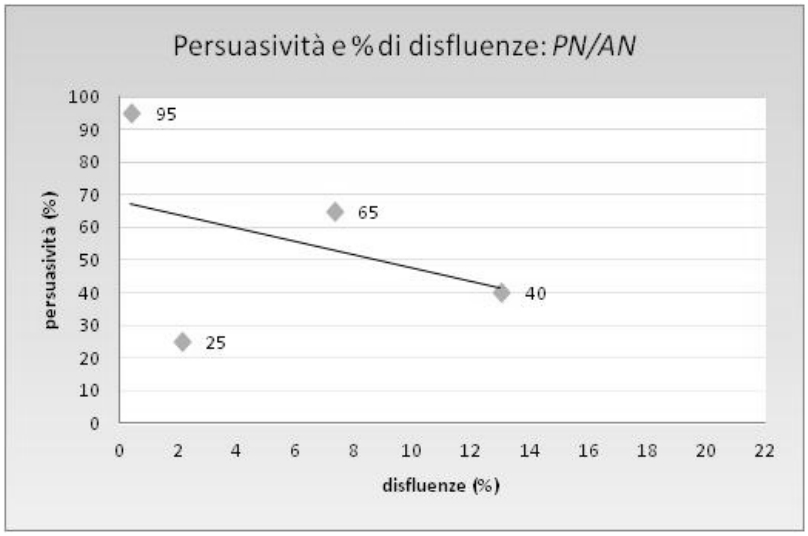


Figura 8: Percentuale di disfluenze

Come nel *corpus* in L1, anche nel corpus in L1/L2, non si riscontrano relazioni significative fra persuasività e velocità di eloquio. Invece, è stata notata una tendenza ad attribuire maggiore persuasività a un parlato poco variato tonalmente (figura 9):

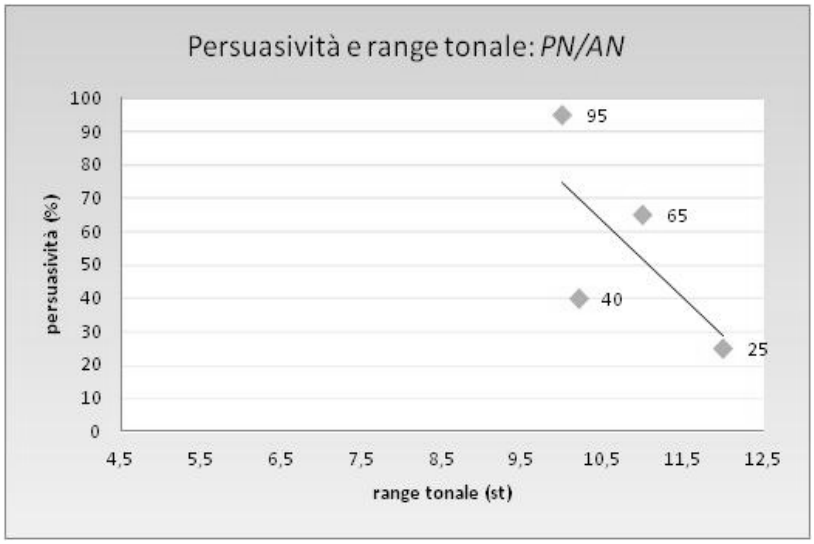


Figura 9: *Range* tonale

Per quanto concerne il rapporto fra persuasività e velocità di articolazione, relativamente ai parlanti non nativi valutati dagli italiani, i risultati emersi confermano lo stesso andamento riscontrato nella coppia parlante nativo/ascoltatore nativo (figura 10):

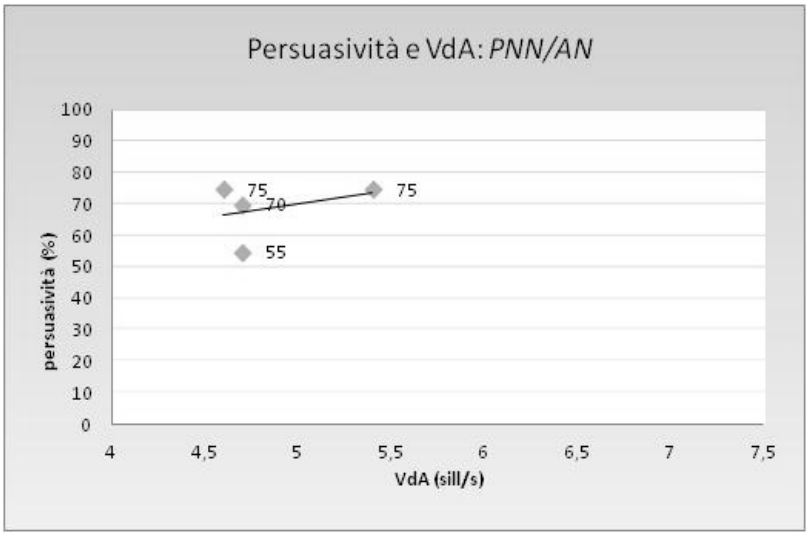


Figura 10: VdA

Il grado di persuasività, nella valutazione parlante non nativo/ascoltatore nativo, aumenta quando vengono prodotte brevi catene foniche (figura 11) e lunghi silenzi (figura 12). Dunque, gli ascoltatori nativi, assistendo al confronto fra nativi e stranieri, .cambiano completamente la loro valutazione, rispetto al parlato in L1.

Una possibile spiegazione è imputabile, ancora una volta, all'ascolto simultaneo di voci native e non native:

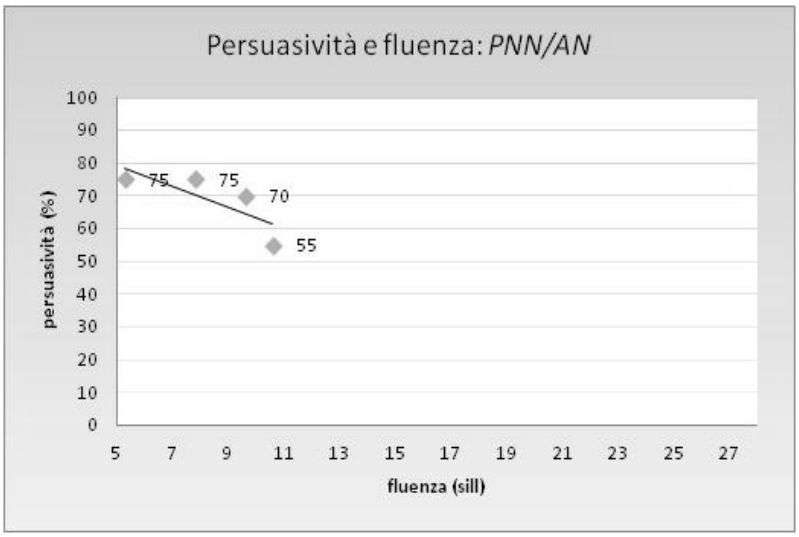


Figura 11: Fluenza

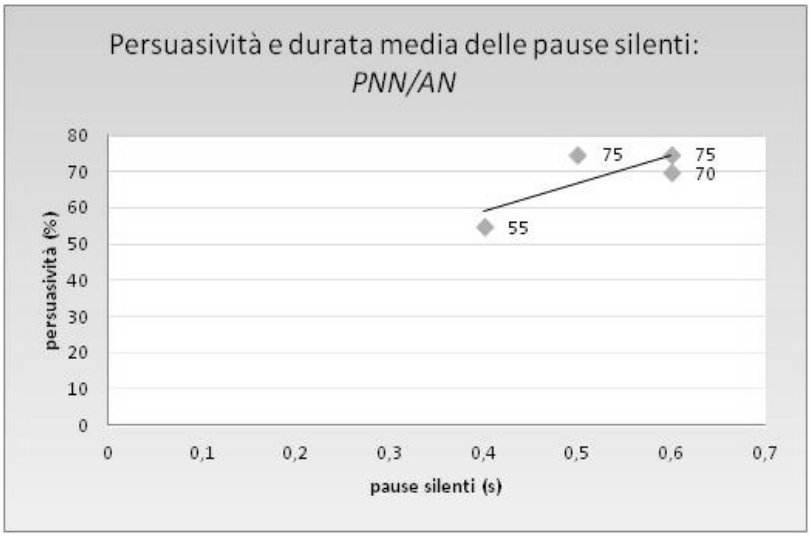


Figura 12: Durata media delle pause silenti

Parimenti a quanto emerso nel confronto fra parlanti e ascoltatori italiani, anche nella coppia parlante straniero/ascoltatore italiano, la percentuale di disfluenze viene percepita come un elemento di disturbo rispetto alla persuasività (figura 13). Lo stesso accade con il range tonale: ancora una volta il parlato poco enfatico, più calmo, sembra convincere di più (figura 14).

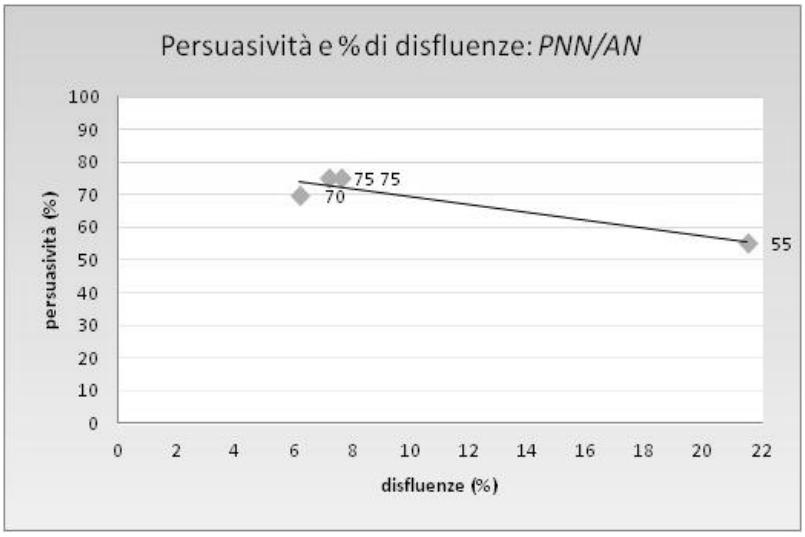


Figura 13: Percentuale di disfluenze

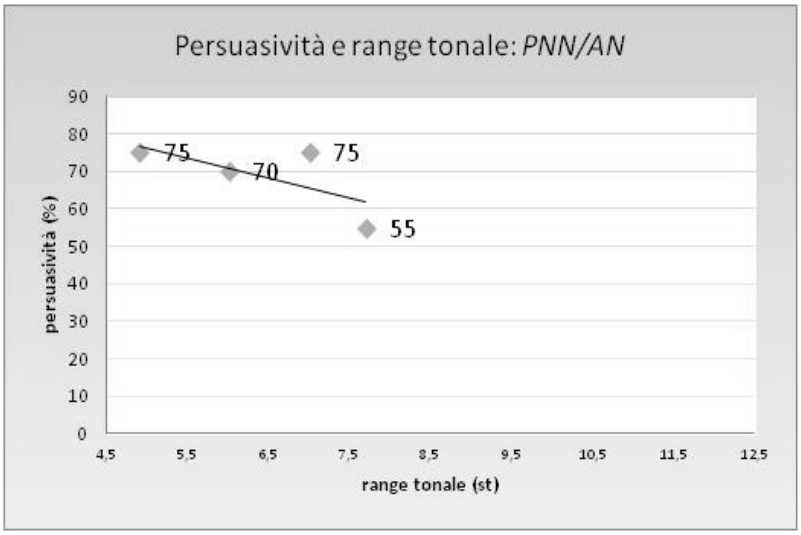


Figura 14: *Range* tonale

Ancora una volta non emerge alcuna particolare relazione fra persuasività e velocità di eloquio nel parlato spontaneo. In una comparazione complessiva dei tre *corpora*, è stato riscontrato un adeguamento della capacità percettiva in relazione alla persuasività, per quanto concerne la velocità di articolazione e i silenzi. Risulta, infatti, che l'ascolto di parlanti italiani e stranieri, che si sono susseguiti a turno esponendo le proprie argomentazioni dinnanzi a un pubblico altrettanto misto, ha determinato un adattamento della percezione degli ascoltatori: l'eccessiva articolazione degli stranieri ha influito sul giudizio degli ascoltatori italiani, che in tutti in casi hanno valutato negativamente una bassa velocità di articolazione. La minore fluenza e la maggiore durata dei silenzi, invece, ritenuti importanti nel *debating* in L1, sono stati considerati dagli ascoltatori italiani come fattori negativi se prodotti dai nativi, probabilmente influenzati dall'eloquio dei non nativi, e come fattori positivi, in quanto funzionali alla comprensione e, dunque, al raggiungimento dello scopo di convincere, se prodotti dai non-nativi. È importante sottolineare, anche, che dalla parte degli ascoltatori non nativi non emerge, ancora una volta, la capacità percettiva di discriminare i tratti soprasegmentali del parlato in quanto il loro giudizio non muta, sia che ascoltino solo parlanti non nativi, sia che ascoltino un'interazione fra italiani e stranieri. Infine, il range tonale, il cui rapporto con la persuasività era risultato poco significativo nel *debating* in L1, nel *debating* in L1/L2, invece, è risultato inversamente proporzionale alla persuasività, poiché il parlato pacato e monotòno degli italiani e dei cinesi ha convinto di più gli ascoltatori nativi.

La produzione prosodica, dunque, incide sensibilmente sulla percezione, il che potrebbe determinare un notevole impatto sulle modalità di sviluppo della competenza prosodica in una seconda lingua, sui contesti di apprendimento linguistico e, infine, sulla comunicazione interculturale.

8. CONCLUSIONI

La presente ricerca si è posta l'obiettivo di analizzare la percezione dei tratti ritmico-prosodici in apprendenti d'italiano L2, in un'ottica comparativa con l'italiano L1,

nell'ambito del parlato argomentativo spontaneo. Nel perseguire tale scopo, si è tentato di delineare una relazione fra gli indici prosodici usati dal parlante e la sua capacità persuasiva: è emerso che gli ascoltatori italiani reputano più persuasivo un parlato in L1 e L2 che sia ben articolato, con brevi catene foniche,pause silenti di lunga durata e poche disfluenze. Tali dati relativi al parlato spontaneo confermano le ricerche condotte da De Meo et al. (2011) relativamente al parlato letto.

La percezione della persuasività negli italiani cambia nel momento in cui ascoltano un dibattito fra italiani e cinesi: probabilmente influenzati dal parlato degli stranieri, lento, iperarticolato, tonalmente molto piatto, l'ascoltatore italiano tende ad attribuire maggiore persuasività a un parlato in L1 che sia veloce, con pochi silenzi e disfluenze, e molto variato dal punto di vista intonativo. Invece, per quanto riguarda gli stranieri, per il parlato argomentativo spontaneo non sono state riscontrate relazioni significative fra persuasività e prosodia, dal momento che i cinesi hanno attribuito valutazioni sempre altamente positive che non consentono di individuare una linea di tendenza che leghi le suddette variabili. Anche in questo caso, possiamo avanzare alcune riflessioni. Se sul piano della produzione non sono state riscontrate differenze al livello soprasegmentale, sul piano della percezione emergono comportamenti linguistici diversi, motivabili con diverse spiegazioni. La prima, di carattere sociolinguistico e pragma-culturale, riguarda l'adeguamento dell'ascoltatore a modelli percettivi diversi: nel caso degli italiani, ciò è accertato dal diverso giudizio di valutazione che conferiscono agli italiani a seconda se essi interagiscono con altri italiani o con cinesi; nel caso dei cinesi, si tratta di un'ipotesi che potrebbe essere verificata applicando la stessa procedura di ricerca a *debating* in cinese L1 ed esaminando le valutazioni del pubblico rispetto ai diversi tipi di interazione.

Un'altra spiegazione è di natura cognitiva: se, a causa di ostacoli legati all'apprendimento linguistico, gli stranieri non riuscissero a oltrepassare la fase della decodifica e comprensione del messaggio, senza giungere al confronto con la propria opinione personale e all'attribuzione di un giudizio, allora è ipotizzabile la presenza di un'interlingua anche sul piano percettivo in relazione a testi diversi. Se ciò fosse approfondito con ulteriori ricerche, ci sarebbero forti ricadute in ambito glottodidattico, dal momento che i testi audio somministrati agli apprendenti dovrebbero essere costruiti, adattati e scelti non solo in base a strutture morfosintattiche e funzioni linguistiche, ma anche in virtù dei vari livelli di competenza percettiva che lo straniero sviluppa in L2. Infine, sarebbe interessante non solo allargare lo studio della persuasività ai suoi legami con le variabili testuale e cinesica, ma anche approfondire quello con la prosodia tentando di annullare le altre variabili attraverso la tecnica del low-filtering.[7] In questo modo si potrebbe calcolare in quale percentuale la prosodia incida sulla persuasività.

BIBLIOGRAFIA

Aoyama, K., Guion S. (2007), Prosody in second language acquisition. Acoustic analysis of duration and F0 range, in Bohn-Munro, Language Experience in Second Language Speech Learning, (ed. by Murray J. Bohn-Munro), Amsterdam-Philadelphia: Benjamins, 281-297.

[7] Il low-filtering è una tecnica che consiste nel cancellare tutte le componenti del segnale che si trovano al di sopra di una determinata frequenza, detta frequenza di taglio (ft). Il risultato, sul piano percettivo, è che l'ascoltatore è in grado di distinguere il suono ma non il testo (Pettorino et alii, in corso di stampa).

Albano Leoni, F., Giordano, R. (a cura di) (2005), Italiano parlato. Analisi di un dialogo, Napoli: Liguori.

Barki Coricelli, P., Diadori, P. (1997), Pro e contro. Conversare e argomentare in italiano. Livello intermedio. Guida per l'insegnante, Roma: Bonacci.

Bazzanella, C. (2002), Sul dialogo. Contesti e forme di interazione verbale, Milano: Guerini.

Birdsong, D. (1999), Second language acquisition and the Critical Period Hypothesis, Mahwah, NJ: Erlbaum.

Burnham, D., Mattock, K. (2007), The perception of tones and phones, in Language Experience in Second Language Speech Learning (ed. by Murray J. Bohn-Munro), Amsterdam-Philadelphia: Benjamins, 259-280.

Chini, M. (2010), Concetti, fenomeni e fattori relativi all'acquisizione di lingue seconde, in Italiano di Cinesi, Italiano per Cinesi, dalla prospettiva della didattica acquisizionale (a cura di S. Rastelli), Perugia: Guerra, 23-43.

Costamagna, L. (2011), L'apprendimento della fonologia dell'italiano da parte di studenti sinofoni: criticità e strategie, in La didattica dell'italiano a studenti cinesi: il programma Marco Polo ed altre esperienze, Atti del XV seminario AICLU (a cura di E. Bonvino, S. Rastelli), Pavia: Pavia University Press, 49-66.

Chun, D.M. (2002), Discourse Intonation in L2, Amsterdam-Philadelphia: Benjamins.

De Meo, A., Pettorino, M. (2011), L'acquisizione della competenza prosodica in italiano L2 da parte di studenti sinofoni, in La didattica dell'italiano a studenti cinesi: il programma Marco Polo ed altre esperienze, Atti del XV seminario AICLU (a cura di E. Bonvino, S. Rastelli), Pavia: Pavia University Press, 67-78.

De Meo, A., Pettorino, M., Vitale, M. (2012), Non ti credo: i correlati acustici della credibilità in italiano L2, in Competenze e formazione linguistiche. In memoria di Monica Berretta, Atti dell'XI Congresso dell'Associazione Italiana di Linguistica Applicata (a cura di G. Bernini, C. Lavinio, A. Valentini, M. Voghera), Perugia: Guerra Edizioni, 229-248.

Flege, J.E. (1987), The Production of "New" and "Similar" Phones in a Foreign Language: Evidence for the Effect of Equivalence classification, Journal of Phonetics, 15, 47-65.

Giacalone Ramat, A. (1994), Il ruolo della tipologia linguistica nell'acquisizione di lingue seconde, in Italiano: lingua seconda/lingua straniera, Atti del XXVI Congresso internazionale della Società di linguistica italiana (a cura di A. Giacalone Ramat & M.Vedovelli) Roma: Bulzoni, 27-43.

Gut, U. (2009), Non-Native Speech: A Corpus-Based Analysis of Phonological and Phonetic Properties of L2 English and German, Frankfurt: Peter Lang.

Pettorino, M., Pellegrino, E., Salvati, L., Vitale, M., De Meo, A., (in corso di stampa), La voce dei media. Trapianti ritmico-intonativi per un'analisi diacronica dell'italiano parlato, in Coesistenze linguistiche nell'Italia pre- e postunitaria, Atti del XLV Congresso Internazionale della Società di Linguistica Italiana, Aosta/Bard/Torino, 26-28 settembre2011.

CARATTERISTICHE TEMPORALI DEL PARLATO ITALIANO E TEDESCO: UN CONFRONTO TRA PARLANTI NATIVI, BILINGUI E NON-NATIVI

Stephan Schmid & Volker Dellwo
Phonetisches Laboratorium der Universität Zürich
schmidst@pholab.uzh.ch, volker.dellwo@uzh.ch

1. RIASSUNTO

Il presente contributo analizza alcune caratteristiche temporali di due lingue tradizionalmente assegnate a 'classi ritmiche' diverse: l'italiano, di solito classificato come 'isosillabico' (*syllable-timed*) e il tedesco, spesso considerato come 'isoaccentuale' (*stress-timed*). Da un lato si tratta di applicare a nuovi dati delle 'metriche' sviluppate nell'ultimo decennio, al fine di verificare la cosiddetta 'ipotesi delle classi ritmiche'; dall'altro lato si cerca di andare un passo oltre la ormai consistente letteratura sull'argomento, tenendo conto anche di altri fattori quali la velocità di elocuzione e, soprattutto, considerando tre tipi diversi di parlanti: nativi, bilingui e non-nativi.

Le ricerche che hanno applicato delle metriche ritmiche al parlato di locutori non-nativi (apprendenti di L2) hanno rilevato sia una specie di 'ritmo intermedio', sia un generale aumento delle durate vocaliche dovuto a fenomeni di esitazione. Invece, gli studi sul ritmo nei parlanti bilingui 'precoci' non sono molto numerosi, ma permettono lo stesso di formulare due ipotesi contrastanti: i) i bilingui parlano con un ritmo diverso nelle due lingue (si comportano cioè come i rispettivi parlanti monolingui), oppure ii) i bilingui si collocano nello spazio ritmico in una posizione intermedia tra parlanti nativi e non-nativi.

Al fine di verificare tali ipotesi è stato allestito, presso il laboratorio di fonetica dell'Università di Zurigo, un corpus di parlato italiano e tedesco: 5 studenti bilingui, 5 studenti italofoni e 5 studenti tedescofoni hanno letto 10 frasi in ciascuna delle due lingue. Gli audio-file delle registrazioni sono stati segmentati in intervalli vocalici e consonantici, onde poter calcolare una serie di metriche ritmiche. I risultati principali forniscono indicazioni contrastanti a più livelli.

Dal punto di vista generale della tipologia ritmica delle lingue si profilano alcune tendenze che sono emerse in studi precedenti: i valori ricavati dalle metriche %V, ΔC, nPVI-V e *%Voiced* confermano in linea di massima l'ipotesi delle classi ritmiche e in particolare il carattere più 'sillabico' dell'italiano di fronte al carattere più 'accentuale' del tedesco. Invece, l'applicazione di una nuova metrica che calcola il rapporto di durata tra sillabe toniche e atone fornisce un elemento contrario alla tradizionale tipologia ritmica, dato che nei nostri dati tale rapporto risulta essere maggiore in italiano che non in tedesco.

Per quanto riguarda la differenziazione dei tre tipi di parlanti in base alle caratteristiche temporali del parlato letto, il rapporto di durata tra sillabe toniche e atone fornisce evidenza a favore della prima ipotesi summenzionata, in quanto i bilingui si comportano in ambedue le lingue come i rispettivi parlanti monolingui. Considerando altri fattori sono però emersi numerosi indizi che depongono a favore della seconda ipotesi, dato che i bilingui si trovano in una posizione intermedia per una serie di parametri tra cui la velocità di eloquio, la variabilità delle durate di intervalli vocalici e la percentuale degli intervalli sonori.

2. INTRODUZIONE

In questo capitolo introduttivo ripercorreremo brevemente alcune tappe della ricerca sul ritmo linguistico. Partendo dalla falsificazione della tradizionale 'ipotesi dell'isocronia' accenneremo alla sua riformulazione come 'ipotesi delle classi ritmiche' e, in particolare, ad alcune delle metriche che sono state elaborate per render conto dei correlati acustici di tale tipologia fonologica (2.1.). Particolare attenzione sarà rivolta alle ricerche empiriche che hanno applicato questi algoritmi alle due lingue che ci interessano in questa sede, ovvero all'italiano e al tedesco (2.2.); riporteremo anche i risultati di alcuni studi che hanno esaminato con questa metodologia il ritmo di apprendenti di L2 (2.3.) e di soggetti che sono cresciuti sin dalla loro infanzia con due lingue (2.4.)[1].

2.1. Dall'ipotesi dell'isocronia all'ipotesi delle classi ritmiche

Com'è noto, la fonetica linguistica del ventesimo secolo ha per lungo tempo sostenuto la cosiddetta 'ipotesi dell'isocronia' proposta da Pike (1945) e da Abercrombie (1967). Secondo tale ipotesi le lingue del mondo possono essere suddivise in due o tre grandi tipi, ovvero in lingue ad isocronia sillabica (*syllable-timed*), accentuale (*stress-timed*) e morica (*mora-timed*). Nelle lingue isoaccentuali come l'inglese, l'unità fondamentale – che ricorre in intervalli di uguale durata – sarebbe costituita dal gruppo accentuale (in altre parole: dal piede metrico); invece, per le lingue isosillabiche – come ad esempio lo spagnolo – si assume come unità fondamentale la sillaba, che quindi ricorrerebbe in intervalli di uguale durata. È altresì noto che l'ipotesi dell'isocronia è stata falsificata empiricamente: in tutte le lingue esaminate, la durata delle sillabe dipende dal numero di segmenti che la compongono, così come la durata dei piedi è determinata dal numero delle sillabe (cfr. Bertinetto, 1989). Di conseguenza, alcuni autori hanno proposto che il ritmo linguistico derivi piuttosto dall'effetto congiunto di una serie di proprietà fonologiche quali la complessità delle strutture fonotattiche oppure il grado di riduzione delle sillabe atone (v. ad esempio Dauer, 1983; Bertinetto, 1989).

Un ritorno alla prospettiva fonetica è invece avvenuto intorno all'anno 2000 con la proposta – paradossalmente legata proprio alla reinterpretazione fonologica del ritmo – di adottare nuove metriche temporali che tengano conto della complessità della struttura sillabica e della riduzione vocalica. Segmentando il segnale acustico non più in sillabe e gruppi accentuali, ma in intervalli vocalici e consonantici, si possono calcolare tre 'metriche ritmiche' (*rhythm metrics*): i) %V ovvero la percentuale degli intervalli vocalici rispetto alla durata totale di un enunciato, ii) ΔC ovvero la deviazione standard delle durate degli intervalli consonantici, iii) ΔV ovvero la deviazione standard delle durate degli intervalli vocalici

[1] Com'è noto, la nozione di 'bilinguismo' può essere definita secondo vari criteri, a seconda che ci si attenga alla competenza o all'uso delle lingue in gioco. A scanso di equivoci occorre quindi precisare che in questo lavoro considereremo come bilingui non coloro che abbiano imparato una seconda lingua in età adolescente o adulta, ma piuttosto degli individui che hanno acquisito in modo spontaneo sin dalla loro primissima infanzia due lingue, usandole continuamente nella loro vita quotidiana. In linea di principio questa definizione secondo l'uso lascia aperta la possibilità di varie configurazioni della competenza bilingue, secondo le due ipotesi enunciate nel riassunto e ribadite in 2.4 – ovvero di un bilinguismo sia 'coordinato' che 'composto' (secondo la classica dicotomia proposta da Weinreich 1963[1974]).

(Ramus et alii, 1999). In particolare la combinazione di %V e ΔC ha permesso di distinguere le lingue tradizionalmente considerate come isoaccentuali da quelle isosillabiche. Come normalizzazione delle misure ΔC e ΔV è stato suggerito di utilizzare i coefficienti di variazione VarcoC e VarcoV piuttosto che le deviazioni standard, dato che le durate effettive degli intervalli consonantici e vocalici sono sensibili anche alla velocità di eloquio (Dellwo & Wagner, 2003; Dellwo, 2006; Ferragne & Pellegrino, 2004). Un'ulteriore modifica dell'approccio di Ramus et alii (1999) consiste nella suddivisione del segnale acustico non in intervalli vocalici e consonantici, bensì in parti periodiche e a-periodiche, cioè in intervalli 'sonori' o 'sordi' di cui si calcolano ad esempio le rispettive percentuali della durata totale degli enunciati (*%Voiced* e *%Unvoiced*; cfr. Dellwo et alii, 2007).

Le metriche considerate sinora possono essere denominate 'globali', in quanto si fondano su calcoli di statistica descrittiva (percentuali, deviazione standard e coefficiente di variazione) di tutti gli intervalli vocalici e consonantici segmentati in un determinato numero di enunciati. Un approccio alternativo considera invece il ritmo come un fenomeno piuttosto 'locale', calcolando attraverso il cosiddetto *Pairwise Variability Index* (PVI) la media delle differenze di durata tra coppie di intervalli vocalici e consonantici successivi; anche in questo caso, il PVI degli intervalli vocalici ha permesso di differenziare le lingue secondo le tradizionali classi ritmiche (Grabe & Low, 2002). Infine, un ulteriore sviluppo del PVI è stato fornito dal cosiddetto *Control and Compensation Index* (CCI) che tiene conto anche del numero di segmenti fonologici che compongono un determinato intervallo vocalico o consonantico (Bertinetto & Bertini, 2008).

Attualmente esistono quindi varie metriche ritmiche in concorrenza tra loro e il dibattito sul loro valore euristico è tuttora aperto (v. Mairano & Romano, 2010, per una rassegna generale e Barry, 2010, per una presa di posizione critica). Tuttavia il nostro obiettivo principale non è tanto argomentare a favore dell'una o dell'altra metrica né tantomeno discutere la fondatezza dell'ipotesi delle classi ritmiche *tout court*; ciononostante è d'uopo accennare almeno brevemente ad alcuni studi che hanno applicato tali metriche alle due lingue che ci interessano in questa sede.

2.2. Caratteristiche ritmiche dell'italiano e del tedesco

Com'è noto, l'italiano viene tradizionalmente annoverato tra le lingue ad isocronia sillabica (Bertinetto, 1977), benché non siano mancate riserve su questa classificazione (Vayra et alii, 1984). Tuttavia, anche nelle ricerche basate sulle metriche ritmiche l'italiano occupa spesso una posizione all'interno delle lingue tradizionalmente caratterizzate come *syllable-timed*, a partire dal lavoro pionieristico di Ramus et alii (1999), dove l'italiano si colloca nella stessa sfera delle altre lingue romanze quali il francese, lo spagnolo e il catalano, fino alle verifiche più recenti di Mairano & Romano (2007, 2010), i quali constatano comunque una lieve variazione interindividuale tra i due soggetti da loro esaminati. L'italiano non fa parte delle lingue esaminate da Grabe & Low (2002), ma un'analisi basata sulle metriche PVI ha ottenuto un quadro simile per le tre varietà regionali di Pisa, Napoli e Bari (Russo & Barry, 2010). Anche da un confronto tra l'italiano regionale siciliano e quello veneto, imperniato sulle metriche VarcoV e VarcoC, non sono scaturite differenze significative (White et alii, 2009). Infine, un esame di ben 15 varietà di italiano regionale permette a

Giordano & D'Anna (2010) di concludere che "%V values are generally consistent with isosyllabic languages" [2].

Per quanto riguarda invece il tedesco, va notato che le lingue germaniche vengono tradizionalmente assegnate alle lingue isoaccentuali, dato il forte peso che vi assumono l'accento di parola e la riduzione – a livello fonologico – delle vocali atone; notiamo che nella lingua tedesca è distintivo anche il contrasto tra vocali (toniche) lunghe e brevi. Nello studio di Ramus et alii (1999) il tedesco è assente, ma in studi successivi con le metriche %V, ΔC e ΔV questa lingua presenta tutto sommato dei valori simili a quelli dell'inglese (v. ad esempio Mairano & Romano, 2010). Molto chiaro in questo senso è anche il quadro che emerge dall'applicazione dei PVI (Grabe & Low, 2002)[3]. Nello studio comparativo di Mok & Dellwo (2008) il tedesco si distingue dall'italiano sia per le durate degli intervalli vocalici (%V, nPVI-V) che per la variabilità degli intervalli consonantici (rPVI-C, VarcoC).

Data la diversità ritmica attestata per l'italiano e il tedesco, questa coppia di lingue si presta molto bene per la finalità della nostra ricerca, ovvero per l'analisi delle caratteristiche temporali presso parlanti nativi, non-nativi e bilingui. Non a caso la diversità prosodica (e in particolare ritmica) tra queste due lingue è stata segnalata da tempo come una delle maggiori difficoltà nella pronuncia del tedesco da parte di apprendenti italofoni (cfr. Missaglia, 1999).

2.3. Caratteristiche ritmiche delle lingue seconde

Tra i primi lavori sperimentali sul ritmo in una seconda lingua figura quello di Gut (2003), che esamina il tedesco parlato da locutori con diverse L1, tra cui anche l'italiano. Benché non vengano ancora applicate le metriche menzionate in 2.1., dalle misurazioni di Gut emerge un dato che corrobora senz'altro l'ipotesi delle classi ritmiche: nel tedesco letto da locutori italiani lo scarto tra le durate medie delle vocali toniche e le durate medie delle vocali atone è chiaramente inferiore rispetto a quello dei parlanti nativi.

In una ricerca contrastiva su inglese, neerlandese, francese e spagnolo, White & Mattys (2007) hanno applicato una serie di metriche ritmiche alla produzione di parlanti che leggevano sia nella loro lingua materna che in una lingua seconda. Tra i principali risultati spicca la minore velocità di eloquio dei parlanti non-nativi nonché un chiaro effetto della lingua materna su VarcoV: ad esempio, l'inglese degli ispanofoni presenta – rispetto a quello dei parlanti nativi – valori di VarcoV minori, probabilmente dovuti ad una minore differenza di durata tra vocali atone e toniche (si tratterebbe quindi di un dato analogo a quello riscontrato per la coppia italiano-tedesco da Gut, 2003). Sempre presso parlanti ispanofoni dell'inglese, anche Dellwo et alii (2009) rilevano una minore velocità di eloquio rispetto ai

[2] Diversa è la situazione dei dialetti italo-romanzi i quali, dal punto di vista della tipologia ritmica, devono essere trattati come lingue autonome e non come varietà di una stessa lingua, dato che ciascun dialetto possiede un proprio lessico con le sue specifiche restrizioni fonotattiche, nonché regole allofoniche che possono incidere ad esempio sulla realizzazione delle vocali atone (cfr. Russo & Barry, 2010, per il ritmo dei dialetti campani e Romano et alii, 2010, per alcune misure ritmiche di dialetti piemontesi; per uno studio contrastivo di vari dialetti italo-romanzi v. ora anche Schmid, 2012).

[3] Com'è noto, molteplici fattori – tra cui il tipo di parlato (letto *vs.* spontaneo) e la velocità di eloquio – sono in grado di influenzare i valori forniti dalle diverse metriche, per cui le varietà regionali del tedesco e dell'italiano si posizionano a volte in zone non previste dall'ipotesi delle classi ritmiche (v. ad esempio Barry et alii, 2003).

parlanti nativi, ma nei loro dati le misure ritmiche che meglio distinguono l'inglese L1 da L2 sono i PVI vocalici e consonantici. Non sempre le misure ritmiche riescono a fornire un quadro chiaro del parlato dei parlanti non-nativi, come mettono in evidenza Mok & Dellwo (2008) per apprendenti dell'inglese con lingua madre cantonese e mandarino; nei loro dati si trovano infatti tanto valori di tipo *syllable-timed* quanto di tipo *stress-timed*. Invece, è stato riscontrato un chiaro influsso del ritmo 'sillabico' del francese L1 sull'inglese L2 per le metriche ΔC, %V e VarcoV nello studio Tortel & Hirst (2010), che mostra peraltro che i valori dei francofoni più competenti nella L2 si avvicinano ai valori dei parlanti nativi. Infine, il fatto di non leggere la propria lingua materna può avere un effetto significativo sulle metriche vocaliche persino quando L1 e L2 appartengono alla stessa classe ritmica, com'è stato mostrato da Dellwo (2010a) per %V e nPVI-V con la coppia tedesco e inglese.

2.4. Caratteristiche ritmiche nel parlato dei bilingui

Riguardo alle caratteristiche ritmiche del parlato dei bilingui 'precoci' si possono avanzare due ipotesi fondamentali (cfr. nota 1): o i bilingui si comportano come i parlanti nativi monolingui in ambedue le lingue del loro repertorio (chiameremo questa ipotesi 'nativa'), oppure essi si discostano alquanto dalle caratteristiche ritmiche dei parlanti nativi in direzione dei parlanti non-nativi, pur non raggiungendo i valori di questo gruppo (chiameremo questa ipotesi 'intermedia')[4]. Teoricamente si potrebbe formulare anche una terza ipotesi, secondo cui i bilingui si comporterebbero piuttosto come degli apprendenti (sarebbe un'ipotesi 'non-nativa'), ma ciò sembra poco plausibile; nei nostri dati ci aspettiamo quindi piuttosto di trovare evidenza per la prima e/o la seconda ipotesi.

Le scarse ricerche sul ritmo nei parlanti bilingui (in senso stretto) forniscono supporto sia per l'ipotesi nativa che per l'ipotesi intermedia. Ad esempio, nei soggetti bilingui svizzero-tedesco e francese studiati da Galloway (2007) non emerge nessuna differenza significativa riguardo ai PVI vocalici e consonantici, per cui la ricercatrice ribadisce a proposito dell'ipotesi nativa che "it is possible for proficient bilinguals to achieve monolingual-like rhythm" (p. 79); la stessa autrice ammette comunque che "individual variation occurs" e che "rhythm can fall somewhere in between" (p. 82). Una chiara evidenza per l'ipotesi intermedia si trova invece nello studio di Carter (2005) sull'inglese parlato da due generazioni di immigrati messicani nel North Carolina: un confronto dei PVI dei nuclei sillabici mostra che i bilingui ottengono in spagnolo valori più alti rispetto ai parlanti monolingui, mentre nell'inglese i loro valori sono sensibilmente più bassi rispetto ai parlanti anglofoni. I dati più interessanti riguardo alle due ipotesi formulate sopra provengono dallo studio di Bunta & Ingram (2007), nel quale si esaminano per le stesse due lingue (inglese e spagnolo) il ritmo di locutori monolingui (solo nella lingua madre) e di locutori bilingui (in ambedue le lingue); in più, per ciascuna delle quattro condizioni vengono analizzate tre fasce di età (bambini piccoli, bambini grandi, adulti). Ora, dai valori dei PVI vocalici emerge un percorso di acquisizione dei bilingui che parte da un ritmo più 'intermedio' nei bambini più

[4] Una certa evidenza empirica per l'ipotesi 'intermedia' proviene da una ricerca condotta con dati molto diversi, ma rilevati presso un campione molto simile al nostro: in un'analisi degli errori riscontrati negli elaborati scritti di studenti di italianistica all'Università di Zurigo (Berruto et alii, 1988) appare che per tutti i livelli di analisi considerati (testualità, sintassi, lessico, morfologia e grafia) il gruppo dei bilingui ha prodotto un numero di errori intermedio tra quello dei parlanti nativi e quello dei parlanti non-nativi.

piccoli per approdare man mano a un ritmo più 'nativo' nei bambini più grandi e nei parlanti adulti.

3. MATERIALI E METODO

Al fine di verificare le due ipotesi formulate in 2.4. è stato allestito, al laboratorio di fonetica dell'Università di Zurigo, un corpus di parlato letto bilingue italiano e tedesco (BiCor). Forniamo di seguito alcune indicazioni relative al campione dei parlanti (3.1.), ai materiali linguistici raccolti (3.2.) e alle procedure di analisi adottate (3.3.).

3.1. I parlanti

Sono stati registrati 15 studenti dell'Università di Zurigo, tutti di un'età compresa tra i 20 e i 30 anni. Il campione può essere suddiviso in tre gruppi in base al repertorio linguistico dei parlanti. Il primo gruppo consiste di 5 parlanti italofoni, tutti nati e cresciuti nel Canton Ticino dove hanno frequentato le scuole elementari, medie e superiori; complessivamente hanno avuto 7 anni di istruzione formale di tedesco. Il secondo gruppo è composto di 5 parlanti tedescofoni, tutti provenienti dalla Svizzera tedesca. In questo caso, il livello di competenza della L2 è meno omogeneo: si tratta di 3 apprendenti principianti che imparano l'italiano per interesse personale, mentre gli altri due soggetti hanno l'italiano o comunque la romanistica come materia di studio, per cui il loro livello di competenza può essere considerato medio se non addirittura avanzato.

Il terzo gruppo, infine, è quello dei parlanti bilingui. 4 studenti hanno delle origine italiane: in due casi entrambi i genitori sono italiani, mentre due parlanti hanno almeno un genitore italiano. Una studentessa è nata in Italia, mentre gli altri tre sono nati in Svizzera. Tutti e quattro sono stati scolarizzati in lingua tedesca, mentre parlano l'italiano in famiglia. In genere, questo tipo di bilinguismo può essere considerato abbastanza 'equilibrato', con un lieve grado di dominanza – almeno dal punto di vista della competenza – del versante tedesco (cfr. De Rosa & Schmid, 2000), per cui questi 4 parlanti presentano una certa somiglianza con quelli del secondo gruppo. La quinta parlante bilingue presenta invece un repertorio per così dire inverso, più simile a quello del primo gruppo: si tratta di una studentessa nata e cresciuta in Ticino che parla tedesco in famiglia (in particolare con la madre, originaria della Germania).

3.2. Le registrazioni

Avendo tutti i 15 soggetti una certa competenza nelle due lingue in esame (chi come prima lingua, chi come seconda lingua), si è proceduto alla costituzione di un corpus bilingue italiano-tedesco (BiCor). Come materiale di lettura sono state scelte le 20 frasi italiane del corpus di Ramus et alii (1999), che sono state tradotte liberamente in tedesco dal primo autore di questo contributo. Per l'analisi si è tenuto conto delle prime 10 frasi in ciascuna lingua (le frasi registrate vengono riprodotte nell'Appendice), per cui il corpus analizzato ammonta a 300 frasi (3 gruppi x 5 parlanti x 10 frasi x 2 lingue).

Le registrazioni si sono svolte in una stanza del laboratorio di fonetica dell'Università di Zurigo mediante un registratore digitale Fostex FR-2LE e un microfono a cravatta Sennheiser MKE 2 (omnidirezionale, gamma di frequenza di 20-20'000 Hz ±23 dB e coefficiente di trasmissione a vuoto di 10 mV/Pa ±2,5 dB).

3.3. Procedura di analisi

Le registrazioni sono state analizzate mediante il programma *Praat* (Boersma & Weenink, 2011). In un primo passo dell'analisi le singole frasi sono state segmentate con una

griglia di testo (*TextGrid*) a un primo livello (*tier*); la segmentazione e l'etichettatura sono state effettuate manualmente, assegnando un simbolo SAMPA ad ogni porzione del segnale acustico che corrisponde ad un fono della lingua italiana o tedesca. Successivamente sono stati aggiunti altri strati mediante una procedura automatica, un *Plugin* di *Praat* dal nome *CVTierCreator* che è stato elaborato dal secondo autore di questo contributo. L'analisi ritmica presuppone infatti almeno quattro livelli: i) un livello *cv-segments* che assegna ogni fono del primo livello a una delle due categorie "c" (consonante) o "v" (vocale); ii) un livello *cv-intervals*$_1$ che riunisce i segmenti successivi della stessa categoria nei rispettivi intervalli consonantici e vocalici pur indicando il numero di foni di cui è composto ogni intervallo; iii) un livello *cv-intervals*$_2$ che rappresenta invece gli intervalli consonantici e vocalici con un'unica etichetta "c" o "v", come di consueto nelle metriche ritmiche più usate; iv) un livello *voicing* che suddivide la catena fonica non in intervalli vocalici e consonantici, bensì in fasi periodiche e aperiodiche, ovvero in intervalli 'sonori' (*voiced, v*) e 'sordi' (*unvoiced, u*), per cui le parti *voiced (v)* contengono – oltre alle vocali – anche delle consonanti sonore (Dellwo et alii, 2007). Infine, sono state calcolate le durate di segmenti e intervalli con l'ausilio di *DurationAnalyzer*, un altro *plugin* elaborato dal secondo autore di questo contributo che fornisce oltre 50 misure temporali[5].

Ai fini della nostra analisi sono stati inseriti manualmente altri due strati: si tratta di un *tier* che suddivide il segnale in sillabe (*syllables*) e di un altro livello che segnala tutte le sillabe portanti un accento lessicale (*stress*).

4. RISULTATI

I materiali raccolti si prestano a vari tipi di analisi, proprio perché sono stati prodotti da tre categorie di parlanti. In un primo momento riporteremo i valori %V, ΔC e PVI dei nostri parlanti nativi, paragonandoli con dati riportati in letteratura per l'italiano e il tedesco (4.1.), dopodiché presenteremo una nuova metrica, calcolando il rapporto di durata tra le sillabe toniche e atone (4.2.). La parte principale dell'analisi sarà dedicata a un confronto dei tre gruppi di parlanti riguardo a due parametri fondamentali, la velocità di eloquio (4.3.) e la variabilità vocalica espressa come ΔVln e nPVI-V (4.4.). Infine daremo uno sguardo alla variabilità interindividuale (vista per il campione intero e all'interno dei tre gruppi di parlanti) attraverso la percentuale degli intervalli sonori (4.5) e la velocità di elocuzione (4.6).

4.1. L'italiano e il tedesco dei parlanti nativi

Una prima analisi alla quale si prestano i dati del corpus BiCor è il semplice confronto dei valori forniti dai nostri locutori nativi di italiano e tedesco con i valori riportati in base alle stesse metriche ritmiche da altri lavori.

Lingua	Metrica	Ramus et alii (1999)	Grabe & Low (2002)	BiCor
Italiano L1	%V	45.2		43.3
	ΔC	0.048		0.048
Tedesco L1	nPVI-V		59.7	59.6
	rPVI-C		55.3	79.1

[5] I due plugin *CVTierCreator* e *DurationAnalyzer* sono disponibili al sito del laboratorio di fonetica dell'Università di Zurigo: http://www.pholab.uzh.ch/leute/dellwo/software.html.

Tabella 1: Valori %V, ΔC e PVI dei parlanti nativi.

Ci limitiamo in questa sede a un confronto con i due studi che hanno introdotto le metriche più usate in letteratura. Non sorprende la coincidenza molto alta per le due metriche %V e ΔC, visto che nei due corpora sono stati lette – almeno in parte – le stesse frasi. Notevole è però anche la coincidenza quasi totale per nPVI-V, benché si tratti di materiali linguistici diversi. Lo scarto per la metrica non normalizzata rPVI-C potrebbe essere dovuto a una diversa velocità di eloquio, ma occorrerebero ulteriori analisi per approfondire questa ipotesi.

Possiamo comunque interpretare la notevole coincidenza tra i nostri valori con quelli trovati nei testi di riferimento come supporto empirico per l'ipotesi delle classi ritmiche: replicando con altri parlanti la stessa procedura sperimentale si ottengono risultati molto simili, il che depone a favore di una certa consistenza delle metriche ritmiche.

4.2. Rapporto di durata tra sillabe toniche e atone

La maggior parte delle metriche ritmiche attualmente adoperate si basano sulla suddivisione del segnale in intervalli vocalici e consonantici (cfr. 2.1.), ma in linea di principio nulla impedisce che non si possano calcolare altre misure, come ad esempio il rapporto di durata tra le sillabe toniche e atone[6].

La figura 1 visualizza tale rapporto per il tedesco (a sinistra) e l'italiano (a destra) nella produzione dei tre gruppi analizzati: i bilingui (boxplot bianchi), i tedescofoni (boxplot a strisce) e gli italofoni (boxplot a puntini).

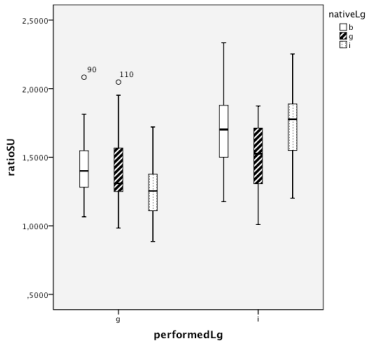


Figura 1: Rapporto di durata tra sillabe toniche e atone in tedesco e in italiano (b = bilingui, g = tedescofoni, i = italofoni).

Un'ANOVA con la variabile dipendente 'rapporto di durata tra sillabe toniche e atone' (*ratioSU*) rivela un'interazione significativa tra i due fattori 'lingua madre' (*native language*) e 'lingua letta' (*performed language*): F[5, 299] = 12.9; p<.001. Dalla figura 1 pare poter dedurre che gli italofoni si distinguano dagli altri due gruppi nella lettura in tedesco (a sinistra), così come i tedescofoni sembrano differenziarsi dagli altri due gruppi nella lettura dell'italiano (a destra). Ebbene, questa impressione viene confermata da due ANOVA separate per la lettura in tedesco (F[2, 149] = 8.85; p <.001) e in italiano (F[2, 149] = 10.1; p<.001); il fattore 'lingua madre' rimane significativo anche dopo la correzione Bonferroni

[6] Questa proposta è stata suggerita nel commento all'abstract di questo contributo da parte di un revisore anonimo, che qui ringraziamo.

con un livello $\alpha = 0.025$. In un test post-hoc (Tukey) gli italofoni si differenziano nella lettura del tedesco in modo significativo sia dai tedescofoni ($p<.008$) che dai bilingui ($p<.001$). Insomma, il tedesco dei bilingui si distingue da quello degli italofoni, ma non da quello dei tedescofoni. Ai nostri fini è importante sottolineare che questo risultato fornisce supporto per la prima delle due ipotesi avanzate in 2.4., cioè per l'ipotesi 'nativa'.

Al lettore attento non sarà sfuggito un altro aspetto che emerge dalla figura 1: complessivamente, e in particolare nel caso dei parlanti nativi italofoni e tedescofoni, il rapporto di durata tra sillabe toniche e atone è leggermente più elevato in italiano (a destra) che non in tedesco (a sinistra). Nell'ottica della tradizionale ipotesi dell'isocronia ci si aspetterebbe invece il contrario, cioè che in una lingua di tipo 'sillabico' – qual è presumibilmente l'italiano – le durate delle sillabe toniche si differenzino meno dalle sillabe atone che non in una lingua di tipo 'accentuale' come il tedesco. È evidente che i dati a nostra disposizione vanno contro questa ipotesi, ma non stiamo a commentare questo risultato difficilmente interpretabile, che sottolinea se non altro la notevole complessità inerente alla fenomenologia del ritmo linguistico.

4.3. Velocità di eloquio (I): confronto tra parlanti nativi, bilingui e non-nativi

Com'è ovvio, i valori ottenuti per la velocità di eloquio dipendono dal tipo di calcolo adottato (cfr. Roach, 1998; Dellwo, 2010a): se usiamo come misura il numero di sillabe prodotte al secondo, possiamo aspettarci di riscontare in una lingua come il tedesco (che ammette strutture sillabiche piuttosto complesse) una velocità di eloquio minore rispetto ad una lingua come l'italiano (che presenta delle strutture sillabiche relativamente meno complesse). D'altro canto, la velocità di eloquio è anche un indice di fluenza del singolo locutore e può quindi essere messa in relazione, in particolare nel caso di un parlante non-nativo, con il suo livello di competenza nella lingua in questione (cfr. Dellwo & Wagner, 2003).

La figura 2 mostra sull'asse delle ascisse i tre gruppi di locutori del nostro campione, da sinistra a destra i tedescofoni (g), i bilingui (b) e gli italofoni (i); le due lingue lette dai tre gruppi sono l'italiano (indicato con un quadratino) e il tedesco (indicato con un cerchietto). Per rappresentare la velocità di eloquio abbiamo scelto sull'asse delle ordinate l'intervallo di confidenza, indicando con una probabilità del 95% la gamma entro la quale si posizionerebbero eventuali dati aggiuntivi.

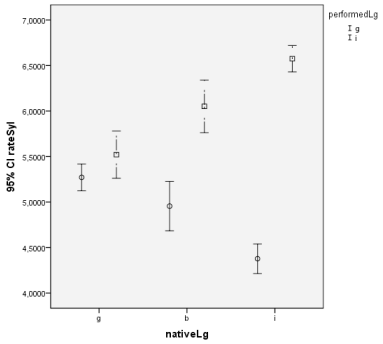


Figura 2: Intervalli di confidenza al 95% per la velocità di eloquio (g = tedescofoni, b = bilingui, i = italofoni).

Il grafico della figura 2 fornisce da un lato evidenza per la dipendenza della velocità di eloquio (espressa in sillabe al secondo) dalla struttura fonotattica delle lingue in gioco: in

effetti, per tutti e tre i gruppi di locutori i quadratini dell'italiano esprimono valori superiori rispetto ai cerchietti del tedesco. Allo stesso tempo, il grafico mostra però anche una chiara differenza fra i tre gruppi di parlanti: i tedescofoni sono appena più 'veloci' in italiano che non nella loro lingua materna, mentre gli italofoni mostrano uno scarto notevole tra la velocità molto elevata nella L1 e quella molto bassa nella L2. Molto interessante ai nostri fini sono gli intervalli di confidenza dei locutori bilingui, che si staccano da quelli dei locutori non-nativi senza tuttavia raggiungere la posizione dei locutori nativi. Possiamo dunque interpretare questo risultato come supporto dell'ipotesi 'intermedia' formulata per i locutori bilingui (cfr. 2.4.).

4.4. Variabilità vocalica: ΔVln e nPVI-V di parlanti nativi, bilingui e non-nativi

Per esaminare la variabilità degli intervalli vocalici realizzati dai tre gruppi di locutori in italiano e in tedesco abbiamo scelto due metriche normalizzate, ΔVln e nPVI-V.

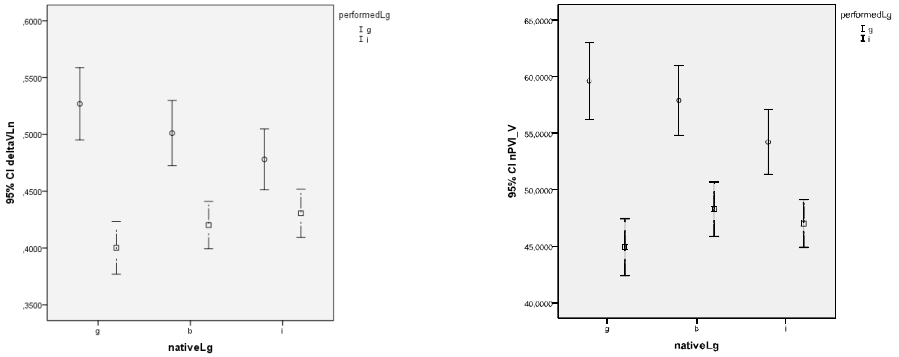


Figura 3: Intervalli di confidenza al 95% per ΔVln (a sinistra) e nPVI-V (a destra) (g = tedescofoni, b = bilingui, i = italofoni).

La metrica ΔVln è una variante dell'originaria metrica ΔV proposta da Ramus et alii (1999), alla quale si applica però una trasformazione logaritmica per ottenere una distribuzione normale dei dati; per una motivazione di questa procedura di normalizzazione rinviamo a Dellwo (2010b).

La figura 3 proietta sui tre gruppi di locutori (in ascissa) gli intervalli di confidenza al 95% per ΔVln (a sinistra) e per nPVI-V (a destra) in italiano (quadratini) e tedesco (cerchietti). I due grafici mostrano a prima vista un'immagine speculare rispetto a quella riportata nella figura 2, rivelando tuttavia delle tendenze analoghe. Per tutti e tre i gruppi di locutori la variabilità vocalica è nettamente maggiore in tedesco che non in italiano, il che costituisce un ovvio riflesso delle differenze fonologiche tra le due lingue riguardo sia alle vocali toniche che alle vocali atone. Notiamo quindi che, diversamente dal rapporto di durata tra sillabe toniche e atone (cfr. 4.2.), le metriche per la variabilità degli intervalli vocalici forniscono dei risultati che sono perfettamente in linea con l'ipotesi delle classi ritmiche.

Per quanto riguarda la differenziazione fra i tre gruppi di parlanti, è interessante osservare che nella lettura in italiano i locutori tedescofoni mostrano una variabilità delle durate vocaliche più ridotta: ciò significa che essi non proiettano affatto il modello ritmico del tedesco sull'italiano, ma che realizzano piuttosto – in modo ipercorretto – un ritmo quasi più

'sillabico' degli stessi locutori italofoni. Questi ultimi, invece, manifestano una chiara interferenza dell'italiano sulla loro lettura in tedesco (rappresentata dalla vicinanza delle due lingue in questo gruppo); probabilmente questo risultato deriva sia dalla mancata riduzione delle vocali atone sia dal mancato allungamento delle vocali toniche tese del tedesco. Il gruppo dei bilingui, infine, si trova tutto sommato di nuovo in una posizione intermedia (con l'eccezione del nPV-V italiano che supera quello dei parlanti nativi), il che depone a favore di una certa interazione – almeno per quanto riguarda il ritmo – tra i due sistemi fonologici in questo tipo di parlante.

4.5. La percentuale degli intervalli sonori (%Voiced)

Un aspetto interessante che caratterizza i nostri dati è che per molte misure temporali emerge una forte variabilità interindividuale sia all'interno del campione intero sia all'interno delle tre categorie di parlanti. La figura 4 mostra la variabilità della percentuale degli intervalli sonori *(%Voiced*; cfr. 3.3.) in tedesco (boxplot bianchi) e in italiano (boxplot a strisce); i 15 parlanti sono suddivisi secondo i tre gruppi (italofoni, bilingui e tedescofoni).

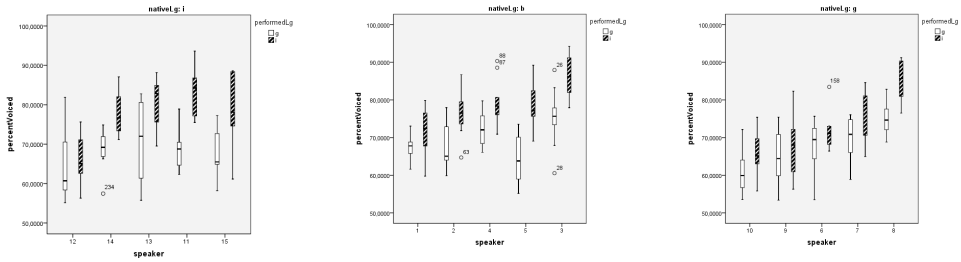


Figura 4: Percentuale degli intervalli sonori *(%Voiced*) per italofoni, bilingui e tedescofoni.

Paragonando i tre gruppi di parlanti si vede che gli italofoni (grafico a sinistra) hanno mediamente delle durate maggiori per gli intervalli sonori; inoltre, questo gruppo tende a separare maggiormente le due lingue. Invece, i tedescofoni (grafico a destra) mostrano i valori di sonorità minori per l'italiano, con una forte sovrapposizione dei boxplot sull'asse delle Y. I parlanti bilingui (grafico al centro) formano un gruppo più variegato: ad esempio, i parlanti 2, 5 e 3 tendono a separare in modo chiaro le due lingue, mostrando un comportamento simile a quello dei parlanti italofoni a sinistra, anche se tutto sommato i boxplot per l'italiano dei bilingui sono caratterizzati da una minore variabilità e da una posizione mediamente inferiore sull'asse delle Y. All'interno del gruppo dei bilingui, un'ANOVA rivela comunque degli effetti altamente significativi per i fattori 'parlante' ($p<0.001$) e 'lingua' ($p<0.001$) considerati separatamente, mentre non vi è nessuna interazione significativa tra i due fattori 'parlante*lingua' ($p=0.065$). Anche per gli altri due gruppi (italofoni e tedescofoni) si ottengono degli effetti significativi per i due fattori 'lingua' e 'parlante' considerati separatamente, di nuovo senza alcuna interazione tra i due fattori.

La minore durata degli intervalli sonori prodotti dai tedescofoni (e in parte anche dai bilingui) sarà determinata almeno in parte da differenze a livello di 'dettaglio fonetico fine' tra le due lingue, legate alla realizzazione delle ostruenti sonore in tedesco e in particolare nella varietà del tedesco standard parlato in Svizzera. Com'è noto, in tedesco standard le ostruenti fonologicamente 'sonore' spesso non vengono realizzate con una vibrazione delle

pliche vocali: ad esempio, nei contesti di coda sillabica, il contrasto di sonorità viene neutralizzato (il che dà luogo ad una realizzazione sorda delle ostruenti fonologicamente sonore); inoltre, le occlusive sonore in posizione iniziale hanno un VOT che si avvicina a 0. Nei dialetti svizzero-tedeschi e nella varietà elvetica del tedesco standard (che è quella in cui sono state lette le nostre frasi), il tratto distintivo [±sonoro] viene sostituito con il tratto [±teso] (per cui anche le ostruenti 'sonore' del tedesco standard non presentano alcuna vibrazione della glottide), il quale ha come correlato fonetico principale una differenza di durata (cfr. De Rosa & Schmid, 2000). Molto probabilmente, la minore durata degli intervalli sonori che emerge dalla produzione dei soggetti tedescofoni visualizzata nella figura 4 è dovuta a questa differenza tra italiano e tedesco; una parziale sostituzione di ostruenti sonore con consonanti 'rilassate' era stata riscontrata non solo nell'italiano di figli di emigrati italiani residenti nella Svizzera tedesca (De Rosa & Schmid, 2000), ma già nel bilinguismo svizzero-tedesco/romancio studiato nel lavoro pionieristico di Weinreich (1963).

Più in generale la metrica *%Voiced* fornisce ulteriore evidenza a favore dell'ipotesi 'intermedia' per i locutori bilingui, ma la figura 4 rivela allo stesso tempo una forte variazione interindividuale all'interno dei tre gruppi. Approfondiamo ulteriormente questo aspetto, tornando alla velocità di eloquio (cfr. 4.3.).

4.6. Velocità di eloquio (II): variabilità interindividuale

La figura 5 mostra la velocità di eloquio, calcolata in sillabe al secondo, per ognuno dei 15 locutori (divisi nei tre gruppi: italofoni, bilingui e tedescofoni).

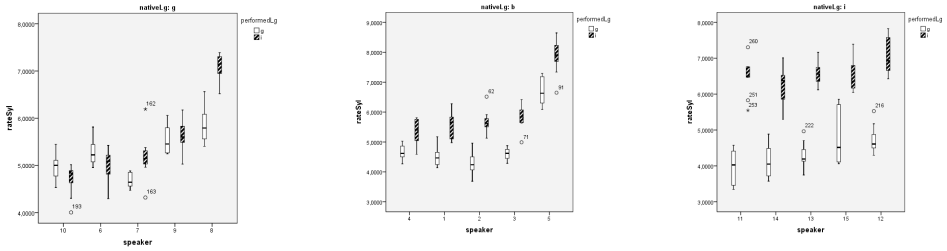


Figura 5: Velocità di eloquio in sillabe/secondo (parlanti tedescofoni, bilingui e italofoni).

Per quanto riguarda gli italofoni (grafico a destra) si nota che tutti i parlanti riescono a separare chiaramente le due lingue: i boxplot per l'italiano (a strisce) si espandono su valori molto più alti e non si sovrappongono mai con i valori per il tedesco. Tutto sommato ciascuno dei cinque locutori conferma il quadro generale della figura 2, ma emergono anche dei pattern individuali: ad esempio, il parlante 12 mostra valori più alti del parlante 13 sia in italiano che in tedesco. Nel confronto con gli italofoni, i boxplot dei bilingui (grafico al centro) mostrano dei valori più bassi per la velocità di eloquio in italiano che si avvicinano – tuttavia senza mai sovrapporsi – ai valori della velocità di eloquio in tedesco; anche in questo gruppo le differenze interindividuali sono notevoli, se si confronta, ad esempio, il locutore 5 con il locutore 1. Ancora diverso è il quadro dei tedescofoni (grafico a sinistra), dove i boxplot delle due lingue possono anche occupare posizioni molto simili, data la velocità di eloquio relativamente alta del tedesco L1 e quella relativamente bassa dell'italiano L2; per i locutori 10 e 6 i valori del tedesco sono addirittura più alti rispetto a quelli dell'italiano. Nel caso dei parlanti tedescofoni la velocità di eloquio rappresenta da un lato un indice per la loro fluenza in italiano: non a caso nei due apprendenti più avanzati (i locutori 7 e

8) le due lingue sono nettamente separate. Dall'altro lato, questi due locutori mostrano come la velocità di eloquio costituisca per certi versi anche una caratteristica personale. Aggiungiamo, a titolo di aneddoto, un'informazione sui due locutori 8 (a destra nel grafico dei tedescofoni) e 5 (a destra nel grafico dei bilingui), ambedue caratterizzati da una velocità di eloquio piuttosto elevata nelle due lingue: si tratta di due studenti fidanzati, per cui viene spontaneo pensare ad un effetto di *speech accomodation* nello stile di parlare all'interno della coppia.

5. OSSERVAZIONI CONCLUSIVE

Fiumi di inchiostro sono stati versati sulla natura del ritmo nelle lingue storico-naturali, e non è certamente nostra pretesa potere esaurire tale questione per l'italiano e il tedesco. Le nostre analisi hanno confermato alcuni risultati forniti da studi precedenti che hanno applicato delle misure ritmiche a queste due lingue, in particolare per quanto riguarda le metriche %V e ΔC dei parlanti nativi dell'italiano e il nPVI-V dei parlanti nativi del tedesco (4.1.), le metriche ΔVln e nPVI-V di tutti e tre i gruppi di parlanti (4.4.) e la percentuale degli intervalli sonori (4.5.). Un indizio contrario alla tradizionale classificazione ritmica è invece emerso dall'applicazione di una nuova metrica che calcola il rapporto di durata tra le sillabe toniche e le sillabe atone (4.2.): contrariamente alle aspettative, nei nostri dati tale rapporto è maggiore in italiano che non in tedesco.

L'obiettivo principale di questo studio è consistito nell'applicazione di misure temporali a un confronto fra tre gruppi di parlanti (nativi, bilingui e non-nativi) delle stesse due lingue, l'italiano e il tedesco. Di particolare interesse si è rivelato in questo contesto il gruppo di parlanti sinora meno studiato, ovvero i bilingui in senso stretto. L'analisi dei nostri dati ha da un lato fornito evidenza per la cosiddetta ipotesi 'nativa' (2.4.), se si considera ad esempio il rapporto di durata tra sillabe toniche e atone (4.2.). Dall'altro lato sono emersi numerosi indizi a favore dell'ipotesi 'intermedia' (2.4.) che derivano dalla velocità di eloquio (4.3.), dalla variabilità delle durate vocaliche (4.4.) e dalla percentuale degli intervalli sonori (4.5.). Riguardo alla velocità di eloquio abbiamo inoltre riscontrato notevoli differenze tra i singoli locutori all'interno dei tre gruppi di parlanti; tale variabilità interindividuale costituirà l'oggetto di future ricerche del nostro gruppo.

RINGRAZIAMENTI

I due autori ringraziano Laura Tramutoli per la registrazione dei parlanti e la realizzazione di una parte della segmentazione dei materiali italiani, This Müller per una prima segmentazione dei materiali tedeschi, nonché i 15 studenti per la partecipazione all'esperimento.

APPENDICE: LE FRASI LETTE

Le 10 frasi italiane

1 La moglie del farmacista sa sempre ciò che vuole.
2 Il teatro ha introdotto molte nuove discipline.
3 Non ha mai voluto rendersi conto dei suoi gran difetti.
4 L'organizzazione dei trasporti collettivi è carente.
5 La situazione della Bilancia dei pagamenti non mi lascia mai tranquillo.
6 I genitori lasciano Marco senza risorse.
7 Le forti piogge della primavera sono dannose.
8 Il treno più rapido resta comunque il pendolino.

9 La ricostruzione della città dovrà farsi lentamente.
10 Il Ministero della Cultura ha scelto la via più semplice.

Le 10 frasi tedesche

1 Die Frau des Apothekers weiss immer was sie will.
2 Das Theater hat viele neue Aufführungen geplant.
3 Er wollte sich seiner Schwächen einfach nicht bewusst werden.
4 Der Öffentliche Verkehr lässt zu wünschen übrig.
5 Die schlechte Zahlungsbilanz lässt mich nicht zur Ruhe kommen.
6 Die Eltern geben ihm keine finanzielle Unterstützung.
7 Die starken Frühlingsregen richten viele Schäden an.
8 Der schnellste Zug ist immer noch der ICE.
9 Der Wiederaufbau der Stadt wird sehr lange dauern.
10 Der Bildungsministerium hat den einfachsten Weg gewählt.

RIFERIMENTI BIBLIOGRAFICI

Abercrombie, D. (1967), Elements of General Phonetics, Edimburgo: University Press.

Barry, W. (2010), Rhythm measures in retrospect. Reflections on the nature of spoken language rhythm, in La dimensione temporale del parlato. Atti del V Convegno Nazionale AISV (S. Schmid, M. Schwarzenbach & D. Studer, editors), Torriana: EDK Editore, 3-12.

Barry, W., Andreeva, B., Russo, M., Dimitrova, S. & Kostadinova, T. (2003), Do rhythm measures tell us anything about language type?, in Proceedings of the 15th International Congress of Phonetic Sciences (M.J. Solé, D. Recasens & J. Romero, editors), Barcelona, Vol. 3, 2693-2696.

Berruto, G., Moretti, B. & Schmid, S. (1988), L'italiano di parlanti colti in una situazione plurilingue, Rivista italiana di dialettologia, 12, 7-100.

Bertinetto, P.M. (1977), 'Syllabic blood' ovvero l'italiano come lingua ad isocronismo sillabico, Studi di grammatica italiana, 6, 69-96.

Bertinetto, P.M. (1989), Reflections on the dichotomy 'stress' vs. 'syllable-timing', Revue de Phonétique Appliquée, 91-93, 99-130.

Bertinetto, P.M. & Bertini, C. (2008), On modeling the rhythm of natural languages, in Proceedings of Speech Prosody 2008 (P. Barbosa, S. Madureira & C. Reis, editors), Campinas, Brazil: Editora RG/CNPq, 427-430.

Boersma, P. & Weenink, D. (2011), Praat: doing phonetics by computer (Versione 5.2).

Bunta, F. & Ingram, D. (2007), The acquisition of speech rhythm by bilingual Spanish- and English-speaking 4- and 5-year old children, Journal of speech, language and hearing research, 50, 999-1014.

Carter, P. (2005), Quantifying rhythmic differences between Spanish, English, and Hispanic English, in Theoretical and experimental approaches to romance linguistics: Selected papers from the 34th linguistic symposium on romance languages (R. S. Gess & E. J. Rubin, editors), Amsterdam: John Benjamins, 63–75.

Dauer, R.M. (1983), Stress-timing and syllable-timing re-analysed, Journal of Phonetics, 11, 51-62.

De Rosa, R. & Schmid, S. (2000), Aspetti della competenza ortografica e fonologica nell'italiano di emigrati di seconda generazione nella Svizzera tedesca, Rivista italiana di dialettologia, 24, 53-96.

Dellwo, V. (2006), Rhythm and Speech Rate: A Variation Coefficient for ΔC, in Language and Language-processing (P. Karnowski & I. Szigeti, editors), Frankfurt am Main: Peter Lang, 231-241.

Dellwo, V. (2010a), Influences of speech rate on the acoustic correlates of speech rhythm, PhDThesis, Univ. of Bonn, Germany
[disponibile al sito: http://hss.ulb.uni-bonn.de:90/2010/2003/2003.htm]

Dellwo, V. (2010b), Choosing the right speech rate normalization method for measurements of speech rhythm, in La dimensione temporale del parlato. Atti del V Convegno Nazionale AISV (S. Schmid, M. Schwarzenbach & D. Studer, editors), Torriana: EDK Editore, 13-32.

Dellwo, V. & Wagner, P. (2003), Relations between language rhythm and speech rate, in Proceedings of the 15th International Congress of Phonetic Sciences (M.J. Solé, D. Recasens & J. Romero, editors), Barcelona, Vol. 1, 471-474.

Dellwo, V., Fourcin, A. & Abberton, E. (2007), Rhythmical classification of languages based on voice parameters, in Proceedings of the 16th International Congress of Phonetic Sciences (J. Trouvain & W. Barry, editors), Saarbrücken, Vol. 2, 1129-1132.

Dellwo, V., Gutiérrez Díez, F. & Gavalda, N. (2009), The development of measurable speech rhythm in Spanish Speakers of English, in Actas del XI Simposio internacional de comunicación social, Santiago de Cuba, 594-597.

Ferragne, E. & Pellegrino, F. (2004), A comparative account of the suprasegmental and rhythmic features of British English dialects, in Proceedings of ''Modelisations pour l'Identification des Langues'', Paris.

Galloway, R. (2007), Bilinguals' interacting phonologies? A study of speech production in French-Swiss German bilinguals, Master Thesis, University of Cambridge.

Giordano, R. & D'Anna, L. (2010), A comparison of rhythm metrics in different speaking styles and fifteen regional varieties of Italian, in Proceedings of Speech Prosody 2010, Chicago [accessibile al sito http://speechprosody2010.illinois.edu/papers/100826.pdf].

Grabe, E. & Low, E.L. (2002), Durational Variability in Speech and the Rhythm Class Hypothesis, in Papers in Laboratory Phonology 7 (C. Gussenhoven, editor), Berlin: Mouton de Gruyter, 515-546.

Gut, U. (2003), Non-native rhythm in German, in Proceedings of the 15th International Congress of Phonetic Sciences (M.J. Solé, D. Recasens & J. Romero, editors), Barcelona, Vol. 3, 2437-2340.

Mairano, P. & Romano, A. (2007), Lingue isosillabiche e isoaccentuali: misurazioni strumentali su campioni di italiano, francese, inglese e tedesco, in Scienze vocali e del linguaggio. Metodologie di valutazione e risorse linguistiche, Atti del III Convegno Nazionale AISV (V. Giordani, V. Bruseghini & P. Cosi, editors), Torriana: EDK Editore, 119-134.

Mairano, P. & Romano, A. (2010), Un confronto tra diverse metriche ritmiche usando Correlatore, in La dimensione temporale del parlato, Atti del V Convegno Nazionale AISV, (S. Schmid, M. Schwarzenbach & D. Studer, editors), Torriana: EDK Editore, 79-100.

Missaglia, F. (1999), Contrastive prosody in SLA – an empirical study with adult Italian learners of German, in Proceedings of the 15th International Congress of Phonetic Sciences (J. Ohala et alii, editors), Berkeley: University of California, Vol. 1, 551-554.

Mok, P. & Dellwo, V. (2008), Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English, in Proceedings of Speech Prosody 2008 (P. Barbosa, S. Madureira & C. Reis, editors), Campinas, Brazil: Editora RG/CNPq, 63-66.

Pike, K. (1945), The Intonation of American English, Ann Arbor: University of Michigan Press.

Ramus, F., Nespor, M. & Mehler, J. (1999), Correlates of linguistic rhythm in the speech signal, Cognition, 72, 1-28.

Roach, P. (1998), Some languages are spoken more quickly than others, in Language myths (L. Bauer & P. Trudgill, editors), London: Penguin, 150-158.

Romano, A., Mairano, P. & Pollifrone, B. (2010), Variabilità ritmica di varietà dialettali del Piemonte, in La dimensione temporale del parlato, Atti del V Convegno Nazionale AISV, (S. Schmid, M. Schwarzenbach & D. Studer, editors), Torriana: EDK Editore, 101-112.

Russo, M. & Barry, W. (2010), Il Pairwise Variability Index (PVI e PVIs): valori ritmici per i dialetti italiani e per l'italiano regionale. Implicazioni tipologiche, in Prosodic universals. Comparative studies in rhythmic modeling and rhythm typology (M. Russo, editor), Roma: Aracne, 185-226.

Schmid, S. (2012), Phonological typology, rhythm type and the phonetics-phonology interface. A methodological overview and three case studies on Italo-Romance dialects, in Methods in contemporary linguistics (A. Ender, A. Leemann & B. Wälchli, editors), Berlin: Mouton de Gruyter, 45-68.

Tortel, A. & Hirst, D. (2010), Rhythm metrics and the production of English L1/L2, in Proceedings of Speech Prosody 2010, Chicago, accessibile al sito [http://speechprosody2010.illinois.edu/papers/100959.pdf].

Vayra, M., Avesani, C. & Fowler, C. (1984), Patterns of temporal compression in spoken Italian, in Proceedings of the 10th International Congress of Phonetic Sciences (M.P.R. Van den Broecke & A. Cohen, editors), Dordrecht: Foris, 541-546.

Weinreich, U. (1963), Languages in contact, 2nd ed., The Hauge: Mouton (trad.it. 1974, Lingue in contatto, Torino: Boringhieri).

White, L. & Mattys, S. (2007), Calibrating rhythm: first language and second language studies, Journal of Phonetics, 35, 501-522.

White, L., Payne, E. & Mattys, S. (2009), Rhythmic and prosodic contrast in Venetan and Sicilian Italian, in Phonetics and Phonology: Interactions and Interrelations (M. Vigario, S. Frota & M.J. Freitas, editors), Amsterdam: John Benjamins, 137-158.

TECNOLOGIE DELLA VOCE

AN ITALIAN EVENT-BASED ASR-TTS SYSTEM FOR THE NAO ROBOT

Piero Cosi*, Giulio Paci*, Giacomo Sommavilla*, Fabio Tesser*
Marco Nalin**, Ilaria Baroni**
*Institute of Cognitive Sciences and Technologies, ISTC, C.N.R., UOS Padova, Italy
** Hospital San Raffaele, Milano, Italy
*[piero.cosi, giulio.paci, giacomo.sommavilla, fabio.tesser]@pd.istc.cnr.it
**[nalin.marco, baroni.ilaria]@hsr.it

1. ABSTRACT

This paper describes an event-based integration approach for building a human-robot spoken interaction system using the NAO robot platform with the URBI middleware within the ALIZ-E project. The ALIZ-E integrated system includes various components but we mainly concentrate on the Automatic Speech Recognition (ASR) and the Text To Speech (TTS) synthesis modules while the following Natural Language Understanding (NLU), Dialog Management (DM) and Natural Language Generation (NLG) ones will be only briefly introduced. We describe these components and how we adapted and extended them for use in the system. We discuss several options that we have considered for the implementation of the interfaces and the integration mechanism and present the event-based approach we have chosen. We describe its implementation using the URBI middleware. The system has been be used for HRI experiments with young Italian users since April 2011.

2. INTRODUCTION

Conversational systems play an important role in scenarios without a keyboard, e.g., when talking to a robot. Communication in human-robot interaction ultimately involves a combination of verbal and non-verbal (gesture) inputs and outputs. Systems capable of such an interaction thus must process verbal and non-verbal observations in parallel, as well as execute verbal and non-verbal actions accordingly in order to exhibit synchronized behaviors. The development of such systems involves the integration of potentially many components and ensuring a complex interaction and synchronization between them.

NAO (Aldebaran Robotics, 2012; Gouaillier et alii, 2008) is one of the most often used humanoid robots for academic purposes worldwide and it is complete with a user-friendly programming environment. From simple visual programming to elaborate embedded modules, the versatility of NAO and its programming environment enables users to explore a wide variety of subjects at whatever level of programming complexity and experience.

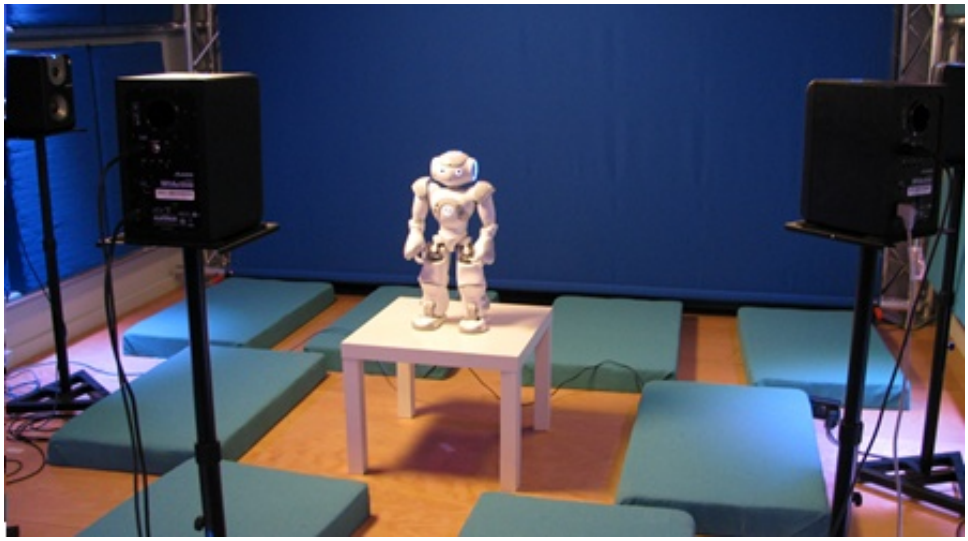


Figure 1: NAO, the ALIZ-E robot.

In this paper we present an event-based approach for integrating a conversational robotic system. This approach has been instantiated using the URBI middleware (URBI Open-Source, 2012; Baillie, 2005) on a NAO robot that is being used as a test bed for investigating child-robot interaction in the context of the ALIZ-E[1] project (ALIZ-E, 2012).

Within the ALIZ-E project the need for a robust software environment capable of performing ASR (Automatic Speech Recognition) and TTS (Text To Speech) synthesis for Italian children users interacting with NAO has been made manifest. In the ALIZ-E project, we choose alternatively SPHINX (CMU SPHINX, 2012; Lee et alii, 1990) or JULIUS (JULIUS, 2012; Lee and Kawahara, 2009) ASR and ACAPELA (ACAPELA, 2012) or MaryTTS (MARY TTS, 2012; Schröder, 2003) TTS, four complete state-of-the-art software suites that implement all the components needed for speech recognition and synthesis respectively. Also a VAD (Voice Activity Detection) module (Dekens and Verhelst, 2011) which triggers all speech processing activities on NAO has been integrated in the system. Moreover, URBI has been elected as the middleware that "orchestrates" the many components that form part of the ALIZ-E project architecture. URBI aims to be a universal, powerful and easy-to-use software for robotic programming. It is based on a client/server architecture providing a high-level interface for accessing the joints of the robot or its sensors.

Using the event-based approach to system integration introduced above, we have been developing a system in the ALIZ-E project integrating the following components: Audio Front-End (AFE), Voice Activity Detection (VAD), Automatic Speech Recognition (ASR), Natural Language Understanding (NLU), Dialogue Manager (DM), Natural Language Generation (NLG), Text-To-Speech Synthesis (TTS), Non-verbal Behavior Planning (NVBP) and motor Control (MC). The integration of the components is shown in Fig. 2: filled boxes indicate components implemented in Java, double-line boxes components in C/C++, and plain boxes components in URBIScript. The TTS component with the marble filled box is either the ACAPELA TTS on the NAO or the Mary TTS implemented in Java.

[1] ALIZ-E develops embodied cognitive robots for believable any-depth affective interactions with young users over an extended and possibly discontinuous period

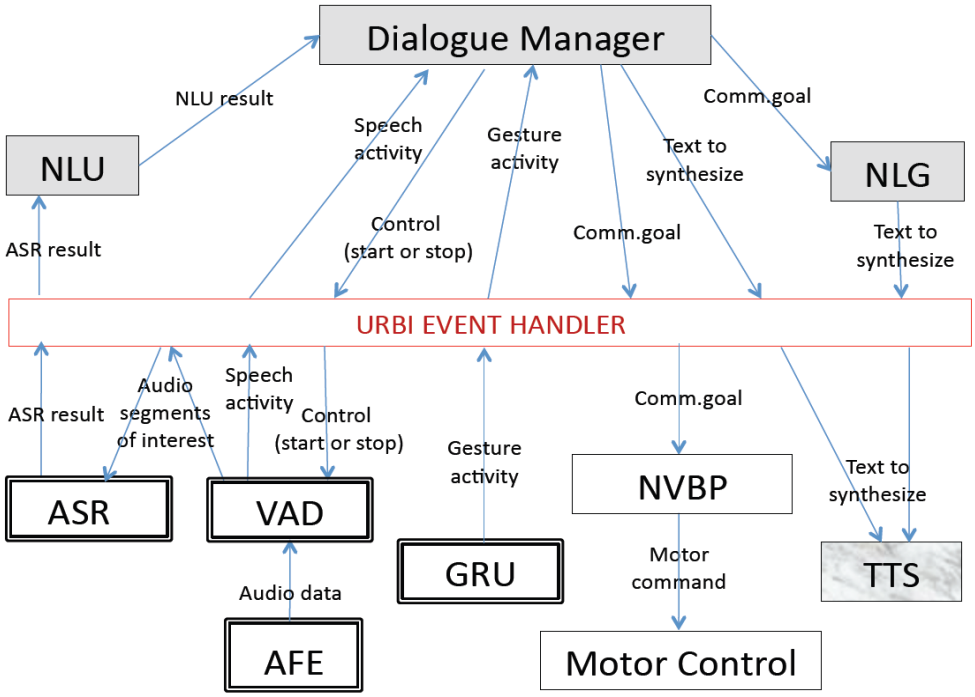


Figure 2: The components of the integrated system.

The system has been so far implemented for three scenarios: an imitation game of arm movements, a quiz game, and a dancing game and this paper focuses on the components developed for the imitation and quiz games, in which verbal interaction has a key role. The system for the dancing game applies the same integration approach. Experiments with Italian children in all three scenarios have been carried out in Milan at Hospital San Raffaele.

3. ASR

Two widely used open-source ASR systems have been considered for integration into the ALIZ-E framework: SPHINX and JULIUS.

SPHINX has been developed at Carnegie Mellon University (CMU) at Pittsburgh, while JULIUS has been developed at Kyoto University and both systems have to be considered as Large Vocabulary Continuous Speech Recognition (LVCSR) systems with real-time decoding capabilities.

Originally in the ALIZ-E project we used CMU SPHINX-3 (CMU SPHINX, 2012; Lee et alii, 1990) for speech recognition of Italian children users interacting with NAO but we are now using JULIUS because, in SPHINX-3, it was difficult to implement run-time features (live decoding) and to handle audio input. Moreover, SPHINX-3 is no longer maintained and, on the other hand, JULIUS has proven to be very easy to be integrated within the ALIZ-E project architecture.

3.1. JULIUS

Open-Source LVCSR Engine JULIUS is a high-performance ASR decoder for researchers and developers, designed for real-time decoding and modularity. Moreover most of the features available in other state-of-art decoders are also available for JULIUS, including major search techniques such as tree lexicon, N-gram factoring, cross-word context dependency handling, enveloped beam search, Gaussian pruning, Gaussian selection, etc.

JULIUS decoder main features are:

- open-source;
- small memory footprint;
- core engine is a separate C library;
- AM Models in HTK (HTK, 2012) ASCII hmmdefs format are supported;
- several LM types are supported: N-gram, grammar, isolated word and user-defined LM function embedding;
- modular configuration file structure (i.e., each config file can embed another one covering only one particular aspect of the configuration);
- it can perform decoding with parallel, multiple LMs and AMs: from the same audio input, a different result is given for every ac./lang. model;
- N-best / Word lattice / Confusion network output;
- word-level confidence scoring;
- real-time, on-the-fly, high speed, two-pass (forward-backward) decoding:
  - word bigrams are used for the first pass, while N-grams are used for the second one;
  - first pass output can be used before second one has finished;
- integrated GMM-based and Energy-based VAD;
- on-the-fly audio normalization: cepstral mean normalization (CMN), cepstral variance normalization (CVN), vocal tract length normalization (VTLN).

Most of these features are desirable in the ALIZ-E project.

3.2. SPHINX-3 and JULIUS-4 Comparison

In Table 1 a comparison between SPHINX-3 and JULIUS-4 is provided. The following elements in particular showed that JULIUS is more suitable than SPHINX-3 for the ALIZ-E project and drove the ASR engine change decision:

- JULIUS decoder API is very well designed (it made integration smoother in comparison with SPHINX-3);
- its system requirements are low (this is important in an integrated system handling several components);
- language models can be swapped at run-time;
- configuration is modular;
- possibility of performing multi-model recognition;
- on-the-fly Input/AM normalization.

Feature	Sphinx-3	Julius4
Open Source	yes	yes
System requirements	Computation and memory intensive (Each word has its own HMM)	Low memory requirement: less than 32MBytes required for work area (<64MBytes for 20k-word dictation with on-memory 3-gram LM)
Decoder API	no	yes
Decoder binary	yes	yes
AM formats	Sphinx	HTK
AM training	Sphinx	HTK
LM formats	ARPA N-gram Finite State Grammar	ARPA N-gram DFA grammar isolated word user-defined functions
Configuration	Monolithic Held for the entire execution	Modular Run-time swapping allowed
Parallel Multi-model recognition	no	yes
Confidence scoring	yes	yes
Integrated VAD	yes	yes
Audio normalization	MLLR VTLN MAP	MLLR (using external tools) VTLN CMN CVN
Audio normalization (on-the-fly)		VTLN CNN CVN
Output	N-best Word lattice	N-best Word lattice Confusion network

Table 1: SPHINX-3 and JULIUS-4 feature comparison.

3.3. Acoustic Model Training with HTK

The LVCSR Engine JULIUS distribution does not include specific training tools for acoustic models, however any tool that create acoustic models in the Hidden Markov Model Toolkit (HTK) format can be used. In the ALIZ-E Project we used the HTK tools (HTK, 2012; Young et alii, 2006) for this task, following the Voxforge HTK training for JULIUS tutorial (VoxForge, 2012b). The same procedure has been used to train an Italian adult acoustic model, using the training data provided for the EVALITA 2011 Forced Alignment task (Cutugno et alii, 2012) (a subset of the CLIPS corpus) and an Italian child acoustic model, using the FBK CHILD-IT corpus (Gerosa et alii, 2007). The training procedure is described in the following Table 2.

1. create a prompts file with the transcription of all audio files;
2. create a pronunciation dictionary that covers the whole prompts file (our own lexicon has been used; the lexicon has been extended with pronunciations of missing words using Sequitur G2P (Bisani and Ney, 2008), grapheme to phoneme tool followed by a quick manual revision);
3. obtain a list of all the used phones, including silence phones;
4. generate a phone level transcription from the orthographic transcriptions, using the first entry of the pronunciation dictionary;
5. extract 13 MFCC coefficients (plus Δ) from audio files (25ms Hamming analysis window, 10ms analysis step, preenphasis 0,97);
6. initialise model parameters for embedded training (13 normalized MFCC features plus Δ, 25ms Hamming analysis window, 10ms analysis step, preenphasis 0,97);
7. perform embedded training using HERest tool (Young et alii, 2006). Embedded training has the advantage that it does not require any prior knowledge of the phone boundaries and thus can be applied when only orthographic transcription is available (only symbolic phonetic transcription is required). HERest implementation of embedded training is implemented using the Baum-Welch algorithm. It starts by initialising to zero the accumulators for all the parameters of all the HMMs and then, for each training utterance, proceeds as follow:
   (a) joins all the HMMs of the phonetic symbols of the phonetic transcription, constructing a composite HMM;
   (b) calculates the forward and backward probabilities for the composite HMM;
   (c) calculates the probabilities of state occupation at each time frame using the forward and backward probabilities;
   (d) updates the accumulators in the Baum-Welch usual way;
   finally it uses the accumulators to estimate new parameters for all of the HMMs;
8. repeat the embedded training steps two times more;
9. add a short pause model by copying the learned silence model parameters and allowing short pauses between words;
10. repeat the embedded training steps two times more;
11. perform forced alignment using the current acoustic model, allowing multiple pronunciations, in order to detect the best one;
12. repeat the embedded training steps two times more using the new phonetic transcriptions;
13. convert the phonetic transcription (that uses context independent phones) to a context dependent one (using left and right phones as context);
14. repeat the embedded training steps two times more using the new phonetic transcriptions;
15. manually create a file with phones features derived from linguistic knowledge;
16. tie phone states according to phonetic questions and a lexicon of all the words to be recognised;
17. update the phonetic transcriptions according to the tied phone states;
18. repeat the embedded training steps two times more using the new phonetic transcriptions;

19. create a map for all the remaining possible phone states, so that words not in the training set nor in the recognition lexicon can be recognised in a later step.

Table 2. Acoustic Model Training Procedure.

3.4. Language modelling

The LVCSR Engine JULIUS supports N-gram, grammar and isolated word Language Models (LMs). Also user-defined functions can be implemented for recognition. However its distribution does not include any tool to create language models, with the exception of some scripts to convert a grammar written in a simple language into the Deterministic Finite Automaton (DFA) format needed by the engine. This means that external tools should be used to create a language model.

3.4.1 N-gram LM

The JULIUS engine supports N-gram LMs in ARPA format. We used SRI-LM toolkit (SRILM, 2012; Stolcke, 2002), to train a simple model for question recognition of the Quiz Game ALIZ-E scenario. The Quiz questions and answers database has been used as training material for this model. The model is very simple and very limited, but it should be enough to recognise properly read questions (the questions to be recognised are expected to be from the training set), especially if used in conjunction with some other, more flexible, model.

3.4.2 Grammar LM

The JULIUS engine distribution includes some tools that allow to express a Grammar in a simple format and then to convert to the DFA format needed by JULIUS. That format, however, has very few constructs that helps writing a proper grammar by hand and writing a non-trivial grammar is very hard. Third-party tools exist to convert an HTK standard lattice format (SLF) to the DFA format and to optimise the resulting DFA [8]. SLF is not suitable to write a grammar by hand, but HTK provides tools that allow a more convenient representation based on the extended Backus-Naur Form (EBNF) (Young et alii, 2006).

A simple model for Quiz answers recognition where written in the EBNF-based HTK grammar language. Part of the grammar was automatically derived by including the answers in the Quiz database. Several rules were added to handle common answers and filler words.

3.5 ASR Resources

In this section the resources used for Acoustic Model building and the work done for expanding them are described. Efforts have been made to create an ALIZ-E system test set, consisting of transcribed spontaneous speech (i.e., a corpus of speech plus transcription data recorded from ALIZ-E experiments).

3.5.1 .Speech Corpora

Two Italian and one English Speech Corpora have been tested so far with HTK and JULIUS:

- the training data provided for the EVALITA 2011 Forced Alignment task (Cutugno et alii, 2012) (this is a subset of the Italian CLIPS Corpus adult voices that counts about 5 hours of spontaneous speech, collected during map-task experiments, from 90 speakers from different Italian areas);
- Italian FBK ChildIt Corpus (Gerosa et alii, 2007) (this is a corpus of Italian children voice that counts almost 10 hours of speech from 171 children; each child reads

about 60 children literature sentences; the audio was sampled at 16 kHz, 16 bit linear, using a Shure SM10A head-worn mic);

- English Voxforge adult voices (Voxforge, 2012a) (the whole 16 kHz, 16 bit linear data set as it was in October 2011 has been used; the data set counts more than 80 hours of read speech from more than 600 speakers).

3.6 ASR Data collection

Both read and spontaneous speech has been collected for the ALIZ-E project. Read speech has been recorded according to guidelines similar to those that were used to collect the FBK ChildIt corpus, with the main goal of extending training material. Spontaneous speech has been collected during real interactions between children and NAO, with the main goal of creating a proper test set for further development.

3.6.1 Read Speech

Data collection of read speech is useful to enlarge our Corpus of audio plus transcription children data. The major advantage of collecting read speech is that obtaining the transcriptions corresponding to the audio is straightforward. Thus is a relatively low time consuming task (compared to transcribing spontaneous recordings). These data are meant to be used to train the Acoustic Model.

Several sessions has been recorded during a summer school near Padova. For the text of the recordings it has been decided to use the FBK Childit's prompts which are phonetically balanced sentences selected from children literature. For each session the input coming from the four NAO microphones, a close talk microphone (Shure WH20QTR dynamic head set) and a panoramic (AKG Perception 200, -10 dB, flat equalisation) one has been recorded. The close talk and the panoramic microphones were connected to a digital audio recorder (Zoom H4n Handy Recorder). Synchronisation of the sources has been granted using a chirp-like sound, played by an external loudspeaker, at the beginning of every utterance.

The goal within this project is to collect read speech for a total amount of ∼10 hours / ∼10000 utt. / 171 children, which is the size of FBK ChildIt. About 2 hours of children speech (∼2k utterances, 32 children) have already been collected and more will be collected in the future. NAO was very useful as its use helped to keep children attention alive.

3.6.2 Spontaneous Speech

Read Speech is very useful for expanding AM training data, but it is not well suited for building a reliable test set for ASR in the ALIZ-E project. A proper test set should consist of audio collected in a scenario as close as possible to the real one. For this reason it has been decided to manually transcribe and annotate some data collected during the Quiz game experiments that took place at Ospedale San Raffaele in December 2011 and March 2012.

The collected audio data consists of spontaneous speech recordings of children utterances produced during real interactions with NAO in a Wizard-Of-Oz modality. The database will be extended with data recorded from new experiments. The experiments consisted basically of a Q&A quiz game between the robot and the child. It starts with the robot asking questions (and subsequently providing possible answers) to the child, then, after more or less four answers, they exchange roles and the child became the asker. In the latter case, the child speech cannot be considered entirely "spontaneous". Anyway, since it is part of the real system interaction, we consider that those data can represent a good test set.

Sometimes the child is not sure if robot heard him/her, and repeat his/her input, thus in those cases we collect more than one sample utterance of the same sentence. Also, we noticed a few cases of barge-ins: the user knew the answer and answered without waiting for the options. The robot gave them anyway, and the child re-answered. Frequent cases of interjections were observed from the children: the user does not understand robot's answer (and usually says "Uh?").

It has been decided not to use NAO built-in microphones for recording, as those are low quality microphones. Moreover, two of them are placed under the speakers, one is placed on the robot's nape, near the fan, and all the four microphones record a lot of noise resulting from motors and electronic circuits. Instead, we used hand-free radio close talk microphone (Proel RM300H, Radio frequency range: UHF High Band 750-865 MHz, Microphone: headset HCM-2). This microphone has been selected in order to interfere as little as possible with child-robot interaction. The microphone has been connected to a Zoom H4n, allowing us to record both using a computer (the Zoom is used as an USB audio input interface) or saving audio data on a SD memory card.

The first experiments showed that there were no problems in convincing the kids to wear the microphone. Telling them that "it helps NAO hearing you" is enough. These experiments also seems to confirm the impression that radio wireless technology can be really useful for (1) giving a reasonable freedom of movement to the users and (2) sending clean speech data to the system PC.

We started transcribing the recorded experiments' utterances with the software Transcriber (Transcriber, 2012; Barras et alii, 1998), version 1.5.1. The fundamental rule for transcribing was to use correct orthographic lowercase words with punctuation, except for the following specific cases:

Numbers: do not transcribe them as digits, but in orthographic form (for ex. if heard /VOGLIO LA 1/, then transcribe: "voglio la uno");

Acronyms: as they are spelled: they can be pronounced:

(1) as a word (for ex. /AIDIESSE/: "AIDS_aidiesse");
(2) as a combination of names of letters and a word (for ex. /GEIPEG/: "JPEG_geipeg")
(3) only as the names of letters (for ex. /AIDS/: "AIDS_aids");

Letters: (for ex. /A/: "A", /BI/: "B", /ZETA/: "Z", /B/: "B b", /C/ (as in "cane"): "C k", /C/ (as in "cielo"): C TS);

Proper Nouns: put initial capital letter (for ex. /MARIO/: "Mario");

Interruptions: insert an hyphen as suffix (for ex. /VADO A CA/: "vado a ca-"); be aware that, depending on the context, there can be differences (for ex. /CHE BELLO IL CE/ (meaning "cesto") transcribe "che bello il ce-", otherwise (meaning "cielo"): transcribe "che bello il cie-");

Wrongly pronounced words: transcribe in uppercase the correct word, fol¬lowed by an underscore and the wrongly spelt word, trying to adhere to heard sounds as much as possible with Italian orthography (for ex. /PROABILMENTE/: "PROBABILMENTE_ proabilmente");

Loanwords: transcribe with correct (foreign) orthography (for ex. /UIKEND/: "week-end" /OCHEI/: "okay");

Foreign words: here again, try to adhere to heard sounds as much as possible with Italian orthography (for ex. /GION/: "JOHN_gion" /GARAGE/: (with the french /Z/) "GARAGE_ garaZ");

Speaker fillers: here follows a list of the most frequent fillers we encountered so far: (a) <EHM>, (b) <mmm>, (c) <mmm>, (d) <EH>, (e) <whispered->, (f) <-whispered>, (g) <lipsmack>, (h) <breath>, (i) <tongueclick>, (j) <laughter>, (k) <unintelligible>;

External noises: here follows a list of the most frequent noises we encountered so far: (a) <robot speaks>, (b) <robot moves>, (c) <garbage>, (d) <background voices>.

3.7 .JULIUS/URBI integration into the ALIZ-E system

One of the main problems in a robotic environment is how to deal with the need of having components that either:

(a) can access to low-level hardware details:
(b) perform heavy computations, and typically run on different, more powerful machines:
(c) should be coordinated concurrently:
(d) should react to (typically asynchronous) events.

Languages such as C/C++ can well suit low-level and heavy computational tasks, but it can be tedious to manage concurrency, network communication and event handling with those programming languages.

The open source URBI (URBI, 2012; Baillie, 2005) environment, a Universal Robotic Body Interface, developed by GOSTAI ALIZ-E project partner, provides a standard robot programming environment made of urbiscript scripting language which can orchestrate complex organisations of components named UObjects in highly concurrent settings.. It is based on a client/server architecture where the server is running on the robot and accessed by the client. The URBI language is a scripted language used by the client and capable of controlling the joints of the robot or access its sensors, camera, speakers or any accessible part of the machine.

It is relatively easy to make a C/C++/Java program accessible to the URBI language. The UObject API defines the UObject class and each instance of a derived class in C++ code will correspond to an URBI object sharing some of its methods and attributes. Basically one needs to "wrap" the upstream program into a C++ (or Java) class inheriting from Urbi::UObject, then bind in urbiscript the methods that should be accessible from there. Eventually, one needs to define events and how to handle methods concurrently.

3.7.1 URBI ASR component

In the ALIZ-E integrated system, ASR is provided as an API whose functions can be accessed by other components (e.g., the DM - Dialog Manager or the NLU - Natural Language Understanding module). When ASR output is available an event is launched and the result is provided as a payload, so that every components that needs this information can access it. ASR component is basically made up of two modules: a configuration structure (that holds data for AM and LM) and a main recognition loop function (also called "JULIUS stream"). The latter contains an internal VAD - Voice Activity Detection module and outputs second pass recognition result as an NBest list.

The principal methods of this component are: load/free/switch configuration; start/stop main recognition loop. A schema of function call and data exchange among ASR, DM and NLU components can be seen in Figure 3.

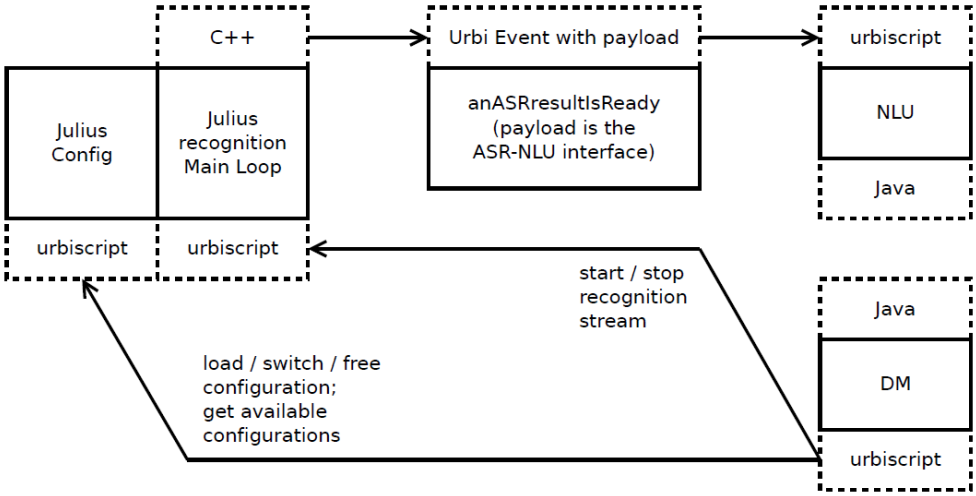


Figura 3. ASR communication through urbiscript.

3.8 ASR-NLU Interface

In order to communicate with other modules such as the NLU Java one, we needed to create an URBI data structure that could be populated by JULIUUS C++ UObject output, and that could be accessed by other components.

3.8.1 Data Exchange Format

Thanks to the collaboration with DFKI project partner, responsible for the ALIZ-E Natural Language Understanding and Dialogue Management tools, we implemented an NBest interface for feeding the NLU component with ASR output. This Data Structure consists of a list of sentences; each of them holds total (i.e. sentence-wide) acoustic and linguistic probabilities. Sentences are actually lists of words, and for every word an acoustic probability and a linguistic one are provided.

It is worth noting that Sphinx-3 actually provides by default acoustic and linguistic probabilities for both word- and sentence-level, while JULIUS outputs sentence-level acoustic and linguistic probabilities, and only a generic confidence score for word-level.

3.8.2 Communication Reliability

In order to implement this data exchange interface in the ALIZ-E integrated system we had to solve 3 major problems:

1. in the ALIZ-E integrated system, components communicate each other over the network, thus we need a way to encode the structured ASR output into something that can be sent through a serial communication channel, and then decode it (i.e. “restore” the data structure) on the receiver side;

2. suppose that the ASR component uses a shared variable to store its computation result and alerts other components which in turn will have to read the content. In this case is not possible to guarantee that the latter are able to do so before the variable is overwritten.
3. to overcome the above problem, one may consider using a complex structure (like a queue) to save successive outputs; in this way, however, there is the problem of emptying this data structure, as it may be difficult to know whether all the components that were to receive the recognition result have read the contents of the variable.

The URBI middleware layer provides features that address these problems. Problem 1 is solved by serializing ASR output data structure. Regarding Problems 2 and 3, we safely implemented this interface thanks to the event-based programming paradigm. With URBI, setting triggering events and reacting to them is straightforward; also, the following conditions are guaranteed:

- messages are passed one-to-many (multicast or broadcast);
- messages are transferred reliably (multicast protocol reliability means that the system ensures total order, atomicity and virtual synchrony);
- messages are guaranteed to be delivered in order;
- messages are passed asynchronously: the sender delivers a message to the receivers, without waiting for them to be ready (this is well suited in robotic environments where messages are generated at irregular intervals).

In our current implementation the result of ASR is sent to NLU component, but, since URBI allows multicast signaling, events can easily be caught by more than one component in the integrated system.

Actually, the C++ JULIUS UObject sends (or "emits") an "ASRresultIsReady" URBI event, which is caught by urbiscript code that triggers NLU functions. The ASR output Data Structure (NBest list) is embedded to the event as a "payload". This means that the event carries a message for the receiving components, ensuring atomicity in the communication. Figure 4 shows a diagram in which the ASR component emits an event with payload that is caught by NLU.

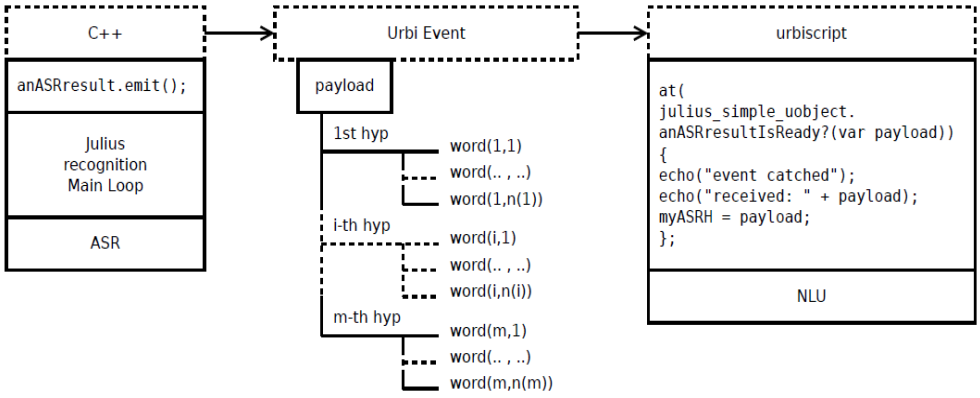


Figure 4. ASR to NLU event with payload.

4. TTS

Within the ALIZ-E project, the robot have to convey different messages to the child. Movements/gestures, lights and audio/speech are the different output channels available on the NAO robot (Gouaillier et alii, 2008) and they are used to communicate the desired message in this project. With reference to the voice channel only, it is known that a lot of messages are contained in the speech signal. Table 3 shows the main messages and the related speech correlates. These messages are encoded by particular acoustic patterns recognizable in the human speech and it would be attractive for a speech synthesizer, to be able to synthesize these patterns in order to communicate to the child all these different messages using the audio channel.

Message	Speech correlates
speaker identity	spectral envelope, voice quality
emotional state of the speaker	spectral envelope, voice quality, prosody
verbal content of the speech	spectral envelope
syntactic information	spectral envelope, prosody
focus	prosody

Table 3. Messages contained in the speech signal and their acoustic correlates. Here the term focus refers to the part of a sentence which expresses the centre of attention.

Moreover emotional speech synthesis must take into account the manipulation of paraverbal parameters like speech rate, voice intensity, pause durations, etc.

4.1 HMM or Statistical Parametric Synthesis and its use on a robotic system

The Statistical Parametric Synthesis (or HMM Speech Synthesis) (Zen et alii, 2009) approach has been chosen for the task of modeling the voice of the robot, because it allows to act on the produced acoustic patterns in various ways and it seems the most suitable solution allowing stronger parameter control than Unit Selection synthesis. For example using HMM speech synthesis technology it is possible to:

- change the speaker identity of the synthetic voice; this is possible changing the vocoder (Imai, 1983; Fukada et alii, 1992) parameters or alternatively using speaker adaptation techniques (Yamagishi et alii, 2009);
- stress the focus of a sentence; applying some particular prosodic patterns;
- change the emotional content of the synthetic speech applying different prosodic settings and patterns;

The prosodic settings and patterns mentioned beforehand can be either the results of previously acquired knowledge and experience (e.g., it is knows that happy voices usually adopt an higher pitch with respect to a normal voice) or they can be the results of modules able to learning these from real data.

Some TTS customizations have been already implemented and used in the ALIZ-E project, the first is used to provide the robot with a child-like voice, the others refer to the possibility to apply prosodic modifications according to the focus or to produce a speech that reflect a particular emotional state of the robot.

Finally, in order to obtain emotive speech, HMM trajectory estimation techniques must be coupled with digital signal processing algorithms and speech models capable of imple-

menting voice quality and timbre modifications (Tesser et alii, 2010) as well as general pitch shifting and time stretching algorithms. In fact, while it is true that HMM-based speech synthesis allows for more flexible voice control, data-driven speech synthesis allows for more natural sounding voice qualities.

4.2 Robot voice identity

Synthesized speech triggers social identification processes, for this reason within the ALIZ-E Project we would like to use the voice of a child for NAO. A vocal tract scaler, which can simulate a longer or shorter vocal tract, has been used in order to obtain a child-like voice, starting from a female voice.

In this implementation the frequency axis warping method has been used. The resulting voice is good for this task because the robot doesn't need to have a realistic child voice and some artifacts can be accepted.

Anyway, using the HMM synthesis approach, some improvements in this task are possible, for example:

- to use a vocal tract scaler effect based on vocoder used in system (MLSA filter (Imai, 1983; Fukada et alii, 1992) with mixed excitation (Yoshimura et alii, 2001);
- to use speaker adaptation techniques for HMM synthesis (Yamagishi et alii, 2009)

4.2 Focus prominence by prosodic modification

The Natural Language Generation module is able to mark focus words during the verbal output generation process. In the speech synthesis process we are interested in emphasizing the focus using adequate speech parameters. As HMM synthesis technology is suited to prosody modifications, a first implementation has been done using some tags able to force the relative prosody changes in the words that bears the focus. After some informal listening test, the prosody on the focus words are forced in the following way:

- the speech rate is decreased of 10% with respect to the normal production;
- the pitch is raised of 25% with respect to the normal production.

4.3 Emotional prosodic modification

The ALIZ-E system is able to decide when the verbal output should be rendered with (non-neutral) emotional coloring, either "sadly" or "happily".

According to this, the speech paralinguistic feedback is realized increasing the speech rate (+5%) and the pitch contour (+25%) in the happy case, while in the sad case the speech rate and pitch contour are both decreased (-20%).

4.3 Italian voice for Mary TTS

MaryTTS (MARY TTS, 2012; Schröder, 2003) is an open-source, multilingual Text-to-Speech Synthesis platform written in Java. It was originally developed as a collaborative project of DFKI's Language Technology lab and the Institute of Phonetics at Saarland University and is now being maintained by DFKI. As of version 4.3, MaryTTS supports German, British and American English, Telugu, Turkish, and Russian and more languages are in preparation. MaryTTS comes with toolkits for quickly adding support for new languages and for building unit selection and HMM-based synthesis voices.

Within the ALIZ-E project we have pursued the route of making available an Italian MaryTTS female voice for the robot. We started with the porting of some of the existing Italian FESTIVAL TTS modules (Cosi et alii, 2001, Tesser et alii, 2005). An Italian lexicon

for MaryTTS has been created converting the Italian FESTIVAL lexicon and an Italian Letter-To-Sound (LTS) module has been obtained together with a first simple Part Of Speech (POS) tagger.

The latest official version of MaryTTS (4.3.1) contains the support for Italian and the istc-lucia-hsmm voice. It can be downloaded from http://mary. opendfki.de/wiki/4.3.1.

4.4 Text corpus selection

In order to select the text scripts for the recordings we have used the automatic procedure for optimal (phonetically/prosodic balanced) text selection made available in MaryTTS (Pammi et alii, 2005), based on the analysis of the freely available Wikipedia dumps for the language taken into consideration.

The original MaryTTS procedure has been modified to select only sentences for whose it was possible to obtain a phonetic transcription using only the lexicon. The final text selection has been obtained by the iteration (4 times) of the following steps:

- ignore all sentences that do not improve the coverage score;
- manual inspection of the selected list and removal of the too-difficult-to-pronounce sentences;
- reiterate the coverage selection procedure.

Table 4 shows the technical details of the obtained text corpus.

Feature	Description
Wikipedia dump date	2011/08/15 (20110815194113134300 62)
DB Size (sent.)	1400
Coverage method	SimpleDiphones+SimpleProsody

Table 4. Description of the Italian Text corpus for Mary TTS.

4.5 Recordings

In order to do not fatigue the speaker's voice, the recordings has been done in several sessions spanned in two weeks. Table 5 shows the technical details of the recordings.

Feature	Description
Speaker	Female
Age	20
Room characteristics	Silent room
Microphone	Shure WH20QTR Dynamic Headset
O.S.	Linux Ubuntu
Soundcard	Focusrite Saffire LE FireWire audio interface
Audio driver	Jack Sound Server trough Pulse Audio
DB Size (sentences)	1400
DB Size (time)	□2 hours
Manually checked segmentation	Only on sentences identified by a quality control check

Table 5. Description of the Lucia TTS recording corpus.

4.6 Building the voices

Both Unit Selection and HMM voices have been created using the Voice Import Tools under the MaryTTS environment. The building process consists of the following steps:

- feature extraction from acoustic data;
- feature vector extraction from text data;
- automatic labeling (ehmm from Festvox);
- Unit Selection voice building;
- HMM voice building (SPTK, 2012; HTS, 2012).

The resulting voices was positively judged by some informal listening test with the following comments:

- the Unit Selection voice has good audio quality, but sometimes the voice is cracked/chunked, probably because of some missing units in the corpus;
- the HMM voice has a lower audio quality, but it has an higher intelligibility and constant quality with respect to the Unit Selection voice.

4.7 TTS Integration into the ALIZ-E system

With regards to integration we wanted to keep the possibility of using both TTS systems: ACAPELATTS (ACAPELA, 2012) and MaryTTS (MaryTTS, 2012). We achieve this implementing a configuration system able to choose between the two different TTS system to use.

Since ACAPELATTS is already available on NAO, most of the work has been done for making available MaryTTS into the NAO/URBI environment.

MaryTTS is a platform written in Java using the client/server paradigm. Due to the NAO CPU resources limitations, it has been decided to run the MaryTTS server on the remote PC while keeping the ACAPELA TTS as a built-in system on the NAO Robot.

When MaryTTS produces the audio stream the resulting speech must be played on the NAO loudspeaker. This has been achieved using a streaming server based on GStreamer (GStreamer, 2012). In order to have a real time interaction, a Real-time Transport Protocol (RTP) (RTP-Wikipedia, 2012) streaming server is active on NAO. URBI is able to manage RTP data as well, but we have decided to follow the GStreamer approach in order to avoid URBI server overloading. Moreover this approach allows to choose among a lot of already available plug-in for audio/video pipelines building.

In order to bring MaryTTS and GStreamer RTP in the URBI world, an URBI Object (UMaryTTS) has been created as the principal Object responsible for routing the synthesis request (Mary TTS client) and for playing the resulting audio to different output channels.

These channels are represented by the following Urbi Objects:

- UMaryTTSAudioPlayer is an UObject that makes a request to the MaryTTS server and play the resulting synthesized voice through the PC loudspeakers (useful for the fake robot simulation);
- UMaryTTSAudioPlayer is an UObject that makes a request to the MaryTTS server and streams the resulting synthesized audio through an RTP connection[2] using the efflux library (efHux-java, 2012);

[2] This connection is created and destroyed every time a sentence is synthesized.

- UMaryTTSGstreamerPlayer is an UObject that makes a request to the MaryTTS server and stream the resulting synthesized audio through an UDP RTP permanent connection[3] using GStreamer-java (GStreamer-java, 2012).

While ACAPELA TTS exposes the functions: say(pText) and stopTalking(), UMaryTTS also exposes sayWithEmotion (pText, pEmotion) that is able to change the global prosody setting according to the emotion taken into consideration. Moreover UMaryTTS is able to manage the focus[4] generated by the NLG module.

7. FUTURE PLAN

As for ASR, since mismatches between training and test conditions can severely degrade the performance, one can apply Speaker Adaptation to ASR, using a small amount of observed data from an individual speaker to improve a speaker-independent model.

Adaptation can be very useful in the ALIZ-E project because children's vocal tract lengths can differ a lot. As the child will interact several times with the robot we can consider of reusing data from previous interactions for adapting the models. Preliminary experiments using SPHINX have shown recognition improvements on ChildIt corpus using for example VTLN (Vocal Tract Length Normalization) and MLLR (Maximum Likelihood Linear Regression) adaptation techniques (Nicolao and Cosi, 2011).

ASR test set such audio recordings have to meet very precise requirements to be considered reliable for our purposes:

- they should be collected from children (as much spontaneous as possible) speech;
- they must reflect a real user case project scenario.

Due to the these reasons, we have not yet been able to evaluate ASR and NLU accuracy because we lack of test data. Nonetheless, this is a major goal for our purposes. We are currently involved in speech data collection in various ALIZ-E Experiments conducted at HSR.

An AI Component based on the Dialogue or the Memory Manager of the ALIZ-E integrated system could be conceived in order to "choose" a specific LM according to its current status. Such specific models should be available before the interaction starts.

For example, if the interaction is in a preliminary stage, the AI component could prefer to load a LM that is built upon greetings and introductory speech forms. Fig. 5 shows a schema of this idea. Also LM online generation which is related to the previous idea cold be exploited assuming that AI components in the system can have access to text data (collected from children real speech) and can build LMs at runtime from specific subsets of those data using AI knowledge, the interaction status and the history (memory).

In order to improve JULIUS recognition performance we aim to collect more data for the ChildIt2 corpus, audio plus transcription, which will be used to train a better, more robust children AM for Italian.

[3] We experienced that a permanent RTP connection is more stable than a single shgot conncetion for each sentence, then UMaryTTSGstreamerPlayer is more stable than UMaryTTSAudioPlayer.

[4] The special characters § and ^ are treated as symbols to identify the focus part in a sentence.

The input of our JULIUS component is currently fed by the capture device of ALSA, and we want to implement a connection between the input of the recognizer and the Audio Front End (AFE) module. This will bring more flexibility to the system, and the possibility of testing the four NAO built-in microphones.

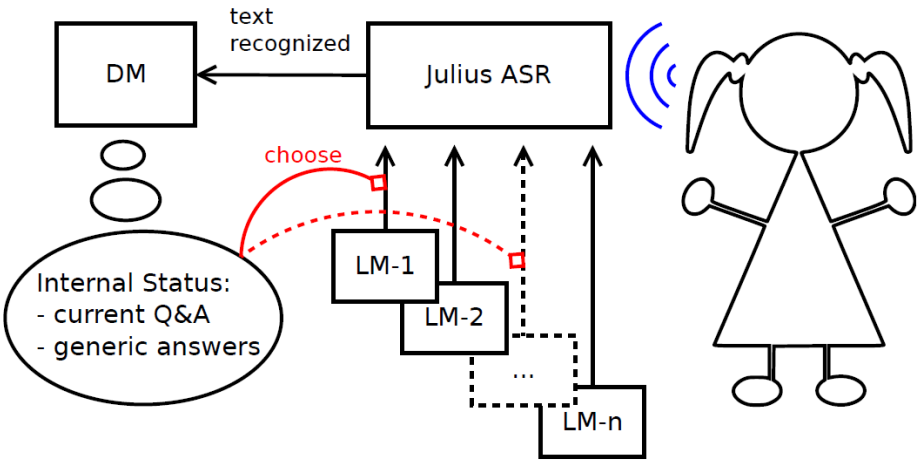


Figure 5. Run-Time LM switching schematic representation.

AM/voice adaptation could be used in our system to improve recognition accuracy. One idea could be to adapt the AM with a portion of speech pronounced at the beginning of a interaction. The adapted AM could be used for the rest of the dialogue.

The recognition engine currently works with a fixed pronunciation dictionary and we want to build a letter-to-sound module for Italian, in order to phonetize (and thus be able to recognize) new words and we are considering to test two l2s engines for the Italian language (trained upon our database):

- MaryTTS letter2sound;
- Sequitur;

As for TTS, in order to improve the current use of speech synthesis inside the ALIZ-E project, the main effort will be spent on improving the coupling between the Natural Language Generation (NLG) module and the speech generation using a more sophisticated input for the TTS. We will test some ideas about the use of automatic prominence detection (Rosenberg, 2010) inside the HMM voice building creation process; this can results in a method that allows to force some prosodic symbols like ToBI (Silverman et alii, 1992) together with text in input to the TTS.

Besides, we would like to investigate strategies to achieve a better child-like voice for the robot inside the ALIZ-E project. For example we plan to evaluate the building of a MLSA-based vocal tract effect, and testing the HMM speaker adaptation/voice conversion algorithms.

Regarding the Italian MaryTTS system, the following improvements will be evaluated:

- to build a more robust POS tagging module for Italian;
- to build a text normalization module able to handle digits, acronyms, etc.;
- to re-check the phonetic prosodic coverage of the corpus;

- to re-run the text scripts selection by using these more evolved Natural Language features;
- if necessary other recordings with the same speaker will be taken into account.

6. CONLUDING REMARKS

We introduced an event-based integration approach for building a human-robot spoken interaction system using the NAO robot platform with the URBI middleware within the ALIZ-E Project. In particular, we focused on how we adapted and extended the JULIUS Large Vocabulary Continuous Speech Recognition (LVCSR) system and the MaryTTS Text To Speech (TTS) synthesis modules. Their final integration into the system is a first important mark toward the implementation of a fully integrated communication system for NAO.

We discussed several options considered for the implementation of the interfaces and the integration mechanism and presented the event-based approach we have chosen. We described its implementation using the URBI middleware.

The system has been be used with success for HRI experiments with young Italian users since April 2011.

ACKNOWLEDGMENTS

Parts of the research reported on in this paper were performed in the context of the EU-FP7 project ALIZ-E (ICT-248116).

INDEX TERMS: Human-Robot Interaction (HRI), integration, NAO, URBI, Italian children Automatic Speech Recognition (ASR) , Italian Text-To-Speech (TTS) synthesis, Voice Activity Detection (VAD), Dialogue Management (DM), Natural Language Generation (NLG), Non-Verbal Behavior Generation (NVBG).

REFERENCES

ACAPELA (2012), URL: http://www.ACAPELA-group.com/index.html.

Aldebaran Robotics (2012), URL:: http://www.aldebaran-robotics.com/en.

ALIZ-E (2012), URL: http://ALIZ-E.org/.

Baillie, J. C. (2005), "URBI: Towards a Universal Robotic Low-Level Programming Language", in IEEE/RSJ International Conference on Intelligent Robots and Systems, 2005, 820-825.

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (1998), Transcriber: A Free Tool for Segmenting, Labeling and Transcribing Speech, Proc. Of the First International Conference on Language Resources & Evaluation (LREC), Granada, Spain, 1998, 1373-1376,

Bisani, M., and Ney, H. (2008), Joint-sequence models for grapheme¬to-phoneme conversion, Speech Communication 50.5 (2008), 434-451.

CMU SPHINX (2012), URL: http://cmuSPHINX.sourceforge.net/.

Cosi, P. and Nicolao, M. (2009), "Connected Digits Recognition Task: ISTC-CNR Comparison of Open Source Tools", in CD Proceedings of EVALITA Workshop 2009, in Post-

er and Workshop CD Proceedings of the XI Conference of the Italian Association for Artificial Intelligence, 2009, Reggio Emilia, Italy, December 9-12.

Cosi, P., Tesser, F., Gretter, R. and Avesani, C. (2001), "Festival Speaks Italian!", in Proceedings of Eurospeech'01, , 2001, Aalborg, Denmark, September 3-7, 509-512.

Cutugno, F., Origlia, A. and Seppi, D. (2012), EVALITA 2011: Forced alignment task. Tech. rep. 2012. URL: http://www.evalita.it/sites/evalita.fbk.eu/files/working_notes2011/Forced_Alignment/FORCED_ORGANIZERS.pdf.

Dekens, T. and Verhelst, W. (2011), "On the Noise Robustness of Voice Activity Detection Algorithms", Interspeech 2011, Florence, Italy, 2649-2652.

efflux development team (2012), efflux: a Java RTP Stack with RTCP Support and a Clean API, Mar. 2012, URL: https://github.com/brunodecarvalho/efflux.

Fukada T., Tokuda, K., Kobayashi, T., and Imai, S.(1992), An Adaptive Algorithm for Mel-Cepstral Analysis of Speech, Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1992, Vol. 92-1, 137–140.

Gerosa, M., Giuliani, D. and Brugnara, F. (2007), "Acoustic variability and automatic recognition of children's speech. Speech Communication, 49 (Feb 2007), 847-860.

Gouaillier, D., Hugel, V., Blazevic, P., Kilner, C., Monceaux, J., Lafourcade, P., Marnier, B., Serre, J. and Maisonnier, B. (2008), "The NAO Humanoid: a Combination of Performance and Affordability", CoRR, 2008, abs/0807.3223 (http://arxiv.org/abs/0807.3223).

GStreamer Development Team (2012), GStreamer Open Source Multimedia Framework, Mar. 2012, URL: http://gstreamer.freedesktop.org/.

GStreamer-java Development Team (2012), Java Interface to the GStreamer Frame-Work, Mar. 2012, URL: http://code.google.com/p/gstreamer-java/.

HTS Working Group (2012), HMM-Based Speech Synthesis System (HTS) version 3.2. Mar. 2012, URL: http://hts.sp.nitech.ac.jp/.

HTK - Hidden Markov Model Toolkit (2012), URL: http://htk.eng.cam.ac.uk/

Imai, S., (1983), Cepstral Analysis Synthesis on the Mel Frequency Scale, Proc. IEEE ICASSP 1983, Vol. 8., 93-96.

JULIUS (2012), URL: http://JULIUS.sourceforge.jp/en_index.php.

Lee, A. and Kawahara, T. (2009), "Recent Development of Open-Source Speech Recognition Engine JULIUS", Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2009.

Lee, A., Kawahara, T. and Shikano, T., (2001), Julius - an Open Source Real-Time Large Vocabulary Speech Recognition Engine, Proc Interspeech 2001, 1691-1694.

Lee, K.F., Hon, H. W. and Reddy, R. (1990), "An Overview of the SPHINX Speech Recognition System", IEEE Transactions on Acoustics, Speech and Signal Processing, January 1990, 38(1): 35-45.

Mary TTS Development Team (2012) The MARY Text-to-Speech System, Mar. 2012, URL: http://mary.dfki.de/.

Nicolao, M. and Cosi, P. (2011), "Comparing SPHINX vs. SONIC Italian Children Speech Recognition Systems", in Abstract Book & CD-Rom Proceedings of AISV 2011, 7th Conference of Associazione Italiana di Scienze della Voce, "Contesto comunicativo e variabilità nella produzione e percezione della lingua", Lecce, Italy, 2011 - Abs: 85 - (CD: 414-425).

Pammi. S., Charfuelan, M., and Schröder, M. (2005), Multilingual Voice Creation Toolkit for the MARY TTS Platform, Proc. Int. Conf. Language Resources and Evaluation, LREC 2005, URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/720_Paper.pdf.

Rosenberg, A. (2010), AuToBI - A Tool for Automatic ToBI annotation, Proc. Interspeech 2010, 146–149.

RTP-Wikipedia (2012), Real-time Transport Protocol. Mar. 2012. URL: http://en . wikipedia.org/wiki/Real-time_Transport_Protocol/.

Schröder M. and Trouvain, J. (2003), "The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching", International Journal of Speech Technology, 6, 2003, 365-377.

Silverman, K., Beckman, M. Pitrelli, J., Ostendorf, M., Wightman, C., Price, P. Pierrehumbert, J., Hirschberg, J. (1992), TOBI: A Standard for Labeling English Prosody, in Ohala, J.J. et al. (eds.), Proceedings ICSLP 92, 867-870.
URL: http://www.cs.columbia.edu/\~{}julia/papers/TOBI_i92_0867.pdf.

SPTK Working Group (2012), Speech Signal Processing Toolkit (SPTK) version 3.2, Mar. 2012, URL: http://www.sp-tk.sourceforge.net/.

SRILM - The SRI Language Modeling Toolkit. URL: http://www.speech.sri.com/projects/srilm/.

Stolcke, A. (2002), SRILM - An Extensible Language Modeling Toolkit, Proc. Intl. Conf. on Spoken Language Processing, Denver, USA, vol. 2, pp. 901-904.

Tesser, F., Cosi, P., Drioli, C. and Tisato, G. (2005), "Emotional Festival-Mbrola TTS Synthesis", in Proceedings of Interspeech'05, 2005, Lisbon, Portugal, 505-508.

Tesser, F., Zovato, E., Nicolao, M. and Cosi, P. (2010), "Two Vocoder Techniques for Neutral to Emotional Timbre Conversion, in Sagisaka, Y. and Tokuda, K., eds., Proceedings of 7th Speech Synthesis ISCA Workshop (SSW), Kyoto, Japan, 2010, 130–135.

Transcriber: a Tool for Segmenting, Labeling and Transcribing Speech (2012), Authors: Boudahmane, K., Manta, M., Antoine, F., Galliano, S., and Barras C. (2012), URL: http://trans.sourceforge.net.

URBI Open-Source (2012), URL: http://www.gostai.com/products/URBI/.

Yamagishi, J., Nose, T., Zen, H., Ling, Z., Toda, T., Tokuda, K., King, S., and Renals S., (2009), A Robust Speaker-Adaptive HMM-based Text-to-Speech Synthesis, IEEE Trans. Audio, Speech, & Language Processing, vol.17, no.6, 2009, 1208-1230.

Yoshimura, T.. “Mixed Excitation for HMM-based Speech Synthesis”. In: Eurospeech. 2001.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2001), Mixed Excitation for HMM-Based Speech Synthesis, Prpc. EUROSPEECH 2001, 2263-2266.

Young, S.J, Evermann, G., Gales, M.J.F., Kershaw, D., Liux, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valchev, V. and Woodland, P. (2006), The HTK Book, version 3.4. Cambridge, UK: Cambridge University Engineering Department, 2006.

Zen, H., Tokuda, K., and Black, A. W. (2009), “Statistical parametric speech synthesis”, Speech Communication, 51(11), 2009, 1039-1064.

VoxForge (2012a), Tutorial: Create Acoustic Model - Manually. Mar. 2012. URL: http://www.voxforge.org/home/dev/acousticmodels/linux/create/htkjulius/tutorial.

VoxForge (2012b), Free Speech Recognition (Linux, Windows and Mac) - vox-forge.org. Mar. 2012. URL: http://www.voxforge.org/.

UN AMBIENTE INFORMATICO PER IL CONTROLLO DEI PROCESSI RELATIVI ALLA CONSERVAZIONE ATTIVA IN UN ARCHIVIO DIGITALE DI CORPORA VOCALI

Federica Bressan[1], Pier Marco Bertinetto[1], Chiara Bertini[1], Cristina Bertoncin[2], Francesca Biliotti[2], Silvia Calamai[2], Sergio Canazza[3], Nadia Nocchi[1]
[1]Scuola Normale Superiore, [2]Università degli Studi di Siena (sede di Arezzo), [3]Università di Padova
grafo@sns.it

1. SOMMARIO

Negli ultimi trent'anni si è andata diffondendo la sensibilità verso il tema della conservazione dei documenti sonori, riconosciuti nel loro valore di bene culturale. Parallelamente, si sono arricchiti il dibattito etico e gli strumenti per attuare le pratiche conservative. Ad oggi, tuttavia, il patrimonio costituito dai documenti conservati in alcune tipologie di archivi, custodi principalmente di registrazioni in ambito etno-musicologico e linguistico, è minacciato dalla progressiva erosione delle informazioni causato dal degrado dei supporti e da inadeguati strumenti metodologici di conservazione. Questo articolo presenta alcune soluzioni proposte dal progetto di ricerca *Grammo-foni. Le soffitte della voce* (Gra.fo) che si occupa di archivi di corpora vocali in area toscana. In particolare, vengono descritti gli strumenti informatici, alcuni sviluppati *ad hoc*, che costituiscono l'attuale ambiente di gestione dell'archivio digitale del progetto. L'elemento innovativo consiste nell'utilizzo di un approccio sistemico, in grado di combinare una serie di procedure semi-automatizzate per la generazione, la descrizione dell'archivio e la distribuzione delle informazioni con strumenti per il controllo di qualità.

2. INTRODUZIONE

Il ruolo fondamentale che le registrazioni sonore rivestono negli studi di linguistica è ormai pienamente riconosciuto. Tali registrazioni sono il prodotto di una tecnologia che si è andata evolvendo negli ultimi centocinquant'anni. Oggi le tecniche di registrazione sono potenzialmente in grado di restituire una qualità sonora percettivamente paragonabile a quella della sorgente acustica originale. Purtroppo i supporti fisici sui quali vengono conservate le informazioni acustiche, per la loro natura chimico-fisica, subiscono un processo di degrado relativamente rapido, misurabile nell'ordine di decenni o talvolta di pochi anni. Inoltre, l'equipaggiamento tecnologico utilizzato per l'ascolto o per il trasferimento analogico-digitale e i sistemi informatici che intervengono nell'eventuale fase di restauro sono soggetti a (micro-)malfunzionamenti che provocano alterazioni spesso difficili da individuare e correggere; i tecnici che adoperano le apparecchiature possono incorrere in errori operativi; inoltre la scelta dei dispositivi e la loro regolazione assecondano i mutamenti dell'ascolto e del gusto estetico del proprio tempo. In altre parole, non è ipotizzabile una neutralità del trasferimento dell'informazione acustica da un *medium* ad un altro *medium* (processo di rimediazione), unica soluzione all'inesorabile erosione delle memorie sonore.
Recenti rapporti di lavoro dell'UNESCO (Edmonson, 2002) affermano che oltre la metà del patrimonio culturale è a forte rischio di scomparsa nonostante l'attenzione per il tema della conservazione dei beni culturali, anche da parte della Comunità Europea che ha dimostrato grande sensibilità nel finanziare numerosi progetti di ricerca in questo campo. Molti fattori ostacolano la salvaguardia dei beni sonori e visivi, principalmente l'ingente investimento di

risorse materiali e umane richieste dalle campagne di digitalizzazione, oltre a gruppi di lavoro con competenze multidisciplinari, difficili e costosi da formare. Il risultato è che oggi molti archivi, soprattutto quelli di piccole dimensioni, sono completamente sprovvisti degli strumenti metodologici e tecnologici per tutelare adeguatamente il proprio patrimonio.
In particolare, gli archivi custoditi da alcuni dipartimenti di ricerca, da soggetti privati, e gli archivi che custodiscono documenti sonori di interesse per una categoria ristretta, sono maggiormente soggetti a pratiche di conservazione inadeguate. Gli archivi composti da documenti di indagine etno-musicologica e linguistica rientrano pienamente in questa casistica. Sul numero totale di progetti di ricerca che hanno ricevuto finanziamenti europei negli ultimi dieci anni, solo un'esigua percentuale si è occupata di tali archivi (ad esempio, Junge Kuenstler Europas gestalten Maerchen, 2001-2002, e Preservation and On-line Fruition of the Audio Documents from the European Archives of Ethnic Music, 2005-2006).
I documenti sonori di interesse linguistico, e in particolare dialettologico, rappresentano un caso particolarmente critico da trattare dal punto di vista conservativo, poiché si tratta di registrazioni raccolte in condizioni ambientali non favorevoli, con una tecnologia di livello non professionale e con supporti ad alto rischio di degrado. Inoltre, la salvaguardia di questo materiale non si limita alle politiche di conservazione a lungo termine, bensì comprende l'insieme delle azioni volte a permettere e incoraggiare l'accesso da parte di utenti esperti e generici ai documenti sonori per scopi di studio, ricerca, reinterpretazione, intrattenimento e altro.
Questo articolo presenta l'esperienza del progetto di ricerca *Grammo-foni. Le soffitte della voce* (*Gra.fo*), 2011-2013, finanziato dalla Regione Toscana (PAR FAS 2007-13, linea d'azione 1.1.a.3), condotto dal Laboratorio di Linguistica "Giovanni Nencioni" della Scuola Normale Superiore di Pisa e dall'Università di Siena. Gli obiettivi del progetto comprendono la raccolta sul territorio nazionale di registrazioni di rilevanza per l'area linguistica toscana, la creazione di un archivio digitale per la conservazione a lungo termine dei documenti trattati in un laboratorio allestito all'interno del Laboratorio di Linguistica, e la catalogazione dei materiali sonori, e la loro parziale trascrizione ortografica e fonetica. Le registrazioni custodite dagli archivi che finora hanno aderito al progetto sono copie uniche nella quasi totalità dei casi e molte di queste sono fondamentali per la ricostruzione del quadro linguistico dell'area regionale toscana a partire almeno dagli anni Sessanta. Il progetto si distingue per essere uno dei primi casi in Italia di integrazione programmata tra un laboratorio di linguistica e un laboratorio di conservazione, equipaggiato con dispositivi di riproduzione e registrazione allo stato dell'arte tecnologica, dove le unità di ricerca linguistica e tecnico-scientifica collaborano a stretto contatto. Questo contesto ha permesso la progettazione e la realizzazione di un ambiente informatico per la gestione dei processi relativi alla conservazione attiva[1] dei documenti sonori, in grado di supportare un insieme di procedure per il controllo di qualità e di coerenza interna dei dati.

[1] Nel campo delle memorie audio, la conservazione si articola in passiva1 (difesa del supporto dagli agenti ambientali, senza alterarne la struttura) e attiva (riposizionamento dei dati su nuovi *media*). La conservazione passiva si articola in indiretta – che non comporta il coinvolgimento fisico del disco – e diretta, nella quale il disco viene trattato, senza comunque alterarne struttura e composizione. Nella conservazione passiva indiretta rientrano: la prevenzione ambientale (che si esplica attraverso il controllo dei parametri ambientali che sono, in ordine decrescente di pericolosità per i dischi: umidità relativa, temperatura, inquinamento, luce), la formazione del personale addetto alla conservazione, l'educazione dell'utente. La conservazione passiva diretta comprende gli interventi di: realizzazione di custo-

Da tempo l'informatica si è imposta pressoché in ogni settore della società industrializzata, e gli operatori dei beni culturali, e in particolare gli archivi aperti al pubblico hanno una gestione informatizzata dei cataloghi e talvolta delle modalità di accesso dell'utenza. Se gli strumenti informatici sul mercato rispondano alle reali esigenze degli archivi e se gli archivi possiedano le competenze per formulare i propri bisogni informatici in maniera efficace, è una discussione aperta. Edwin Van Huis, basandosi sull'esperienza maturata in oltre dieci anni a capo dell'archivio audiovisivi dei Paesi Bassi, critica l'inerzia delle istituzioni archivistiche di fronte alle nuove tecnologie, sintomo che ne venga ignorato, secondo Van Huis, il potenziale nel raggiungere gli utenti ormai avvezzi all'immediatezza, a fronte dell'incerta autorevolezza, di fenomeni come Google (Van Huis, 2009).
Nel caso particolare del progetto Gra.fo, si è ritenuto necessario sviluppare strumenti informatici *ad hoc* perché: 1) nel panorama degli archivi italiani è emersa la mancanza di un'architettura in grado di gestire sistemicamente il processo di conservazione attiva e di supportare procedure per il controllo della qualità, ossia di gestire in modo automatico il controllo di processi concorrenti (acquisizione del segnale audio, estrazione dei metadati, creazione copie conservative, ecc.); 2) lo schema catalografico definito da Gra.fo per la descrizione del materiale linguistico è originale (Calamai, 2012) e pertanto non supportato da strumenti esistenti di archiviazione dei dati; e infine 3) l'implementazione del flusso di lavoro che caratterizza il processo di conservazione è vincolato dalla struttura fisica del laboratorio, ossia dalla configurazione hardware, dalle interfacce standard e dai protocolli di trasferimento adottati.

Il contributo è articolato come segue: la sezione 3 presenta il progetto Gra.fo, i suoi obiettivi, le risorse impegnate e la rilevanza nella comunità linguistica. La sezione 4 introduce alcuni problemi legati alla conservazione delle memorie sonore; la sezione 5 presenta i problemi aperti nel campo della conservazione delle memorie sonore dal punto di vista del restauro fisico dei supporti e degli strumenti software legati alla gestione dei dati nonché della loro diffusione presso il pubblico generico e specializzato. Nella sezione 6 sono descritte le soluzioni proposte ai problemi descritti nella sezione 5, ovvero le strategie adottate dal progetto Gra.fo e maturate grazie all'esperienza di precedenti progetti di ricerca nazionali e internazionali.

3. LE SOFFITTE DELLA VOCE: IL PROGETTO GRA.FO

Il progetto *Grammo-foni. Le soffitte della voce* (*Gra.fo*), condotto dalla Scuola Normale Superiore di Pisa e dall'Università degli Studi di Siena, finanziato dalla Regione Toscana (PAR FAS 2007-2013 Regione Toscana Linea di Azione 1.1.a.3.) è il primo progetto che intende restituire alla comunità, non solo dei linguisti, quanto negli anni è stato raccolto da parte di studiosi, di appassionati, di cultori delle tradizioni popolari sul territorio della regione. Esso mira pertanto a censire, raccogliere, salvaguardare, analizzare documenti vocali d'interesse linguistico, antropologico, storico, etnografico presenti in Toscana attraverso la loro ri-mediazione ad alta definizione (e, se del caso, attraverso un restauro): ciascun documento sonoro viene catalogato, descritto e – per i materiali più significativi – anche trascritto ortograficamente e foneticamente e reso infine disponibile – previa identificazione

---

die di protezione; spolveratura delle raccolte; disinfestazione degli archivi con gas inerti (Canazza, 2007).

dell'utente – presso un archivio *on-line* in un sito dedicato all'interno del dominio della Scuola Normale Superiore (*http://grafo.sns.it*).

Nella costruzione dell'archivio sonoro si sono susseguite diverse fasi, di cui rendiamo conto in altra sede (Calamai et alii, in c.d.s.; Calamai, 2012): in un primo tempo, il gruppo di ricerca si è concentrato sulla localizzazione e la scelta dei singoli archivi, definendo al contempo gli aspetti legali e operativi in merito alla cessione temporanea dei beni vocali da parte dei possessori/detentori, sia pubblici che privati; in un secondo momento sono state definite alcune procedure legate alla catalogazione e alla trascrizione dei singoli documenti sonori. Nonostante l'esistenza di un censimento dettagliato proprio per quanto concerne i beni vocali di Toscana (Andreini et alii, 2007) abbia favorito il lavoro di reperimento dei archivi, il numero degli archivi 'da salvare' appare in continua crescita.

Le tipologie di supporti sui quali sono generalmente memorizzate le indagini sul campo, tipicamente in ambito linguistico, antropologico ed etno-musicologico, differiscono da quelle sui quali sono memorizzate registrazioni di musica d'arte e di ricerca prodotte per lo più in studi di produzione e custoditi in archivi d'importanza riconosciuta, come quelli di enti nazionali o di fondazioni teatrali. I supporti impiegati per le indagini sul campo dovevano soddisfare due requisiti fondamentali, la maneggevolezza e l'economicità, in contrapposizione a quelli delle produzioni musicali dove il compromesso con l'economicità era limitato dall'esigenza di preservare la qualità della resa sonora. Di pari passo con l'evoluzione delle tecnologie per la memorizzazione e per la riproduzione acustica, queste diverse esigenze hanno portato lungo i decenni alla scelta di tipologie di supporti e in secondo luogo di formati di registrazione molto diversi. Conseguentemente, la conservazione degli archivi sonori necessità di approcci altrettanto diversi, ovvero di interventi mirati a seconda della tipologia di archivio.

Ripercorrendo cronologicamente l'evoluzione delle tecnologie per la registrazione sonora, le tipologie di supporti più comunemente trattati da Gra.fo sono audiobobine, audiocassette, Digital Audio Tape (DAT), Compact Disc (CD) e dispositivi di memoria di massa. Ciascuna tipologia di supporti presenta caratteristiche fisico-chimiche peculiari, e corrisponde ad una rosa di formati di registrazione e di lettura legata all'apparecchiatura a disposizione, alle scelte consapevoli e agli errori di chi ha effettuato la registrazione, cui si sommano una serie di varianti non convenzionali nell'utilizzo dei supporti per massimizzare il tempo di registrazione disponibile (vedi paragrafo 4.1).

4. CONSERVARE LE MEMORIE SONORE

Nel decreto legislativo n. 112 del 1998 (in attuazione della legge n. 59 del 1997), al Titolo IV (Servizi alla persona e alla comunità), capo V (Beni e attività culturali), per la prima volta viene data una precisa definizione di "tutela dei beni culturali" (art. 148 "Definizioni", comma 1, lettera c): «ogni attività diretta a riconoscere, conservare e proteggere i beni culturali e ambientali», ossia il complesso delle azioni, dirette e indirette, volte a rallentare gli effetti della degradazione causata dal tempo e dall'uso sulle componenti materiali dei beni culturali.

La conservazione dei documenti sonori si è naturalmente avvantaggiata dell'esperienza secolare maturata nell'ambito della conservazione libraria e di altre tipologie di opere d'arte, tuttavia gli operatori del settore si sono dovuti arrendere ben presto all'evidenza che i supporti su cui vengono memorizzate le informazioni sonore sono destinati ad un degrado inarrestabile causato alla loro struttura fisico-chimica. In altre parole, si è abbandonato il

concetto di "conservazione dell'originale" (Schüller, 2006). Senonché la duplicazione di un documento sonoro altera necessariamente alcune caratteristiche della copia, ad esempio per le possibili imperfezioni del sistema di riproduzione, senza contare le implicazioni di un mutato rapporto tra il bene e la sua materialità, riflessione non di non secondaria importanza e che pertiene all'indagine estetico-filosofica (Brandi, 2000). Nonostante alcune resistenze nei confronti della tecnologia digitale a scopi conservativi all'epoca del suo avvento, oggi appare universalmente condiviso che essa è quella più adatta al mantenimento dei dati nel lungo termine – idealmente "per sempre" (Edmonson, 2002), ovvero per "un lasso di tempo sufficientemente ampio da sollevare problemi di obsolescenza della tecnologia, e ciò può significare decadi o secoli"[2] (The National Science Foundation and The Library of Congress, 2003). Se da un lato la tecnologia digitale consente di produrre duplicati ennesimi di un documento sonoro senza modificarne (i.e., diminuirne) il rapporto segnale-rumore (SNR), dall'altro non è immune dai problemi legati all'invecchiamento dei supporti e dei formati. Per compensare a questi svantaggi, le linee guida proposte dall'International Federation of Library Associations and Institutions (IFLA, 2002) e dall'International Association of Sound and Visual Archives (IASA, 2005) incoraggiano l'adozione di software e di formati *open source* per l'archiviazione delle informazioni, e l'installazione di procedure di controllo, *refresh* e *migration* da applicare periodicamente ai dispositivi di memorizzazione per garantirne l'integrità nel tempo.

Nonostante il supporto fisico costituisca solo il mezzo attraverso cui il documento sonoro si esprime, esso è portatore di informazioni utili all'interpretazione del documento stesso. Pertanto è indispensabile conservare, nella forma più opportuna, anche le informazioni relative al supporto, agli eventuali allegati e alle loro imperfezioni: in altre parole è imperativo rispettare, e conservare, l'*unità documentale*. Per questo motivo la *copia d'archivio* o *copia conservativa* ("archive copy" o "preservation copy") di un documento sonoro è definita come *un insieme organizzato di dati che rappresenta tutta l'informazione portata dal documento originale nella sua complessità, unitamente alla loro descrizione e alla documentazione relativa al processo di conservazione* (Canazza et alii, 2011). La copia conservativa non è destinata alla diffusione dei contenuti per la loro fruizione, essendo finalizzata principalmente alla conservazione a lungo termine (IASA, 1999). Dalla copia conservativa è possibile ottenere una *copia d'accesso* caratterizzata da una qualità sonora pari o inferiore, accompagnata da un insieme selezionato dei dati di descrizione. Su questa tipologia di documento è possibile sperimentare interventi di restauro di entità proporzionale ai risultati desiderati, fermo restando che sull'audio della copia conservativa è interdetto qualsiasi tipo di elaborazione che renda il documento *altro* rispetto al segnale audio così come estratto dal supporto originale, ossia interventi di restauro che modificano lo spettro del segnale. Per un quadro approfondito del dibattito storico sull'etica della conservazione e del restauro, si veda Canazza et alii (2011).

4.1. Struttura di una copia conservativa

La copia conservativa ha lo scopo di minimizzare la perdita di informazione che interviene inevitabilmente durante il processo di ri-mediazione. La Figura 1 rappresenta l'organizzazione logica del contenuto di una copia conservativa: oltre al segnale audio, ci

[2] "*Long-term* may simply mean long enough to be concerned about the obsolescence of technology, or it may mean decades or centuries".

sono due livelli di metadati (nel campo degli archivi sonori, i *metadati* sono le informazioni che si possono estrarre automaticamente dal segnale) e le informazioni contestuali (nella terminologia informatica si tratta tecnicamente di metadati, ma nel campo degli archivi sonori le informazioni contestuali si distinguono perché si tratta di informazioni che *non* possono venire estratte automaticamente dal segnale, ma provengono dal documento originale, ovvero fotografie degli allegati e delle eventuali annotazioni, nonché la documentazione relativa alle corruttele del supporto e al processo di conservazione). In conformità alle principali linee guida internazionali (IASA, 2005), la copia conservativa definita dal progetto Gra.fo contiene il *checksum* del segnale audio ottenuto con tre metodi di calcolo diversi. I *checksum* sono sequenze di bit comunemente utilizzati nel campo delle telecomunicazioni che, in questo contesto, servono a verificare l'integrità dei dati a distanza di tempo o dopo una o più duplicazioni. Nell'ambito del progetto Gra.fo vengono considerati metadati di primo livello in contrapposizione alla documentazione dei formati dei documenti contenuti nella copia conservativa, considerata di secondo livello. Tale documentazione è utile per una corretta interpretazione dei flussi di bit con cui sono rappresentati i file laddove i formati fossero obsoleti e/o non fosse disponibile un programma in grado di interpretarli automaticamente in modo corretto. Infine, una scheda descrittiva descrive il contenuto della copia conservativa nel suo complesso: la scheda comprende un elenco dei documenti presenti, informazioni circa la provenienza del documento originale, la paternità e le caratteristiche della copia conservativa generata a partire da tale supporto originale, e ulteriori dati di natura tecnica che riguardano la conservazione a lungo termine.

Figura 1: Rappresentazione logica di una copia conservativa.

4.2. Formati di registrazione

Il grado di libertà nella personalizzazione dei formati di registrazione dipende dall'equipaggiamento a disposizione, dalle competenze di chi lo utilizza e dalla tipologia di supporto, ed è progressivamente maggiore se si retrocede sulla linea del tempo (le audiobobine e le audiocassette, tipologie di supporti più lontane nel tempo, consentono usi non convenzionali più di quanto non facciano i nastri digitali e i Compact Disc). Ad esempio, un magnetofono dotato di testina di registrazione stereo (*half-track*) prevede nel caso base che sulla larghezza del nastro trovino posto due tracce, una per il canale destro e una per il canale sinistro, in un solo senso di scorrimento (Figura 2, Caso 1). Se durante la fase di re-

gistrazione si esclude arbitrariamente il segnale in ingresso in uno dei canali, al termine della lunghezza del nastro è possibile invertirne il senso di scorrimento e registrare un altro segnale sulla traccia precedentemente esclusa, ottenendo così due segnali monofonici e un nastro a due sensi di scorrimento (Figura 2, Caso 2). Senza invertire il senso di scorrimento, è altresì possibile registrare un secondo segnale monofonico riavvolgendo il nastro ed escludendo il complementare del canale escluso in precedenza, ottenendo così due segnali monofonici e un nastro ad un solo senso di lettura (Figura 2, Caso 3). In funzione della velocità di scorrimento del nastro, questi espedienti permettono di aumentare fino a raddoppiare il tempo di registrazione disponibile, tuttavia complicano le operazioni di analisi del formato di lettura: per i soli casi considerati dall'esempio, che non esauriscono i casi possibili, le configurazioni che si possono ottenere con un magnetofono dotato di testina di registrazio-

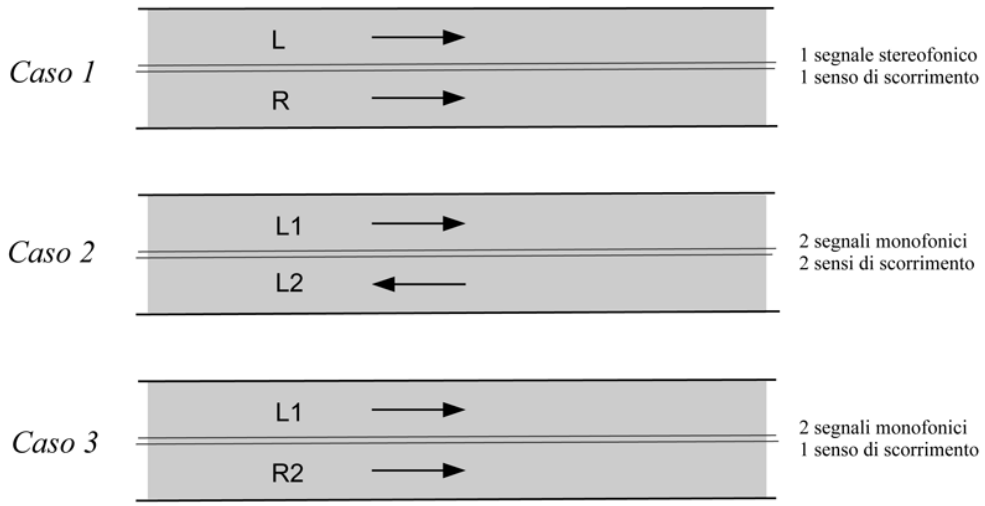


Figura 2: La figura presenta alcune configurazioni che è possibile applicare alle tracce di un nastro con un magnetofono dotato di testina di registrazione stereo. Il caso 1 è il caso base, i casi 2 e 3 sono alcune varianti possibili.

ne stereo e di due velocità di scorrimento sono una decina[3]. Sul materiale di corredo che accompagna i nastri di materiale linguistico ed etno-musicologico raramente compare un'indicazione utile circa il formato di registrazione, che va quindi individuato secondo una procedura che prevede ripetute sessioni di lettura e l'analisi comparativa dei segnali estratti, fino all'esclusione delle combinazioni non previste dagli standard e dalle varianti permesse dall'equipaggiamento usato in fase di registrazione, di cui spesso non si possiede la documentazione. Inoltre, le varianti possono aumentare e/o confondersi in formati ibridi la cui natura è solamente ipotizzabile se i dispositivi di registrazione presentavano difetti (ad esempio testina di cancellazione difettosa) e/o se sul nastro era già presente una precedente registrazione, effettuata con lo stesso equipaggiamento o un altro, applicando lo stesso for-

[3] Con riferimento alla Figura 2, il numero di velocità che si può applicare in maniera indipendente a ciascun segnale è 1 nel caso 1, e 2 nei casi 2 e 3. Se il dispositivo di registrazione supporta fino a 2 velocità, le combinazioni possibili sono $2^1+2^2+2^2=10$.

mato o un altro. Ciò costituisce un serio problema nel caso di nastri che sopportano male o non sopportano affatto ripetute sessioni di lettura.

5. TIPOLOGIE DI PROBLEMI

Il processo di conservazione attiva (schematizzato in Figura 3), finalizzato al trasferimento dell'informazione acustica nel dominio digitale, prevede tre fasi consecutive: la preparazione del supporto (ottimizzazione per la lettura e quindi restauro fisico del supporto); la lettura e l'elaborazione dell'informazione per mezzo di un dispositivo adeguatamente regolato; il trasferimento del segnale e la sua archiviazione su nuovi supporti.

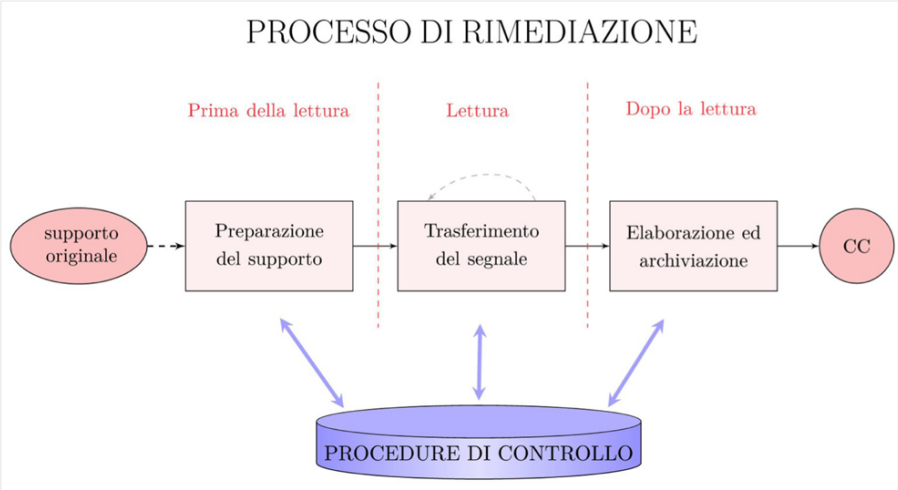


Figura 3: Schema del controllo di qualità del processo di conservazione attiva.

Lungo la catena del processo di conservazione, uno dei primi problemi da affrontare è quello del ripristino della funzionalità del supporto, ovvero dell'ottimizzazione del suo stato di conservazione prima della fase di estrazione del segnale (ossia della lettura). Questa fase comprende anche la verifica del corretto funzionamento del sistema di lettura e di registrazione, quindi la sua manutenzione e specifici test di controllo. Le criticità che si incontrano a questo punto riguardano il trattamento dei documenti sonori nella loro fisicità e richiedono competenze di chimica dei materiali e di storia delle tecniche di registrazione. Si tratta di un passo fondamentale perché un trattamento inadeguato potrebbe compromettere in maniera irreversibile l'integrità del documento e quindi le possibilità di estrazione del segnale.

Una volta portata a termine la fase di lettura del supporto originale, sia esso analogico o digitale, il segnale audio è sempre digitale. Da questo momento in poi, la manipolazione avviene esclusivamente con strumenti informatici, dalla predisposizione dei dati descrittivi da allegare alla copia conservativa alla sua archiviazione sul supporto di memorizzazione prescelto per la conservazione a lungo termine. Tutta l'informazione che viene generata e/o manipolata dalla fase di lettura in poi ha carattere puramente immateriale, ovvero si costituisce di flussi di bit che rappresentano di volta in volta testo, immagini, suoni, e pone quindi problemi che vanno risolti nella dimensione dell'informatica.

6. LE SOLUZIONI PROPOSTE

Uno degli obiettivi principali del progetto Gra.fo è stato l'allestimento di un laboratorio di restauro dotato dell'equipaggiamento tecnico necessario a svolgere tutte le operazioni previste dal processo di conservazione. L'unicità del laboratorio consiste nel fatto di essere collocato all'interno di un laboratorio di linguistica, inserendosi così nelle attività ivi condotte e permettendo la permeabilità di competenze tra il dominio linguistico e quello della conservazione, nelle sue manifestazioni estese dalla filologia all'informatica. Un tratto distintivo del laboratorio consiste nell'approccio al metodo di lavoro interdisciplinare, che trasforma il processo di raccolta dei requisiti in un dialogo quotidiano a due sensi tra le unità di ricerca informatica e linguistica. Questo dialogo è possibile solo laddove le unità lavorino a stretto contatto per diversi mesi, ragionando in modo collaborativo sugli obiettivi e superando i limiti dovuti alla concezione che discipline distanti tra loro hanno del medesimo oggetto di studio. A fronte di un'iniziale dilatazione dei tempi rispetto alla canonica raccolta dei requisiti, un lavoro di questo tipo permette di comprendere meglio la logica e i vincoli della disciplina dell'altro, che è l'unica via per giungere alla definizione di concetti e strumenti originali in senso autenticamente multidisciplinare Tale approccio è stato teorizzato solo recentemente (Agosti, 2008) e gli autori credono sia un modo di lavoro innovativo e vantaggioso in ambito informatico. In particolare, nei primi mesi di progetto il dialogo si è concentrato sulla modellizzazione degli oggetti sonori, cercando di conciliare i punti di vista linguistico, archivistico e informatico. Il risultato è un insieme di concetti e di attributi in grado di rappresentare le entità documentali e le risorse linguistiche, mantenendone la storia di trasmissione e le relazioni. In concreto, queste informazioni vengono mantenute in una base di dati, progettata per rispettare tali concetti, descritta nel paragrafo 6.1 assieme agli strumenti per il suo popolamento. Inoltre, da una riflessione sul processo di conservazione, anche alla luce dell'esperienza maturata in passati progetti di ricerca, è emersa la necessità di applicare procedure per il controllo (Figura 4) della qualità finalizzate alla validazione dei dati durante l'attività laboratoriale e durante la fase di archiviazione. Ne è risultato un ambiente informatico ad architettura modulare, descritto nel paragrafo 6.2.

Attualmente il laboratorio ospita due postazioni di riversamento parallele, attrezzate per trattare tipologie di supporti come (micro)audiocassette e audiobobine. Per il trattamento termico dei nastri magnetici è presente un incubatore di precisione in grado di variare e mantenere temperature fino ai 60 gradi centigradi (Figura 5, sinistra). Per l'estrazione dell'informazione contestuale è stata allestita una postazione fotografica stabile. Nelle immagini dei documenti vengono incluse un'unità di misura di riferimento (Galasso & Giffi, 1998) e una filigrana digitale con gli estremi del progetto e dell'archivio di appartenenza.

Le principali attività che hanno luogo nel laboratorio sono riassunte nello schema logico della Figura 6. I documenti sonori vengono consegnati dagli archivi di provenienza e temporaneamente custoditi in laboratorio. Al termine del processo di conservazione, al cui vertice in Figura 6 si colloca la fase di estrazione del segnale audio, i documenti originali vengono riconsegnati ai possessori assieme ad una copia dell'audio riversato. L'output del processo sono delle copie conservative complete che vengono archiviate nel sistema di memorizzazione a lungo termine, nel caso del progetto Gra.fo una batteria di Hard Disk Drive (HDD) in configurazione RAID-5 accessibili tramite una macchina server virtuale affidata ad una gestione Unix-based.

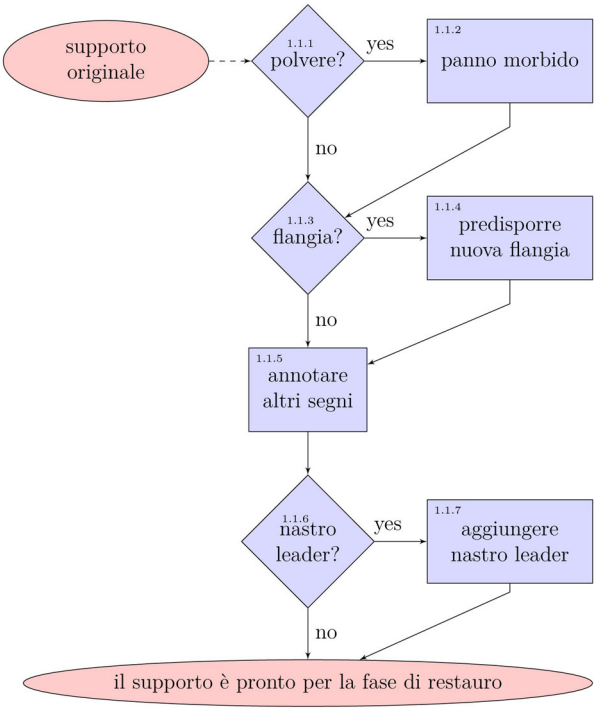


Figura 4: Diagramma di flusso che descrive la prima parte della procedura di ottimizzazione del supporto che precede la lettura per l'estrazione del segnale.

6.1. Struttura della base di dati e strumenti di popolamento

Affinché le unità di consultazione (etichetta stabilita in Gra.fo per descrivere il materiale sonoro dopo la sua riorganizzazione) costituiscano fonti documentali dotate di autorevolezza, è necessario mantenere la loro relazione con i documenti da cui derivano, in modo da poter risalire con certezza alla loro provenienza e alle elaborazioni cui sono state sottoposte. La base di dati destinata ad ospitare queste informazioni è stata progettata per mantenere i dati prodotti durante il processo di conservazione, le descrizioni previste dallo schema catalografico, e le loro relazioni. Per il popolamento della base di dati, progettata in MySQL, sono stati sviluppati due applicazioni *stand-alone* con tecnologia Java: *Audiografo PreservationPanel* e *Audiografo CataloguingPanel*.

Le prime informazioni di un documento sonoro che vengono prodotte riguardano il documento originale (durante la fase di analisi e restauro del supporto). Il programma *Audiografo PreservationPanel* (Figura 8) è stato sviluppato 1) per inserire nella base di dati a) tutti i dati prodotti durante la fase di lavorazione in laboratorio; b) i metadati noti a priori o derivabili, in modo trasparente per l'utente; nonché c) dati utili ai controlli di coerenza interna sull'archivio; e 2) per gestire l'archivio in modo automatizzato a) creando la struttura delle copie conservative a partire dai dati inseriti dall'operatore; e b) gestendo in modo automatico la collocazione dei documenti all'interno di ogni copia conservativa. A differenza

di una generica interfaccia per il popolamento della base di dati, *Audiografo Preservation-Panel* è stato sviluppato considerando l'insieme delle informazioni modellate nella base di dati e analizzando il rapporto di questi ultimi con il flusso di lavoro del processo di conservazione. In questo modo si ottengono diversi vantaggi: in primo luogo l'operatore è guidato lungo le fasi del processo di conservazione e quindi evita di discostarsi dal protocollo o di dimenticare alcune operazioni. Inoltre il programma è in grado di filtrare tutte le proprietà non associabili alla tipologia di supporto momentaneamente in lavorazione, oppure di filtrare i valori non validi per le proprietà comuni a più tipologie di supporto (e.g., la velocità di scorrimento del nastro). Il tempo risparmiato nella creazione manuale dei file e nella loro copiatura è per sé un vantaggio desiderabile dal momento che permette di incrementare la quantità di lavoro prodotta per ogni postazione di riversamento, ma l'obiettivo più importante reso possibile da un'applicazione studiata con l'approccio descritto è il controllo costante sulla coerenza interna dei dati. In particolare, la coerenza tra l'archivio e le informazioni ad essi associate nella base di dati.

Figura 5: A sinistra, incubatore per il trattamento termico dei nastri magnetici; a destra, residuo di pasta magnetica depositato sul magentofono dopo la lettura di un nastro che manifesta una perdita di pasta magnetica (*Sticky-Shed Syndrome,* SSS).

Solo una volta completato il processo di conservazione, può iniziare la fase di interpretazione e catalogazione dei contenuti. In questo momento avviene un passaggio di consegne tra operatori con competenze diverse, tecnico-informatiche da un lato e linguistiche dall'altro. A questo passaggio corrisponde un cambiamento di prospettiva nell'approccio al materiale sonoro: archivistico-filologico prima (processo di conservazione, necessario alla sopravvivenza del bene documentale) e orientato al contenuto specialistico poi (propedeutico alla fruizione da parte degli utenti dell'archivio).

Le registrazioni considerate dal progetto Gra.fo sono per lo più state raccolte sul campo e memorizzate su supporti sonori sfruttando la maggior parte dello spazio disponibile, per motivi legati ai costi dei supporti e in parte alla loro struttura. Raramente quindi un evento considerato come unità indipendente e fruibile per l'utente (intervista, canto, racconto, …), è contenuto in un unico supporto sonoro: più spesso esso è frammentato su più supporti (Figura 7). La fase di catalogazione sarà quindi preceduta da un'analisi delle registrazioni nella loro organizzazione originaria e in successive selezione e riorganizzazione del *continuum sonoro* in funzione del contenuto. Il risultato di questa riorganizzazione è un insieme

di nuovi documenti digitali che per numero e per durata sono completamente indipendenti dal numero e dalla durata dei riversamenti ottenuti dai supporti originali. Questo compito deve essere eseguito da operatori con una solida conoscenza della natura del contenuto – in questo caso di linguistica e di dialettologia. Da loro dipendono l'assetto dell'archivio e l'interpretazione delle risorse che verranno presentati gli utenti finali.

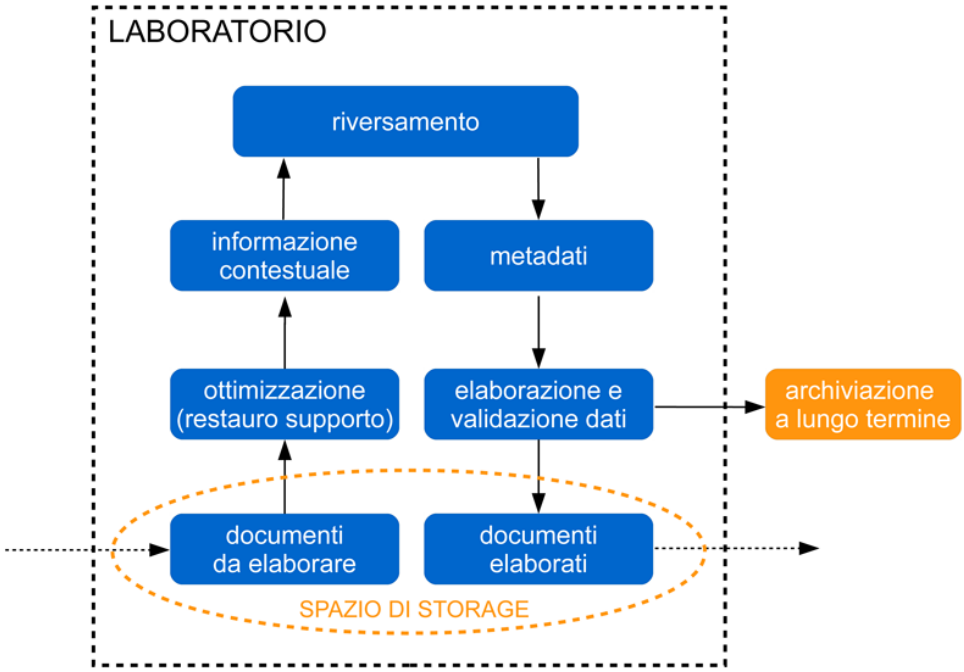


Figura 6: Schema logico delle attività svolte all'interno del laboratorio di restauro.

I documenti generati durante la fase di riorganizzazione vengono catalogati secondo uno schema definito dal gruppo di ricerca (per una descrizione dettagliata e per una discussione sulla proposta catalografica, v. Calamai (2012)), e in alcuni casi essi vengono trascritti ortograficamente e/o foneticamente. Tutte queste informazioni confluiscono nella base di dati in cui precedentemente sono stati inseriti i dati relativi ai supporti originali. La discendenza delle unità di consultazione viene mantenuta nella base di dati per mezzo di riferimenti ai documenti originali che sono stati impiegati interamente o in parte per comporre le unità. La finalità del programma *Audiografo CataloguingPanel* è quello di permettere, facilitare e controllare i dati inseriti dagli operatori nella base di dati, secondo i medesimi criteri adottati per *Audiografo PreservationPanel* ma in relazione alla scheda catalografica. Inoltre *Audiografo CataloguingPanel* è pensato per un gruppo di lavoro distribuito sul territorio nazionale, permettendo a ciascun operatore di interagire con l'archivio di copie d'accesso, la base di dati e con il lavoro dei colleghi in tempo reale anche in remoto.

6.2. Strumenti di lavoro nell'ambiente informatico

I passi necessari alla composizione di una copia conservativa consistono nella creazione e nella manipolazione di molti documenti digitali di tipo testuale, di immagine e natural-

mente audio. Per questo motivo la quantità di tempo impiegata nella gestione dei documenti digitali è molto elevata, ed è prevedibile che diverse patologie dell'attenzione possano indurre l'operatore a provocare errori dall'effetto a cascata sul processo di conservazione e il conseguente malfunzionamento degli algoritmi di controllo sulla coerenza interna dell'archivio e/o di reperimento dell'informazione.

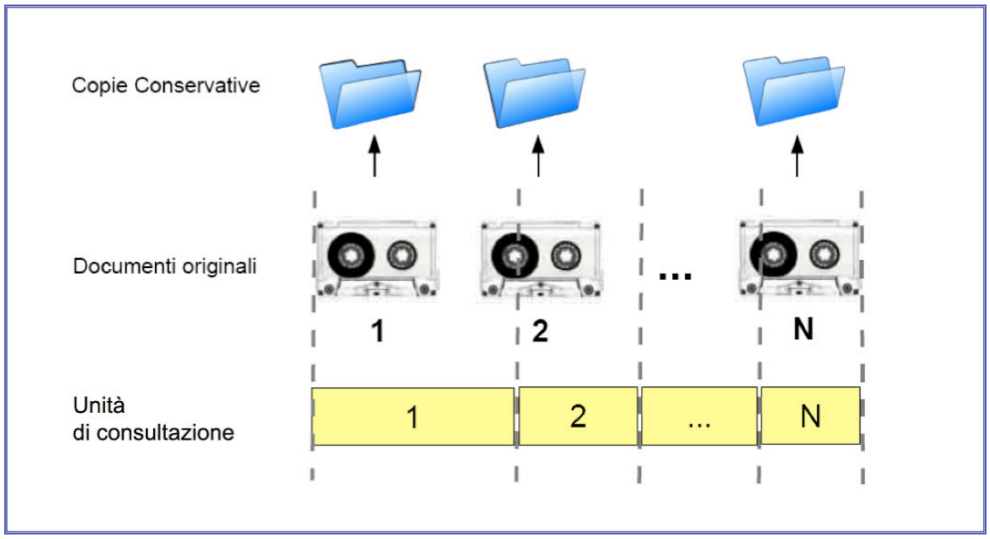


Figura 7: Rappresentazione del rapporto che sussiste tra i documenti originali, le copie conservative (sempre 1:1) e le unità di consultazione (complesso).

La verifica manuale di tutti i documenti che compongono l'archivio è chiaramente impraticabile, e una verifica a campione è altrettanto insoddisfacente. Durante il progetto Gra.fo sono stati sviluppati degli strumenti software *ad hoc* per la gestione del flusso di lavoro all'interno del laboratorio, riducendo il tempo necessario al completamento di una sessione di riversamento, abbreviando sensibilmente il tempo necessario a produrre copie d'accesso a partire dalle copie conservative, e introducendo una serie di controlli automatizzati che garantiscono in modo pienamente affidabile l'integrità dei dati e dell'archivio. Tali strumenti si dividono, in funzione della finalità d'uso, in strumenti di lavoro e di controllo:

1. Strumenti di lavoro
    a. Allineamento degli archivi (laboratorio, server, backup)
    b. Creazione e condivisione di copie d'accesso
    c. Base di dati
    d. Programmi per il popolamento della base di dati

2. Strumenti di controllo
    a. Monitoraggio dei processi
    b. Validazione dei dati (medio/lungo termine)
    c. Procedure di backup (base di dati, sito web, …)
    d. Monitoraggio dell'incremento dei dati

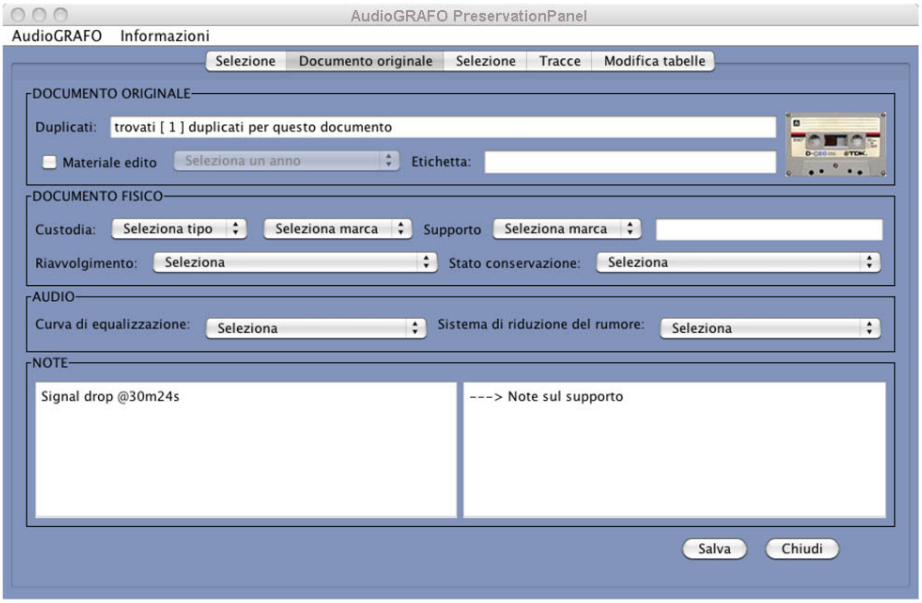


Figura 8: Interfaccia dedicata alle descrizione del documento originale nel programma *Audiografo PreservationPanel.*

L'archivio di copie conservative (generato con l'ausilio di *Audiografo PreservationPanel*) viene trasferito quotidianamente dal laboratorio ad una macchina server, che ospita il sistema di archiviazione a lungo termine e dove ha luogo il resto delle routine di elaborazione. Per ciascuna copia conservativa vengono create una copia di sicurezza e una copia d'accesso, accompagnata esclusivamente dal materiale fotografico, e condivisa via web con accesso regolamentato. Al termine della conversione e quindi della creazione delle copie d'accesso, il gruppo di lavoro riceve una mail di notifica automatica che informa circa la presenza di nuovi documenti da analizzare, comunica l'indirizzo a cui è possibile ottenerli e fornisce l'elenco completo dei documenti, in modo tale che i membri del gruppo possano decidere a seconda dell'incarico assegnato se i documenti sono di loro interesse. Il percorso dal laboratorio al web si ripete quotidianamente, minimizzando così il tempo necessario a completare il passaggio dei materiali audio dall'archivio di provenienza ai curatori dell'archivio.
Al termine della creazione dell'archivio per la conservazione a lungo termine, il processo di conservazione attiva è completato. Ora, per la sopravvivenza dell'archivio è necessario programmare, e sono previsti a breve, i processi di *checking*, *copying* e *migration*, poiché la tecnologia digitale è anch'essa soggetta all'azione degradante del tempo e all'obsolescenza dei formati e dei supporti.
Infine, è stata sviluppata una serie di programmi di controllo per monitorare la corretta esecuzione del flusso di lavoro e dei dispositivi necessari all'esecuzione, in particolare la macchina server (base di dati, web server, servizio di posta per l'invio delle notifiche automatiche, …). La maggior parte di queste funzionalità è programmata per l'esecuzione automatica a intervalli predefiniti e genera una reportistica testuale salvata in loco oppure inviata via posta elettronica. A cadenza giornaliera viene calcolata la durata temporale (in hh:mm:ss) del segnale audio digitalizzato, dettagliato per ogni archivio: dopo sei mesi

dall'allestimento del laboratorio, il monte ore complessivo supera le 400 ore. Inoltre, ogni mese è programmato il monitoraggio di alcune voci incrementali della base di dati, la cui reportistica è inviata automaticamente agli operatori incaricati della sua validazione. Per garantire la sicurezza dei dati ospitati sul server (principalmente copie d'archivio, copie d'accesso, base di dati e sito web), è stato programmato un sistema che crea una copia di backup dell'insieme totale dei dati a rotazione giornaliera o settimanale, mantenendo inoltre informazioni di tipo storico sull'avanzamento del progetto.

7. CONCLUSIONI

Gli archivi che custodiscono documenti sonori hanno dovuto affrontare, negli ultimi decenni, nuovi problemi posti dalla natura fisico-chimica dei supporti impiegati per la memorizzazione delle informazioni di tipo acustico, legati principalmente alla scarsa aspettativa di vita dei supporti stessi e alla rapida obsolescenza dei formati e dei dispositivi di riproduzione. Discipline scientifiche tradizionalmente lontane dalla biblioteconomia e dall'archivistica, come la chimica dei materiali e l'informatica, offrono soluzioni efficaci a tali problemi. In questo articolo sono stati descritti la metodologia adottata e gli strumenti sviluppati dal progetto di ricerca Gra.fo per la creazione e la gestione di un archivio digitale di *corpora* vocali. L'elemento innovativo consiste nell'utilizzo di un approccio sistemico alle attività di conservazione dei documenti sonori, in grado di combinare una serie di procedure semi-automatizzate per la generazione, la descrizione dell'archivio con strumenti per il controllo di qualità, e all'impiego di programmi informatici sviluppati per la gestione e la distribuzione delle informazioni. I risultati raggiunti in precedenti progetti di ricerca e quelli raggiunti allo scadere del primo anno del progetto Gra.fo dimostrano che il supporto di strumenti informatici sviluppati in stretta collaborazione tra esperti di aree disciplinari diverse offre soluzioni efficaci ai problemi della conservazione delle memorie sonore e allo stesso tempo implica una trasformazione nell'approccio all'universo degli archivi sonori, nel quale ormai la tecnologia si è inserita in maniera irreversibile così come in altri settori del patrimonio culturale.

BIBLIOGRAFIA

Agosti, M. (2008), Information Access using the Guide of User Requirements, in M. Agosti (a cura di.), Information Access thought Search Engines and Digital Libraries, Heidelberg: Springer-Verlag Berlin Heidelberg, 1-12.

Andreini, A. & Clemente, P. (a cura di) (2007), I custodi delle voci. Archivi orali in Toscana: primo censimento, Firenze, Regione Toscana.

Brandi, C. (2000), Teoria del restauro, Einaudi.

Calamai, S. (2012), Ordinare archivi sonori: il progetto Gra.fo, Rivista Italiana di Dialettologia, 35.

Calamai, S & Bertinetto P. M. (in c.d.s), Per il recupero della Carta dei Dialetti Italiani, XLV Congresso Internazionale della Società di Linguistica Italiana, settembre 2011, Aosta-Bard-Torino.

Canazza, S. (2007), Note sulla conservazione attiva dei documenti sonori su disco, in 'Il suono riprodotto. Storia, tecnica e cultura di una rivoluzione del Novecento', a cura di A. Rigolli and P. Russo, pagg. 87–111, ed. EDT.

Canazza, S. & De Poli, G. & Vidolin A. (2011), La conservazione dei documenti audio: un'innovazione in prospettiva storica. Archivi, VI(2):7-56, luglio-dicembre 2011.

Edmonson, R. (2002), Memory of the World: general guidelines to safeguard documentary heritage, UNESCO.

Galasso, R. & Giffi, E. (1998), La documentazione fotografica delle schede di catalogo - metodologie e tecniche di ripresa. Tech. rep., Istituto Centrale per il Catalogo e la Documentazione (ICCD) - Ministero per i Beni e le Attività Culturali.

IASA (1999), The IASA Cataloguing Rules, IASA Editorial Group.

IASA-TC 03 (2005), The Safeguarding of the Audio Heritage: Ethics, principles and preservation strategy. Tech. rep.

IFLA (2002), Audiovisual and Multimedia Section: Guidelines for digitization projects: for collections and holdings in the public domain, particularly those held by libraries and archives. Tech. rep.

International Federation of Library Associations and Institutions (IFLA) (March 2002)

Schüller, D. (2006), The Ethics of Preserving Audio and Video Documents, in 'Information for All' Program (IFAP - report 2004-2005), UNESCO, Paris, 78-80.

The National Science Foundation and The Library of Congress (2003), It's about time - research challenges in digital archiving and long-term preservation. Tech. rep.

Van Huis, E. (2009), What makes a good archive? In IASA Journal, 34, december 2009.

"COSA POSSO FARE PER LEI?". UN SISTEMA PER L'ACQUISIZIONE DI CORPORA DI PARLATO ATTRAVERSO UN'APPLICAZIONE WEB

Franco Cutugno[1], Maria Palmerini[2], Gianluca Mignini[2], Ruben Cerolini[2]

[1] LUSI-Lab, Dipartimento di Scienze Fisiche, Università Federico II di Napoli
[2] Cedat85 Roma
cutugno@unina.it, *{m.palmerini, g.mignini,r.cerolini}@cedat85.com*

1. INTRODUZIONE

La raccolta di un corpus di parlato, sia per scopi di ricerca linguistica sia per scopi applicativi e tecnologici, è un compito che si rende necessario sempre più spesso. Specialmente nei casi in cui occorra tenere sotto controllo un certo numero di parametri sociolinguistici (a partire dalla variazione diatopica), può essere difficile che un singolo ricercatore sia in grado di reperire, fra la cerchia dei suoi conoscenti, un numero sufficiente di informatori che rispondano alle caratteristiche richieste. Le moderne tecnologie basate su servizi web e su applicazioni ad intelligenza distribuita permettono la realizzazione di piattaforme di acquisizione e gestione di un grande numero di registrazioni audio, sfruttando in maniera massiva gli aspetti cosiddetti sociali del web 2.0

Possono dunque facilmente essere realizzate delle applicazioni che consentano di progettare la raccolta di un corpus di parlato gestita attraverso un servizio ospitato su un server, chiedendo poi agli informatori di collegarsi come client al servizio attraverso un qualsiasi browser, aprire una o più sessioni di registrazione audio, registrare localmente il materiale audio richiesto e trasferirlo poi sul server.

Iniziative di questo tipo già presenti in rete sono ad esempio:

1) Speak everywhere (http://speak-everywhere.com/): è un portale a pagamento che offre un grande numero di servizi collegati all'apprendimento orale delle lingue straniere. Sviluppato presso il Centro per "Technology-Enhanced Language Learning" presso la Purdue University (Indiana – USA), è pensato prevalentemente per migliorare la competenza delle lingue attraverso una piattaforma di e-learning che si avvale anche di tecnologie per la registrazione e la riproduzione di segnali vocali. Il servizio offre come valore aggiunto la possibilità di raccogliere corpora di parlato.

2) WikiSpeech (http://webapp.phonetik.uni-muenchen.de/wikispeech): è un Content Management System (CMS) specificamente concepito per la raccolta di corpora di parlato online. Il servizio supporta pienamente le fasi sia di raccolta, che di annotazione del materiale audio. La preparazione di tutte le fasi del progetto avviene utilizzando un servizio configurabile dall'utente attraverso la creazione statica di un progetto sul server stesso. Allo stato attuale il WikiSpeech supporta solo le lingue inglese, tedesco, rumeno e russo.

Il presente contributo mostra un progetto realizzato nell'ambito di una collaborazione fra il LUSI-Lab dell'Università Federico II di Napoli e la società Cedat 85, nel corso del quale è stata realizzata un'applicazione per consentire la raccolta di corpora di parlato on line. Nel seguito di questo articolo illustreremo l'architettura software del sistema proposto (sia rispetto all'organizzazione del back-end che di quella del front-end) e ne illustreremo le principali funzionalità.

2. ARCHITETTURA DEL SISTEMA.

Il sistema proposto è un servizio web attraverso il quale è possibile definire alcuni criteri progettuali di un corpus di parlato e predisporre una specifica modalità di registrazione basata sulla somministrazione di istruzioni (prompt) a un insieme di volontari. I prompt possono essere semplicemente letti o possono essere usati per ispirare i parlanti a formulare liberamente degli enunciati che rispettino l'argomento suggerito, eventualmente chiedendo di impiegare termini specifici. I volontari possono effettuare le registrazioni vocali via web, a patto di possedere una connessione di rete di velocità affidabile e un microfono di buona qualità. Il servizio è stato testato e funziona con tutti i principali browser.

In termini generali, l'architettura del sistema può essere riassunta dallo schema in Fig. 1:

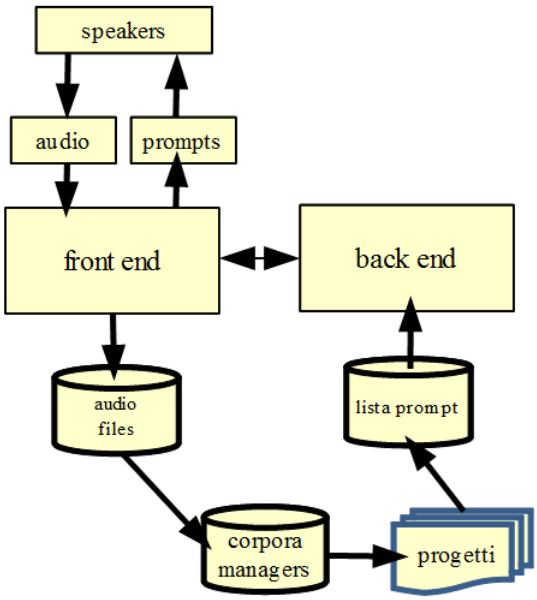


Figura 1. Architettura del sistema

2.1. Descrizione del front-end

In questa applicazione è possibile simulare l'ambiente all'interno del quale vogliamo raccogliere il corpus di parlato. In particolare, l'applicazione ricrea una situazione di dialogo uomo-macchina (almeno nella sua fase iniziale) in cui all'utente umano viene proposta una serie di prompt; questi non sono altro che brevi descrizioni di scenari nei quali l'utente dovrà immaginarsi inserito per poi esprimere la sua richiesta al sistema.

In questo modo, è possibile raccogliere un corpus di frasi che sono espresse in modo spontaneo, ma al tempo stesso sono elicitate in un contesto guidato, così da garantire la pertinenza della frase pronunciata e la copertura di un'ampia casistica. Al contempo è possibile richiedere esplicitamente agli speaker di leggere direttamente i prompt.

Accedendo alla pagina iniziale l'utente trova un'area di login, con la possibilità di richiedere username e password, per gli utenti non registrati. La pagine è mostrata nella figura 2.

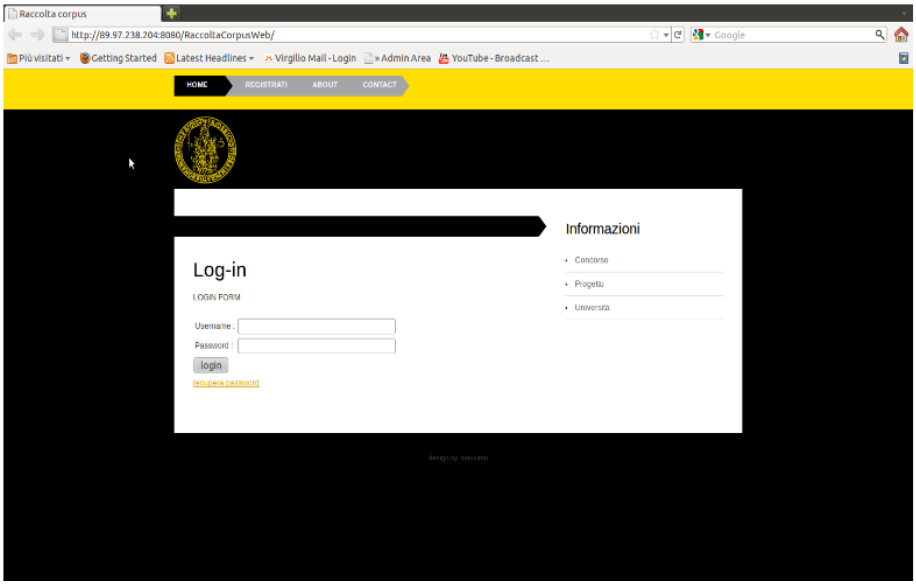


Figura 2. Schermata di login

Dopo aver inserito le credenziali ogni utente accede alla sua area personale dove può iniziare a effettuare le registrazioni. Come mostra la figura 3, all'utente viene proposto un prompt che descrive una situazione-tipo: questo sarà il contesto immaginario all'interno del quale dovrà essere formulata la prima richiesta.

Normalmente, all'utente viene chiesto di leggere con calma il prompt e pianificare la prima registrazione che potrà avere inizio premendo il tasto 'start' sulla sinistra. Dopo aver effettuato la prima registrazione, l'utente può:

- ascoltare quello che ha appena registrato (per verificare che la registrazione sia andata a buon fine, che il microfono fosse acceso, che il volume fosse adeguato etc.) ed eventualmente ripetere la registrazione;
- inviare la registrazione al sistema e procedere oltre.

Si accede così a un secondo prompt, poi a un terzo e così via, fino a raggiungere il numero di prompt indicati in fase di progettazione dell'esperimento.

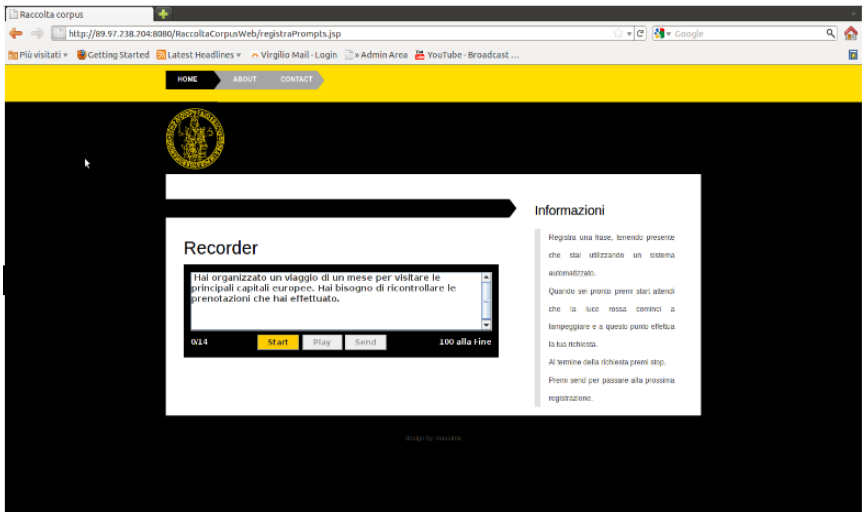


Figura 3. Schermata di descrizione di una tipica sessione di registrazione

2.2 Il back-end

Gli utenti del servizio possono appartenere a due categorie, i corpora manager (o designer) e gli speaker. Per ogni corpus manager sono definiti uno o più progetti; per ogni progetto il manager definisce una serie di prompt. Il back-end registra le scelte del manager, consente la creazione dell'ambiente di lavoro per il progetto selezionato e, successivamente, l'upload sul server della lista dei prompt, come mostrato in figura 2. In conclusione di questa fase, il back-end fornisce al manager l'URL da comunicare agli speaker per l'avvio della fase delle registazioni audio.

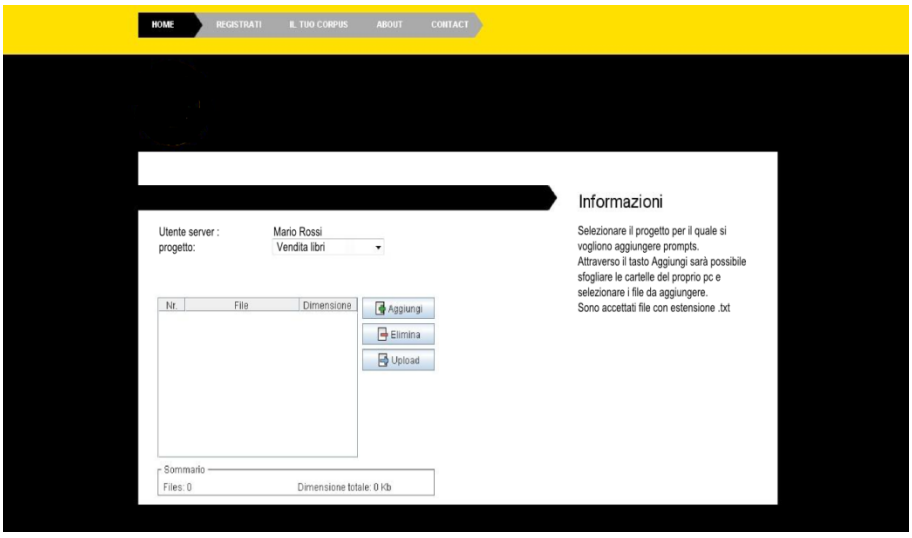


Figura 4. Una schermata del back-end dell'applicazione

Esternamente all'applicazione poi, il manager individua e contatta gli speaker, e gli comunica l'URL del progetto. Un utile suggerimento è quello di utilizzare i social network – come ad esempio Facebook – per selezionare in poco tempo un buon numero di parlanti.

3. L'ATTIVITÀ DI PERSONALIZZAZIONE

Una delle caratteristiche dell'applicazione è di essere facilmente personalizzabile a seconda del dominio sul quale si intende lavorare.

Nella nostra attività di test, abbiamo simulato uno scenario di call center automatico di una compagnia aerea che abbiamo chiamato Volare Web. Abbiamo coinvolto circa 15 persone che hanno partecipato alla nostra raccolta fingendosi utenti del call center.

Lo scopo principale di questa attività è stato raccogliere un corpus di richieste possibili, elicitate per mezzo di una serie di prompt appositamente preparati.

All'inizio della registrazione (attivata premendo il pulsante 'start'), una voce sintetica apriva l'interazione con un'offerta generica: "*Volare Web. Cosa posso fare per lei?*". Questo input aveva lo scopo di facilitare l'elicitazione della richiesta da parte dell'utente.

Le situazioni descritte nei prompt appartenevano sostanzialmente a quattro macro tipologie:

- richiesta di informazione/prenotazione
- cambio di prenotazione
- cancellazione
- consultazione dei dati personali

Poiché abbiamo ritenuto un elemento decisivo riuscire ad avere frasi formulate nel modo più spontaneo possibile, nei prompt che abbiamo preparato abbiamo cercato di descrivere situazioni realistiche e dettagliate ma senza dare indicazioni che potessero suggerire il tipo di richiesta che ci si aspettava dall'utente. In altre parole, abbiamo cercato di descrivere la situazione di base, lasciando che fosse poi l'utente a decidere quale richiesta fare.

Ad esempio, per elicitare una richiesta di cancellazione di un volo, non abbiamo dato un input del tipo "*devi cancellare il volo che hai acquistato*", ma piuttosto:

"*avevi prenotato un volo per andare a un convegno ma questo è stato annullato*"; oppure, con una formulazione più estesa e dettagliata, ma ugualmente non esplicitamente suggestiva:

"il corso professionale al quale ti eri iscritto è saltato perché non ha raggiunto il numero sufficiente di partecipanti; al momento è rinviato a data da stabilirsi, quindi il volo per Milano che avevi prenotato non è più necessario".

In effetti, il corpus di richieste che abbiamo raccolto si è rivelato molto vasto e diversificato. Abbiamo potuto constatare che a partire da uno stesso prompt come quello citato sopra, i nostri utenti hanno formulato richieste estremamente diverse una dall'altra.

Alcuni hanno formulato richieste estremamente brevi ma che contenevano le informazioni indispensabili per essere intercettate da un sistema automatico, come "*vorrei cancellare la mia prenotazione per un volo da Roma a Milano*", o solo "*devo annullare un volo da Brindisi per Milano*". In altri casi, invece, siamo stati sorpresi di scoprire che le richieste risultavano molto estese e anche arricchite di particolari che non erano stati forniti o richie-

sti nel prompt, come: "*buongiorno signora senta dovrei annullare un volo che ho prenotato da Milano Brindisi ieri sera il codice prenotazione è PNR sette quattro otto nove G grazie*".

Inoltre, come quest'ultimo esempio mostra, sebbene fosse stato loro spiegato che avrebbero dovuto simulare l'interazione con un call center automatico, alcuni utenti hanno aperto la richiesta con formule di saluto come "*Sì, buonasera signora*", oppure "*mi perdoni*", o anche "*cortesemente, potrebbe...*", oltre a chiudere la registrazione con "*grazie*", formulazioni adeguate solo in un contesto comunicativo fra esseri umani.

Questi elementi sono senz'altro di disturbo ai fini dell'estrazione automatica delle informazioni necessarie per indirizzare le richieste degli utenti; tuttavia abbiamo ritenuto il corpus raccolto estremamente utile proprio perché ben rappresentativo di un'ampia casistica di richieste che sono effettivamente plausibili, se si pensa a una possibile applicazione utilizzata nel mondo reale da utenza di varia età, cultura e non sempre a conoscenza delle modalità di funzionamento di sistemi automatici di riconoscimento del parlato.

L'applicazione si è, dunque, rivelata un utile strumento per i seguenti motivi:

- possibilità di coinvolgere un alto numero di partecipanti
- possibilità di diversificazione del campione di partecipanti
- facilità di personalizzazione da parte del ricercatore
- facilità di uso da parte dell'utente
- possibilità di raccogliere un gran numero di dati in tempi estremamente ristretti (si veda la tabella 1)
- possibilità di raccolta di un corpus estremamente realistico

Nella tabella 1 sono riassunti alcuni dati quantitativi:

Tabella 1

Numero totale di partecipanti	19
Partecipanti di sesso maschile	4
Partecipanti di sesso femminile	15
Età dei partecipanti	28-55
Durata dell'attività di raccolta	5 giorni
Numero di prompt proposti	100 (divisi in 4 macro tipologie
Numero di registrazioni raccolte	1516
Durata singola registrazione	Da 4" a 1'
Caratteristiche acustiche	file wav mono a 8 KHz

CONCLUSIONI

L'applicazione, con la possibilità di personalizzazione dello scenario e dei prompt, sarà a disposizione di chi ne faccia richiesta, in forma gratuita, se utilizzata a fini di ricerca; sarà invece previsto un costo per un impiego a fini commerciali da parte di soggetti privati[i]

Infine, saranno proposte prossime attività di miglioramento dell'applicazione e suggerite linee di ricerca ulteriori.

L'attività di raccolta del corpus si è rivelata di grande importanza per avere un'idea della varietà di richieste possibili nel contesto di interesse. Nel nostro caso, il corpus è stato impiegato per diverse attività fondamentali per la costruzione di una nuova applicazione:

- addestramento dei modelli (acustico e di linguaggio) del sistema di riconoscimento vocale da integrare in una applicazione su dispositivi mobili (Android) in corso di realizzazione;
- individuazione delle parole chiave e della loro distribuzione all'interno della frase, in una vasta gamma di casi;
- gestione (individuazione e annotazione) delle parti 'di scarto' all'interno della richiesta (saluti, esitazioni, ringraziamenti e tutte le informazioni non utili ai fini del riconoscimento della richiesta, ma comunque presenti negli enunciati ed interessanti per studi collegati alle disfluenze).

Poiché riteniamo che la difficoltà di raccolta dei corpora di parlato sia uno dei problemi cruciali nell'ambito della ricerca e sviluppo di applicazioni nel campo dello speech processing, intendiamo dare un nostro contributo per agevolare questa importante attività.

Il progetto che qui presentiamo è in continua evoluzione, fra gli sviluppi futuri che intendiamo perseguire possiamo rapidamente elencare i seguenti possibili spunti:

1) fornire una procedura di trascrizione e forced alignement per foni e parole a tutto il materiale registrato, alle stesse condizioni offerte per la fruizione dell'intero sistema (gratuito o con un contributo simbolico per gli enti di ricerca e a pagamento per scopi commerciali);
2) aggiungere una funzione per registrare dialoghi ricorrendo ad una piattaforma VOIP integrata nella nostra applicazione;
3) fornire una procedura in cloud di trascrizione e allineamento dei segnali vocali con il relativo testo per coloro che preferiscono trascriversi i dati personalmente.

NB Questo lavoro descrive uno strumento di lavoro che si inserisce in un oltremodo consolidato settore di ricerca, quello della raccolta di corpora di parlato. Gli autori hanno deciso di non completare il testo con una sezione bibliografica, rinviando il lettore a studi di uso più generale. L'intento di questa comunicazione rimane quello di presentare lo strumento non i suoi risvolti scientifici.

[i] Per informazioni sulla disponibilità contattare MP.

SUL MIGLIORAMENTO DELL'INTELLIGIBILITÀ SOGGETTIVA E OGGETTIVA OTTENUTO CON TECNICHE DI SPEECH ENHANCEMENT

Giovanni Costantini[1,2], Andrea Paoloni[3], Massimiliano Todisco[1,3]
[1]Università di Roma "Tor Vergata", Dip. di Ingegneria Elettronica, Roma, Italia
[2]Istituto di Acustica e Sensoristica "O. M. Corbino", Roma, Italia
[3]Fondazione "Ugo Bordoni", Roma, Italia
costantini@uniroma2.it, pao@fub.it, massimiliano.todisco@uniroma2.it

1. INTRODUZIONE

Riteniamo opportuno ricordare che con il termine "intelligibilità" si intende la capacità da parte di un ascoltatore di comprendere correttamente frasi e parole pronunciate da un parlante. L'intelligibilità non va confusa con la qualità, in quanto con quest'ultimo termine si vuole valutare la gradevolezza e la naturalezza della voce. La norma ISO 9921 (ISO 9921-1, 1996) definisce l'intelligibilità come "la misura dell'efficacia nel comprendere il parlato". Diremo ad esempio che siamo in presenza di un'intelligibilità del 90% se siamo in grado di trascrivere correttamente nove parole su dieci.

La misura dell'intelligibilità ha particolare rilievo per verificare l'idoneità di un sistema di trasmissione o di memorizzazione della voce. Quando la voce del parlante giunge all'ascoltatore tramite un sistema elettroacustico, quale un telefono oppure un impianto di amplificazione, le caratteristiche del sistema influenzano l'intelligibilità del segnale. In particolare, l'intelligibilità è influenzata dalla presenza di disturbi di tipo additivo o moltiplicativo (rumori ambientali o riverberazioni) e dalla riduzione della banda audio operata dal sistema. Quando il segnale audio reso disponibile è affetto da disturbi di varia natura che ne compromettono qualità e intelligibilità, si ricorre alle tecniche impropriamente chiamate di filtraggio o di ripulitura del segnale.

Le tecniche di elaborazione del segnale applicate al restauro audio si basano, per lo più, sulla sottrazione spettrale per attenuare il rumore di fondo a larga banda e sull'interpolazione nel dominio del tempo per sostituire il segnale corrotto da disturbi impulsivi. Queste tecniche, su cui torneremo, sono utilizzate con successo per migliorare la qualità delle registrazioni storiche contenute in supporti di diverso tipo e sono spesso utilizzate su di un audio di buona qualità affetto tuttavia da rumori come sibili, ronzii, fruscii, rumore di fondo, click (Brookes, 2008), (Loizou, 2007).

Un'altra applicazione, più recente, del restauro audio è quella volta a migliorare il rapporto segnale rumore per rendere il parlato più comprensibile ai sistemi di riconoscimento. Esempi comuni sono il miglioramento del segnale vocale per i sistemi di comando vocale del telefono e di alcuni dispositivi dell'auto (Alexander, 2011).

Nelle applicazioni forensi è frequente che, a seguito di un'intercettazione, specialmente se ambientale (Cerrato, 1996), (Cerrato, 1999) si registrino segnali fortemente degradati da rumori quali il motore di un'auto o di altre voci che parlano nello stesso ambiente. L'obiettivo del restauro, in questo caso, non è quello di migliorare la qualità audio del segnale, ma di migliorarne l'intelligibilità, ovvero la possibilità di comprendere e trascrivere correttamente quanto viene detto.

In questa particolare applicazione. La qualità dei segnali, in termini di larghezza di banda e rapporto segnale/rumore, è particolarmente scadente, tanto da provocare il fallimento dei sistemi di restauro.

In generale, l'esperto non si confronta solo con i rumori (additivi), ma con disturbi di varia natura, come la riverberazione o la saturazione. Ciò significa che l'esperto forense non ha a che fare unicamente con il miglioramento SNR.

È esperienza comune che, nella maggior parte dei casi, con le tecniche di restauro o di speech enhancement non si ottiene alcun reale incremento dell'intelligibilità del segnale ma piuttosto un peggioramento, più o meno lieve.

Obiettivo di questo lavoro è verificare questo assunto, ovvero se le tecniche di speech enhancement, adottate dai laboratori forensi, siano o meno in grado di migliorare l'intelligibilità di un segnale molto degradato come quello spesso proveniente dalle intercettazioni ambientali.

2. IL FILTRAGGIO DEL RUMORE

Per filtraggio del rumore o denoising si intende quel ramo dell'elaborazione numerica di segnali che studia i sistemi di cancellazione e di attenuazione del rumore. Il rumore è presente a diversi livelli in ogni ambiente ed è quindi un elemento di disturbo in ogni tipologia di comunicazione. I metodi di denoising sono stati elaborati per ridurre questo problema. In particolare essi devono conformarsi in base alla tipologia di rumore e alle applicazioni a cui si riferiscono.

Un segnale vocale può essere reso inintelligibile dalla presenza di rumore o a causa di fenomeni di varia natura. I rumori possono essere additivi, quando si sommano al segnale utile mascherandolo, o moltiplicativi, quando sono correlati al rumore stesso; il rumore additivo non varia in funzione del segnale, ad esempio è presente anche nelle pause del segnale, mentre quello moltiplicativo è presente solo in presenza del segnale.

I sistemi di denoising si possono distinguere in due categorie principali a seconda del campo di applicazione, ovvero il settore delle telecomunicazioni e quello del restauro di documenti audio.

Nel settore delle telecomunicazioni, il denoising viene utilizzato per rimuovere o attenuare la presenza del rumore d'ambiente in comunicazioni radio o telefoniche, ad esempio quando si utilizza un telefono cellulare in auto o in una sala affollata. In questo caso, lo scopo principale del sistema di denoising è quello di rendere comprensibile la comunicazione. Sono quindi accettate delle distorsioni, purché esse non condizionino l'intelligibilità del segnale.

Il restauro di registrazioni sonore su diversi supporti, come vinili e nastri magnetici, è un altro importante campo di applicazione del denoising. In questo caso non è permessa la presenza di distorsioni che, anche a livelli minimi, compromettano la percezione di qualità nel segnale. Si noti che, in genere, il valore di SNR in questo tipo di applicazione è sensibilmente maggiore rispetto al caso delle telecomunicazioni.

Questa distinzione sulla base del campo di applicazione non comporta necessariamente l'elaborazione di metodi diversi, considerato che, sia nelle telecomunicazioni che nelle registrazioni audio, si verificano le medesime condizioni di disturbo. Per tali tipi di circostanze condivise, è quindi auspicabile lo sviluppo di soluzioni valide in entrambi i casi.

SUL MIGLIORAMENTO DELL'INTELLIGIBILITÀ SOGGETTIVA E OGGETTIVA OTTENUTO CON TECNICHE DI SPEECH ENHANCEMENT

Affinché un sistema di denoising sia efficace, è necessario classificare e modellare le varie tipologie di rumore, nonché i processi che sono all'origine della loro manifestazione. In questo modo si possono usare queste informazioni per meglio identificare le componenti indesiderate e facilitarne l'eliminazione.

Esistono diversi modi per classificare anche i sistemi di denoising; si distingue tra sistemi a singolo canale e sistemi a canali multipli (Vaseghi, 1996). Questi ultimi fanno uso di informazioni ricavate dalle posizioni spaziali delle sorgenti. Per esempio, il metodo "beam-forming" determina la direzione da cui proviene un suono per mezzo di un array di microfoni.

Nella tecnica di "cancellazione adattativa del rumore" (ANC) viene analizzata una coppia di canali: nel primo canale è presente il segnale sommato al rumore, mentre nel secondo solo il rumore, e viene imposta l'ipotesi che le due sorgenti di rumore siano correlate tra loro.

Questi metodi non possono essere utilizzati in applicazioni forensi, poiché il segnale proveniente da un'intercettazione è sempre acquisito da registrazioni ad un solo microfono.

Uno dei metodi di denoising a singolo canale è quello a sottrazione spettrale. Esso si basa sulla STFT (Short-Time Fourier Trasform), cioè su una stima dello spettro d'ampiezza di un segnale osservato ad instanti consecutivi e ravvicinati nel tempo. Uno dei vantaggi principali della sottrazione spettrale è che essa può essere applicata anche quando si ha a disposizione soltanto il segnale disturbato. Essa non richiede, infatti, diversamente da altri metodi, l'accesso diretto alla fonte del disturbo.

Le tecniche di sottrazione spettrale sono state introdotte da Boll (Boll, 1979) che intendeva utilizzarle all'interno di un algoritmo per la compressione di segnali vocali nei sistemi di comunicazione. L'algoritmo era destinato ad attenuare il rumore che fosse sia a banda stretta periodico sia a banda larga. In particolare, i primi algoritmi di sottrazione spettrale sono stati sviluppati in ambito militare, con l'intento di migliorare l'intelligibilità del parlato in condizioni di trasmissione fortemente disturbata, ad esempio in situazioni in cui il livello di SNR oscilla tra i ±5 dB. Si è constatato, però, che a valori così bassi di SNR, la sottrazione spettrale non produce risultati soddisfacenti. In alcuni casi essa addirittura peggiora l'intelligibilità del segnale (Lim, 1978), (Lim, 1979), (Hilkhuysen, 2010). Tuttavia questa tecnica permette di migliorare la qualità percepita del segnale. Essa può quindi essere utilizzata nel caso in cui i livelli di SNR siano alti (in genere +15 dB), ovvero in una condizione di poco rumore, per migliorare la qualità del segnale.

Gli algoritmi di sottrazione spettrale, come si è detto, si fondano sull'analisi dello spettro di ampiezza del segnale, mentre ignorano quello di fase. Essi si differenziano proprio in base alle modalità di stima dello spettro d'ampiezza.

La tecnica di filtraggio adattivo a singolo canale è utilizzata per identificare le componenti di un segnale che sono correlate con campioni precedenti del segnale stesso; tali correlazioni derivano dalla periodicità del parlato o del rumore. Questa tecnica può essere vista come un caso particolare, più complesso e generale, di filtraggio adattivo multicanale.

Con altri metodi di denoising si cerca di modellare il segnale o il rumore attraverso una conoscenza a priori di questi segnali. Nel caso di segnali vocali, si utilizzano modelli a coefficienti cepstrali, Hidden Markov Model o modelli basati sul pitch traking.

Un altro disturbo che riduce in maniera significativa l'intelligibilità del segnale audio è il "clipping". Si tratta di un processo di degradazione che si manifesta quando il segnale

viene troncato in quanto ha superato il livello massimo di input di un sistema di acquisizione. Questa distorsione è frequente in sistemi telefonici e registratori di basso costo. All'ascolto questo tipo di distorsione produce una grave compromissione della qualità sonora e anche dell'intelligibilità.

In Dahimene (Dahimene et alii, 2008) l'algoritmo di declipping è basato sulla predizione lineare: i coefficienti del filtro, calcolati a partire dai campioni precedenti il clipping e quindi senza distorsioni, vengono utilizzati per ricostruire i campioni tagliati.

Nell'uso delle tecniche sopra esposte, per ciascun particolare tipo di miglioramento del segnale di interesse deve essere operato un compromesso tra la quantità di rumore rimosso e la distorsione introdotta come un effetto collaterale. Se viene rimosso troppo rumore, la distorsione può essere peggiore per l'ascoltatore che la permanenza del rumore.

Le soluzioni sopra elencate permettono, con efficienza maggiore o minore a seconda dell'importanza e della tipologia del disturbo, del tempo impiegato per correggerlo e della qualità degli strumenti di elaborazione del segnale a disposizione, di migliorare la "qualità" del segnale audio. Con una migliore qualità audio è sicuramente più semplice effettuare la trascrizione del segnale registrato. È tuttavia probabile che, in quasi tutti i casi, non si sia ottenuto alcun reale incremento della intelligibilità ma piuttosto un peggioramento, più o meno lieve.

Boll (Boll, 1991) osserva che: «... il gruppo ha concluso che non vi è attualmente un approccio che migliori l'intelligibilità misurata sulla base di un test diagnostico come il DRT». Tale affermazione probabilmente è ancora valida: infatti, in un recente lavoro di Hu e Louzou (Hu & Loizou, 2007), nessuno degli 8 sistemi di restauro del segnale da loro utilizzati è riuscito a migliorare l'intelligibilità.

3. OBIETTIVO DEL LAVORO

Come abbiamo precedentemente osservato, in molti casi, specialmente in presenza di segnale incomprensibile o parzialmente incomprensibile perché fortemente degradato, le diverse tecniche di riduzione del rumore o più in generale di elaborazione del segnale non si sono rivelate in grado di operare un significativo miglioramento dell'intelligibilità. Obiettivo di questo lavoro è verificare se questa affermazione sia da considerarsi generalmente valida, con particolare riferimento ai segnali normalmente disponibili nelle applicazioni forensi. Inoltre si è ritenuto opportuno verificare se il sistema oggettivo di valutazione dell'intelligibilità da noi proposto sia in grado di valutare correttamente l'intelligibilità del segnale prima e dopo un restauro. Il metodo denominato Single-Side Intelligibility Measures è descritto in Costantini (Costantini et alii, 2010) e si basa sullo Speech Transmission Index (STI), modificato in modo da poter essere utilizzato con un approccio di tipo Single-Sided, cioè a partire dal solo segnale rumoroso. Questa caratteristica, peraltro rara nei sistemi di valutazione oggettivi dell'intelligibilità, è indispensabile in ambito forense in quanto non è mai disponibile il segnale "non rumoroso" da utilizzare per il confronto. La misura oggettiva dell'intelligibilità consentirebbe di valutare l'efficacia dei sistemi di "speech enhancement" in modo più rapido e molto meno costoso di quanto avviene utilizzando un gruppo di ascoltatori.

4. IL CORPUS

Il corpus, impiegato per i test soggettivi e oggettivi, è stato tratto dal materiale audio del progetto europeo SAM EUROM1 (Chen et alii, 1995); in particolare sono state estratte 24

frasi italiane, con o senza significato, lette da 4 voci diverse, due maschili e due femminili. Questo materiale fonico, che presenta il vantaggio di essere stato equalizzato in ampiezza, è stato poi degradato sia inserendo rumore additivo tipo Babble, sia inserendo un disturbo convolutivo. Il rumore additivo è aggiunto a tre diversi livelli di rapporto segnale/rumore (S/N = 4, 0, -4 dB), mentre il disturbo convolutivo (riverberazione), che simula due diversi ambienti, Office e Lobby, ha un tempo di riverberazione T60 pari a 0.95s e 2.03s rispettivamente. In tal modo, si sono ottenute 24 frasi, ciascuna con un diverso livello di degradazione, come illustrato in Tabella I.

I segnali relativi alle 24 frasi variamente degradate sono stati inviati a 4 diversi esperti nel settore dello speech enhancement in applicazioni forensi chiedendo loro di operare un restauro del segnale volto a migliorarne l'intelligibilità attraverso le metodiche da loro abitualmente adottate. Non è stato loro chiesto di dare una descrizione dettagliata delle procedure, sia perché alcuni operatori non conoscono approfonditamente gli algoritmi utilizzati dai sistemi di speech enhancement, sia perché ciò che si voleva valutare era il sistema esperto + sistema filtraggio. Tuttavia si ritiene che abbiano operato al meglio con l'obiettivo di migliorare il più possibile l'intellegibilità del segnale reso disponibile. Inoltre, soprattutto per permettere ai citati esperti un'autovalutazione dei miglioramenti ottenuti, è stato chiesto loro di procedere alla trascrizione delle frasi. Le trascrizioni fornite dagli esperti non sono state poi utilizzate nella valutazione. Attraverso questa procedura si sono ottenuti, in totale, 5 corpora: quello degradato iniziale e 4 diversi corpora, ciascuno dei quali è stato restaurato da un differente esperto (o meglio da un differente laboratorio).

S/N	+ 4 dB	0 dB	- 4 dB
Office	HO CANTATO TANTO CHE SONO RAUCO E SENZA FIATO	HA AVUTO L'INTUITO DI RIMUOVERE TUTTI I POSSIBILI OSTACOLI	MI HA ZITTITO CON UN SUONO GUTTURALE, QUASI MAGNETICO
	SONO STANCO DI IMMETTERE DATI NEL COMPUTER	CHE TI SALTA IN MENTE DI ORDINARE SOLO PER TE?	IN FONDO, E' PIU' SIMPATICO IL GUFO CHE IL LEONE
	MI SONO ARRABBIATO CON LUI E HO URLATO A LUNGO	SUONA ANCHE IL LIUTO, MA UN PO' MALE	CHISSA' SE E' MEGLIO L'OLIO DI SOIA O QUELLO DI MAIS
	DALL'ODORE SI DIREBBE COGNAC DENATURATO	LO AGITI UN PO' E HAI GIA' OTTENUTO UN COCKTAIL SCECHERATO	PER LE GOCCIOLE DI CREMA SERVONO MOLTI TUORLI
Lobby	E' IL PERIODO PIU' IELLATO DEI MIEI ULTIMI ANNI	E' UN VERO AMATORE DI PESCA SUBACQUEA	COGLIETE L'OCCASIONE PER IMPIANTARE UNA MAGLIERIA
	GLI HO DETTO LA VERITA' E LUI SE NE E' ANDATO MOGIO MOGIO	NON LO VEDO ARRIVARE: SARA' ULTIMO	IL GALLO SI E' AVVENTATO PER GHERMIRE LA PREDA
	FINISCI COLL'AVERE UN ALGORITMO SDOPPIATO	CI SONO MOMENTI IN CUI SEI ANNOIATO DI TUTTO	TUTTA LA ZONA DELL'OLGIATA E' MOLTO RICCA
	ALT, FERMATEVI O MI SENTO MALE	LA REGIA MI E' SEMBRATA ACCURATA, MA NON BRILLANTE	C'E' UNO SCREZIO SERIO CON TUTTA LA MIA FAMIGLIA

Tabella I: Corpus utilizzato nelle prove

4. MISURE SOGGETTIVE DI INTELLIGIBILITÀ

Per valutare attraverso le misure soggettive di ascolto l'intelligibilità dei segnali resi disponibili, quello originale da restaurare e quelli elaborati dai 4 esperti, tali segnali sono stati sottoposti ad un gruppo di 12 ascoltatori normudenti utilizzando un software sviluppato appositamente per questo scopo in ambiente Max/MSP (Cycling74, Inc.).

Il software consente l'ascolto del segnale e la sua trascrizione in una finestra denominata "insert your answer here". Si procede nel seguente modo: si scrive il proprio nome, si seleziona la sessione, composta da 8 stimoli a differenti livelli di degradazione, quindi si attiva il tasto "play" per la riproduzione dello stimolo sonoro, che può essere ascoltato più volte, infine si scrive la frase che si ritiene aver compreso. Al termine di ogni sessione viene registrato un testo contenente i risultati prodotti dal soggetto. Esiste, inoltre, la possibilità di usufruire di una fase di addestramento che consente di comprendere meglio lo svolgimento della prova e di regolare il livello del segnale audio. La Fig. 1 mostra l'interfaccia dell'applicazione.

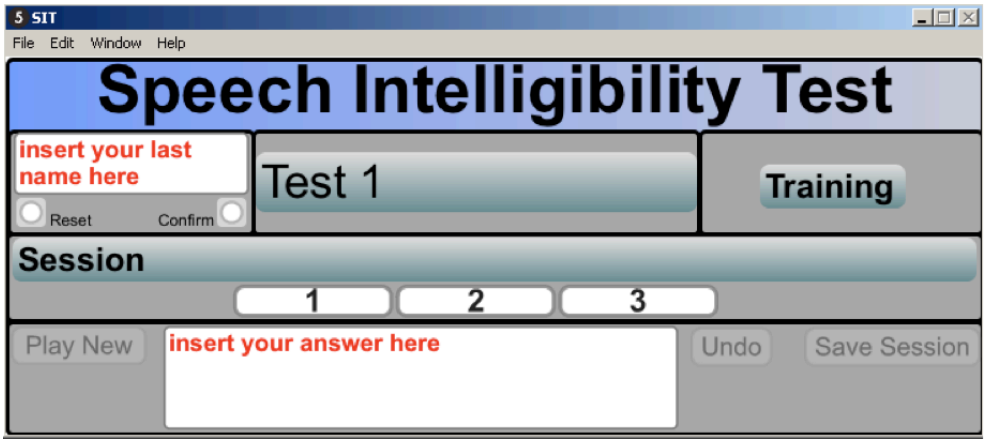


Figura 1: Interfaccia per le prove soggettive

5. MISURE OGGETTIVE DI INTELLIGIBILITÀ

La misura dell'intelligibilità, sia soggettiva che oggettiva, ha dato origine a numerose proposte, ciascuna volta a rispondere ad una particolare esigenza (Kitawaki & Yamada, 2007). Per quanto attiene alle misure soggettive, è stato subito evidente che il risultato dipende strettamente dal materiale audio utilizzato per le prove (Romito 2004), (Steeneken, 2002). A causa delle restrizioni imposte dalla lingua, l'intelligibilità delle frasi sarà maggiore, a parità di rapporto S/N, di quella delle parole e quella delle parole maggiore di quella dei logatomi (gruppi di fonemi, appartenente a un elenco stabilito per ogni lingua, che non costituisce una parola essendo priva di significato). Inoltre, se si vuole stimare correttamente le prestazioni di un sistema, è opportuno utilizzare un insieme di parole che rappresentino statisticamente l'occorrenza dei fonemi nella lingua di interesse. Infine sono

stati proposti test soggettivi diagnostici, volti ad indagare quali siano le coppie di fonemi che vengono maggiormente tra loro confusi (DRT). Numerose anche le proposte di misure oggettive per stimare l'intelligibilità (Ma et alii, 2009; Kryter, 1969), a cominciare dall'Articulation Index, o indice di articolazione, calcolato dividendo lo spettro acustico in bande a terzi di ottava e calcolando per ogni banda il contributo all'intelligibilità nelle condizioni di disturbo presenti (Liu et alii, 2008). Una rassegna dei principali metodi proposti è riportata nella Tab. II.

Method	Standard ref.	Type	Comments
PB - Phonetically Balanced Word	ISO TR 4870	Subjective	
MRT - Modified Rime Test		Subjective	
DRT - Diagnostic Rime Test		Subjective	
AI - Articulation Index	ANSI S 3.5	Objective	instrument based
STI - Speech Transmission Index	IEC 60268-16 - 1998	Objective	instrument based
%AL - Articulation loss of Consonants	Peutz	Objective	

Tabella II: Metodi più diffusi della misura dell'intelligibilità

Uno dei parametri più accurati per valutare l'intelligibilità è lo Speech Transmission Index (STI), basato sulla misura della Modulation Transfer Function (MTF).
Le caratteristiche acustiche dell'ambiente e il rumore di fondo determinano una riduzione delle MTF del segnale test, dalla sua emissione alla sua ricezione, e di conseguenza una riduzione dello STI. Le metodologie di misura e le tecniche per il calcolo di MTF e STI sono regolate dalla normativa IEC-60268-16 (2003).
L'indice STI è un valido indicatore dell'intelligibilità media del parlato, adatto per misurare oggettivamente l'intelligibilità di un canale di trasmissione perché utilizza un approccio di tipo Double-Sided, ovvero sulla base di un confronto tra il segnale vocale pulito e il segnale trasmesso. In un contesto forense, questo approccio non è utilizzabile, perché il perito ha solo la versione rumorosa del segnale, quella che proviene da intercettazioni ambientali o telefoniche. È importante, quindi, poter valutare l'intelligibilità con un approccio di tipo Single-Sided, cioè basato sul solo segnale rumoroso.
La misura STI Single-Sided è calcolata come segue. Il segnale rumoroso viene filtrato passa-banda in sette bande di ottava a partire da banda 125 Hz a 8000 Hz. L'inviluppo di ciascuna banda è stato calcolato utilizzando la potenza del segnale. In particolare,

consideriamo un segnale discreto nel dominio del tempo x(n) filtrato nella banda d'ottava k, si definisce la funzione inviluppo come

$$Env_k(m) = \frac{1}{N_e - 1} \sum_{n=mh}^{mh+N_e-1} h(n-mh)\left[x(n)\right]^2 \tag{1}$$

dove N_e è la dimensione della finestra, h è l' hop size, m □ {0, 1, 2,…, M} l'hop number, h(n) è la finestra di Hanning e n la variabile somma. Quindi, calcoliamo l'inviluppo spettrale normalizzato come

$$s_{k,f_i} = \frac{\left| \sum_{p=0}^{N_s-1} w(p) Env_k(p) \cdot e^{-\frac{i2\pi p f_i}{F_s}} \right|}{\sum_{p=0}^{N_s-1} Env_k(p)} \tag{2}$$

dove *Ns* è la dimensione della finestra, *Fs* è la frequenza di campionamento, *fi* sono le 14 frequenze nel range 0,63 Hz a 12,5 Hz ad 1/3 di ottava, *w(p)* è una finestra rettangolare e p è la variabile somma. L'SNR in ciascuna banda è calcolato come

$$SNR_{k,f_i} = 10 \log_{10}\left(\frac{s_{k,f_i}^2}{1 - s_{k,f_i}^2}\right) \tag{3}$$

e successivamente limitato tra [-15, 15] dB. L'indice di trasmissione (TI) in ciascuna banda è calcolato secondo la seguente equazione

$$TI_{k,f_i} = \frac{SNR_{k,f_i} + 15}{30} \tag{4}$$

Per ogni banda di ottava, il valor medio di TI per una frequenza specificata, dà la Modulation Transfer Index (MTI)

$$MTI_k = \frac{1}{n}\sum_{i=1}^{n} TI_{k,f_i} \tag{5}$$

Infine, la misura basata sullo STI è ottenuta come media ponderata degli MTI su sette bande di ottava

$$STI = \sum_{k=1}^{7} W_k \cdot MTI_k \tag{6}$$

La somma di W_k è 1 e i valori numerici possono essere trovati in (Payton, 1999). La Fig. 2 mostra l'intero diagramma a blocchi della misura basata sullo STI.

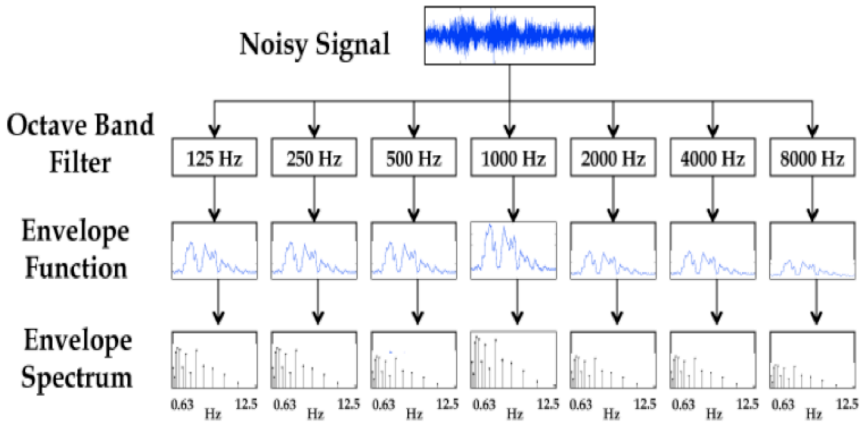


Figura 2: Diagramma a blocchi delle misure basate sullo STI

È stata sviluppata in ambiente MatLab (The MathWorks, Inc) un'applicazione per sistemi Windows, che calcola localmente lo STI a partire da un segnale vocale rumoroso, operando su finestre temporali di 500 ms. In particolare, si calcola la Trasformata di Fourier dell'inviluppo del segnale normalizzata all'area dell'inviluppo del segnale stesso. Inoltre, per avere un'idea globale dell'intelligibilità del segnale considerato, l'applicazione calcola una statistica sui valori di STI ottenuti, ricavando media, varianza e distribuzione.

6. RISULTATI

I risultati delle misure soggettive ed oggettive dell'intelligibilità sono riportati rispettivamente nelle Tabb. III e IV. Nella Tab. III i valori di intelligibilità stimati sono stati ottenuti mediando i valori delle misure sulle frasi appartenenti alla stessa classe di degradazione; inoltre, i 4 esperti sono contrassegnati dalle etichette Expert1,..Expert4, mentre il segnale originale è contrassegnato dall'etichetta Original. L'intelligibilità è misurata in termini di Word Accuracy Rate (WAR). Dall'esame della tabella, o meglio dalla rappresentazione dei dati riportata la Fig. 3, si evidenzia che le metodiche di "enhancement" utilizzate dagli esperti non sono state in grado di migliorare in misura significativa l'intelligibilità del segnale, anzi in alcune condizioni l'operazione di enhancement porta ad un peggioramento significativo dell'intelligibilità. Ad esempio, l'intelligibilità della condizione di minor degrado (+4dB, office), pari al 90% circa, viene ridotta a poco più del 50% dal sistema di enhancement utilizzato dall'esperto 4. Nella Tab. IV sono riportati i valori di intelligibilità ottenuti utilizzando il SW SSIM descritto nel precedente paragrafo. Dall'esame della tabella e dalla Fig. 4 si osserva che il sistema SSIM fornisce risultati che portano a sovrastimare l'intelligibilità, soprattutto a livelli alti di rumore (-4 dB S/N). Si osserva inoltre che il sistema oggettivo SSIM viene "ingannato" dalle metodiche di enhancement e stima un miglioramento di intelligibilità di circa 10 dB che non corrisponde ai dati soggettivi. E' infine degno di nota il fatto che gli esperti ottengano dalla valutazione oggettiva risultati quasi identici.

SUBJECTIVE WAR [%]	SIGNAL PROCESSING				
	ORIGINAL	EXPERT 1	EXPERT 2	EXPERT 3	EXPERT 4
LOBBY -4dB	17.45	8.13	20.25	8.82	9.09
0FFICE -4dB	26.77	10.70	11.83	14.75	15.96
LOBBY 0dB	60.00	31.39	55.41	39.19	51.02
0FFICE 0dB	56.89	46.59	46.49	51.02	45.69
LOBBY +4dB	66.67	57.93	58.78	61.25	51.25
0FFICE +4dB	87.80	72.08	79.86	77.38	56.58

Tabella III – Risultati delle prove soggettive

OBJECTIVE WAR [%]	SIGNAL PROCESSING				
	ORIGINAL	EXPERT 1	EXPERT 2	EXPERT 3	EXPERT 4
LOBBY -4dB	57.30	70.07	68.42	66.76	71.13
0FFICE -4dB	55.25	74.30	70.82	69.52	74.77
LOBBY 0dB	66.71	78.90	77.35	73.58	78.43
0FFICE 0dB	64.92	82.70	81.73	78.83	83.09
LOBBY +4dB	72.83	84.03	84.72	81.46	84.42
0FFICE +4dB	77.23	88.26	89.98	85.96	89.11

Tabella IV - Risultati delle prove oggettive

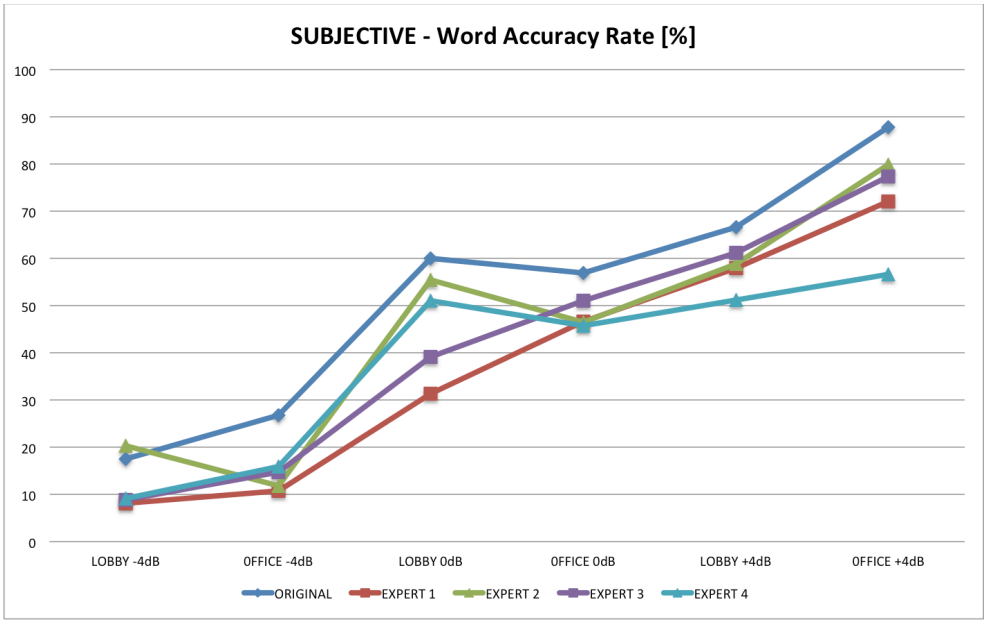


Figura 3: Prove soggettive – I diversi sistemi di enhancement a confronto.

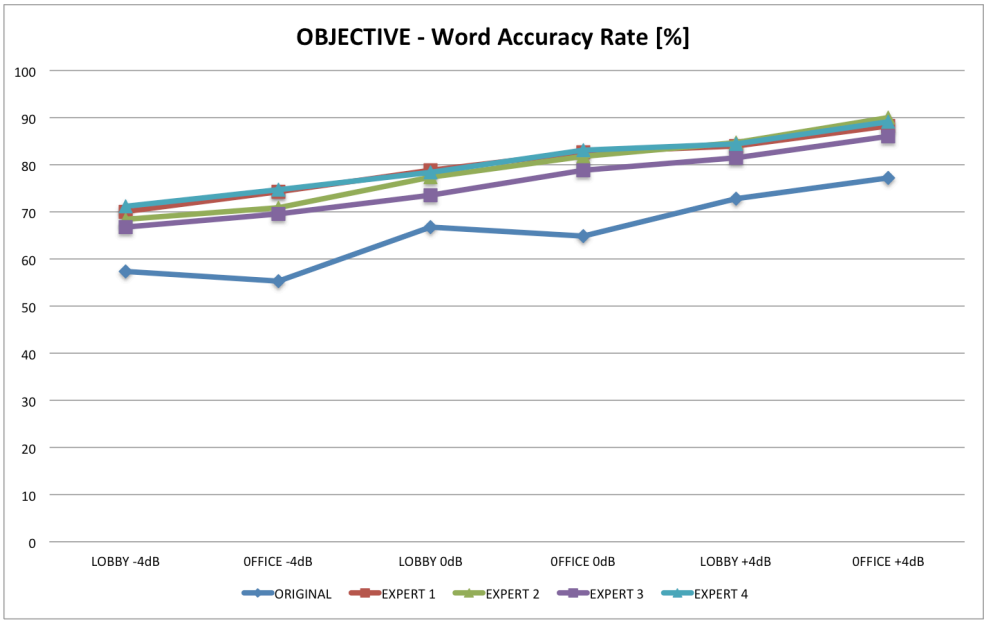


Figura 4: Prove oggettive – I diversi sistemi di enhancement a confronto.

BIBLIOGRAFIA

Alexander J.M., Jenison R.L., Kluender K.R. (2011), Real-Time Contrast Enhancement to Improve Speech Recognition, *Plosone*.

Boll S.F. (1979), Suppression of acoustic noise in speech using spectral subtraction, *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. ASSP-27, pp. 112-120.

Boll S.F. (1991), Speech Enhancement in the 1980s: Noise Suppression with Pattern Matching, in *Advances in Speech Signal Processing*, Dekker.

Brookes M., Gaubitch N., Hukvale M., Naylor P. (2008), Speech Cleaning literature review, *Clear Project.*

Cerrato L., Paoloni A. (1999), Are Transcriptions of Speech Material Recorded by Means of Bugs Reliable?, in *Sixth European Conference on Speech Communication and Technology* (EUROSPEECH'99) Budapest.

Cerrato, L., Paoloni, A. (1996), La "situazione comunicativa" nelle intercettazioni ambientali, in *Atti delle VII Giornate del Gruppo di Fonetica Sperimentale*, Napoli, 221-229.

Chen D., Fourcin A., et alii, (1995), EUROM A spoken language resource for the EU, *ESCA EUROSPEECH '95*, Madrid.

Costantini G., Paoloni A., Todisco M. (2010), Objective Speech Intelligibility Measures Based on Speech Transmission Index for Forensic Applications, *39th International AES Conference on Audio Forensics: Practices and Challenges*, Hillerød, Denmark, June 17–19, pp. 182-188.

Dahimene A., Noureddine M., Azrar A. (2008), A simple algorithm for the restoration of clipped speech signal, *Informatica*, vol. 32, pp. 183-188.

Hilkhuysen G., Huckvale M. (2010), Signal proprieties reducing intelligibility of speech after noise reduction, *EUSIPCO 2010*.

Hu Y., Loizou, C. (2007), A Comparative Intelligibility Study of Speech Enhancement Algorithms, *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007*. IEEE International Conference on Volume 4, Issue, 15-20 April, Page(s): IV-561 - IV-564.

ISO 9921-1 (1996), Ergonomic assessment of speech communication – Part 1: Speech interference level and communication distances for persons with normal hearing capacity in direct communication (SIL method), International Standards Organization.

Kitawaki N., Yamada T. (2007), Subjective and Objective Quality Assessment for Noise Reduced Speech, *ETSI Workshop on Speech and Noise in Wideband Communication*, Sophia Antipolis, France.

Kryter K., (1962), Methods for the calculation and use of the Articulation Inde, *JASA 34*, 1689–1697.

Lim J.S. (1978), Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise, *IEEE Trans. Acoust, Speech. Signal Processing*, vol. ASSP-26, no. 5, pp. 471-472.

Lim J.S., Oppenheim A. V. (1979), Enhancement and bandwidth compression of noisy speech", *Proc. IEEE*, vol. 67.

Liu W. M., Jellyman K. A., Evans N. W. D., and Mason J. S. D., (2008), Assessment of Ob-jective Quality Measures for Speech Intelligibility, INTERSPEECH 2008, *9th Annual Conference of the International Speech Communication*, Association Brisbane, Australia September 22-26.

Loizou P.C. (2007), Speech Enhancement: Theory and Practice, *Taylor & Francis*.

Ma J., Hu Y., Loizou C. (2009), Objective measures for predicting speech intelligibility in moist conditions based on new band importance functions" *JASA 125*.

Payton K. L., (1999), A method to determine the speech transmission index from speech waveforms, *JASA 106*, 3637-3648.

Romito L. (2004), Il contesto, l'intelligibilità, il rapporto segnale-rumore, *Atti AISV 2004*, Padova, 2-4.

Steeneken Herman J.M., (2002), The Measurement of Speech Intelligibility ", *TNO Human Factors*, Soesterberg, the Netherlands.

Vaseghi S. V. (1996), Advanced Signal Processing and Digital Noise Reduction, *Wiley Teubner*.

UN'INTERFACCIA ACUSTICA INTELLIGENTE PER COMUNICAZIONI IMMERSIVE IN AMBIENTI NON STAZIONARI

Danilo Comminiello[*], Michele Scarpiniti[*], Raffaele Parisi[*], Albenzio Cirillo[†], Mauro Falcone[†], Aurelio Uncini[*]
* DIET, "Sapienza" Università di Roma - † Fondazione Ugo Bordoni
{*danilo.comminiello, michele.scarpiniti, raffaele.parisi*}*@uniroma1.it, albenzio.cirillo@gmail.com, falcone@fub.it, aurel@ieee.org*

1. ABSTRACT

Oggigiorno, le comunicazioni vocali immersive stanno diventando sempre più parte integrante del nostro modo di interagire a distanza con altri utenti. Tuttavia i nuovi scenari immersivi sono molto sensibili alle sorgenti interferenti presenti nell'ambiente in cui avviene la comunicazione, per cui la qualità del parlato viene compromessa ancora di più. Allo scopo di mantenere elevata la qualità percepita della comunicazione nuovi sistemi intelligenti vengono applicati ai dispositivi di comunicazione. Questo lavoro propone una nuova interfaccia acustica intelligente in grado di migliorare la qualità del parlato in applicazioni di teleconferenza in condizioni ambientali avverse. In particolare, la novità di tale interfaccia intelligente consiste in una tecnica originale di *beamforming* adattativo progettata *ad hoc* per risultare robusta alle sorgenti interferenti non stazionarie in ambienti riverberanti e rumorosi. Il sistema adattativo proposto è basato su un'*architettura combinata* in grado di offrire complessivamente capacità di tracciamento e prestazioni a regime superiori ai sistemi convenzionali. Inoltre, la flessibilità del sistema proposto permette di elaborare le informazioni acquisite offrendo prestazioni ottimali in presenza di qualsiasi tipo di variazione indotta dalle sorgenti interferenti sull'ambiente di ascolto. La valutazione del sistema è condotta rispetto a diverse condizioni di lavoro e quindi rispetto a specifiche problematiche che possono verificarsi in scenari reali di teleconferenza. Le prestazioni sul miglioramento della qualità del segnale derivato dall'utilizzo del sistema proposto sono valutate attraverso la stima del *rapporto segnale-rumore* (SNR), che quantifica in particolare la riduzione di rumore interferente ottenuta. Gli esperimenti mostrano come l'interfaccia acustica intelligente proposta, in presenza di sorgenti interferenti non stazionarie, riesca ad ottenere un miglioramento qualitativo del parlato rispetto ai sistemi convenzionali utilizzati.

2. INTRODUZIONE

Da sempre l'obiettivo principale dei sistemi di telecomunicazione è quello di abbattere le barriere naturali imposte dalla lontananza fra gli utenti. Negli ultimi anni, notevoli progressi si sono riscontrati nella riduzione dei limiti spaziali e temporali; tuttavia la lontananza fisica fra utenti in una comunicazione a distanza rimane un evidente ostacolo. Inoltre, attualmente le esigenze degli utenti volgono verso un tipo di comunicazione interattiva che permetta loro di condividere il medesimo ambiente. A tale scopo, la ricerca si sta focalizzando su una nuova esperienza di comunicazione, detta "*immersiva*", che consente ad un utente situato in un determinato ambiente di essere calato "percettivamente" in una comunicazione vocale con una o più persone situate in ambienti diversi, avendo così la sensazione di condividere lo stesso spazio acustico.

L'importanza della qualità dei segnali vocali in una comunicazione immersiva è dettata dalla necessità di ricreare artificialmente la percezione uditiva che l'utente avrebbe se si trovasse realmente nell'ambiente remoto. A tale scopo, l'attenzione è volta allo sviluppo di tecniche di interfacciamento acustico ed elaborazione dei segnali vocali necessarie per la realizzazione di sistemi di comunicazione immersivi.

Un'*interfaccia acustica intelligente* è generalmente composta da uno o più altoparlanti e da una schiera di microfoni associata ad un sistema di elaborazione dei segnali vocali acquisiti. Tuttavia, in questo lavoro, l'attenzione è focalizzata sulla fase di acquisizione. Lo scopo principale infatti, è quello di acquisire le informazioni ambientali rifacendosi all'ascolto binaurale caratteristico del sistema uditivo umano, in grado di offrire all'utente una percezione realistica dell'ascolto (Blauert, 1997). Tuttavia ciò che rende "intelligente" un'interfaccia acustica è la sua capacità di adattarsi ai requisiti dell'utente e alle condizioni dell'ambiente. A tale scopo, un'interfaccia acustica intelligente deve disporre di un sistema di elaborazione di segnali acustici e vocali in grado di far fronte alla presenza di rumore, e/o di sorgenti interferenti, che sovrapposte al segnale vocale principale compromettono l'intelligibilità di quest'ultimo e dell'intera comunicazione (Huang, 2011).

Nel corso degli ultimi anni sono state sviluppate diverse tecniche di elaborazione dei segnali vocali che traggono origine dagli studi sull'apparato binaurale umano (Lotter & Vary, 2006; Stern et al., 2008). Utilizzando un'interfaccia microfonica è possibile elaborare i segnali catturati da ciascun microfono secondo le tecniche di *beamforming adattativo*, ovvero combinandoli in un'unica forma d'onda in cui è esaltato il segnale vocale proveniente da una specifica direzione e, pertanto, è attenuato ogni rumore proveniente dalle altre direzioni (Brandstein & Ward, 2001; Li & Stoica, 2006). Il beamforming può essere interpretato come il risultato di un filtro spaziale, in quanto prevede un trattamento del suono differente a seconda del punto nello spazio dove questo viene acquisito. Gli studi presenti in letteratura mostrano come l'efficacia del beamforming cresca all'aumentare del numero di microfoni utilizzati nella schiera microfonica (Brandstein & Ward, 2001; Benesty et al., 2010). Tuttavia vi è ancora necessità di creare algoritmi di beamforming sempre più robusti alle variazioni dovute allo spostamento delle sorgenti presenti all'interno dell'ambiente e alle condizioni di rumore (Li & Stoica, 2006).

Un classico sistema di beamforming adattativo è il *Generalized Sidelobe Canceller* (GSC) (Griffiths & Jim, 1982), composto da un *delay-and-sum beamformer* (DSB), che ha lo scopo di focalizzare la sorgente vocale principale, e da un *blocco di cancellazione adattativa del rumore* (ANC, *adaptive noise canceller*), che riduce la potenza del rumore di fondo nel segnale di uscita. Il blocco ANC nelle applicazioni acustiche comporta l'utilizzo di filtri digitali FIR dell'ordine di centinaia o anche migliaia di coefficienti, i cui valori devono essere stimati adattativamente in modo da ridurre il rumore quanto più possibile (Sayed, 2008; Uncini, 2010). La scelta dell'algoritmo adattativo è la parte critica di un sistema GSC, in quanto deve essere garantita una stima rapida ed efficace dei valori del filtro.

Pur essendo una tecnica molto consolidata, il GSC soffre particolarmente quando l'ambiente in cui si trovano le sorgenti vocali sono molto rumorose e altamente non stazionarie (Gannot et al. 2001; Gannot & Cohen, 2004). Uno scenario classico di questo tipo si verifica quando nello stesso ambiente, oltre alla sorgente vocale desiderata, sono presenti altre sorgenti vocali che producono un effetto rumoroso di tipo "cocktail party" che interferisce con la sorgente desiderata provocando un deterioramento dell'intelligibilità del segnale vocale desiderato (Comminiello et al., 2011). Quando poi queste sorgenti interferenti si

muovono nello spazio l'interfaccia acustica si adatta con più difficoltà all'ambiente circostante, generando così un peggioramento delle prestazioni o, al più nessun miglioramento qualitativo del parlato acquisito (Gannot et al. 2001).

Per ovviare al problema descritto, in questo lavoro verrà proposta una tecnica originale di beamforming adattativo progettata *ad hoc* per risultare robusta alle sorgenti interferenti non stazionarie in ambienti riverberanti e rumorosi. La tecnica proposta è basata su un'*architettura combinata* in cui, per ciascun canale microfonico, viene effettuata una combinazione convessa di due filtri adattativi appartenenti a due classi differenti in modo da ottenere un algoritmo in grado di offrire complessivamente capacità di tracciamento superiori ai singoli filtri (Martinez-Ramón et al., 2002; Arenas-García et al., 2006; Silva & Nascimento, 2008). La flessibilità del sistema proposto permette dunque di regolare in maniera ottimale i parametri dei filtri utilizzati, così che il sistema possa elaborare le informazioni acquisite offrendo prestazioni ottimali nel medio/lungo periodo, ossia a convergenza o quando si verifica una situazione di stazionarietà, e al tempo stesso garantendo dei rapidi tempi di risposta nel breve periodo quando si verifica una non stazionarietà nell'ambiente, il che si traduce con una velocità di convergenza elevata (Arenas-García, 2006). L'interfaccia acustica intelligente descritta risulta così una soluzione robusta alle repentine variazioni indotte dalle sorgenti interferenti sull'ambiente di ascolto.

La valutazione del sistema descritto è condotta rispetto a diverse condizioni di lavoro e quindi rispetto a specifiche problematiche, focalizzando l'attenzione sul caso in cui siano presenti sorgenti non stazionarie nell'ambiente in cui avviene la comunicazione immersiva.

Il lavoro è organizzato nel seguente modo: nel Paragrafo 3 vengono descritte le problematiche relative alle nuove comunicazioni vocali immersive e viene definito il ruolo delle interfacce acustiche intelligenti all'interno di questo contesto applicativo; il Paragrafo 4 descrive il sistema di beamforming utilizzato che integrerà la nuova architettura adattativa proposta, che verrà presentata nel Paragrafo 5. Il Paragrafo 6 contiene un'ampia descrizione degli scenari sperimentali, delle analisi prestazionali e dei risultati ottenuti. Infine, nel Paragrafo 7 sono riportate le conclusioni finali.

3. LE INTERFACCE ACUSTICHE INTELLIGENTI

Allo scopo di rendere più chiare le motivazioni sulle quali è basato il presente lavoro, introduciamo in questa sezione le problematiche principali che vengono affrontate, partendo dal definire il termine "*interfaccia acustica intelligente*" (IAI) e comprendere cosa rende intelligente un'interfaccia acustica. Approfondiremo, inoltre, il loro ruolo all'interno delle comunicazioni immersive.

3.1. L'intelligenza delle interfacce acustiche

Un'interfaccia acustica fornisce un mezzo di scambio di informazioni acustiche fra due o più entità attraverso un'elaborazione del segnale acustico. Più esattamente, un'interfaccia acustica è il *front-end* di un sistema di elaborazione di segnali audio e vocali finalizzato all'estrazione e alla riproduzione delle informazioni (Comminiello, 2011). Una interfaccia acustica è generalmente composta da una schiera di microfoni e da uno o più altoparlanti, come rappresentato in Figura 1. Le schiere microfoniche sono generalmente più performanti di un singolo microfono quando si vuole controllare e/o ridurre il rumore ambientale, le riverberazioni acustiche e il parlato indesiderato (Flanagan et al., 1985).

Figura 1: Un'interfaccia acustica.

Nelle interfacce intelligenti, l'intelligenza può consistere nel predire ciò che l'utente vuole fare, e presentare le informazioni desiderate in base alle esigenze dell'utente (Hefley & Murray, 1993). Le interfacce intelligenti, inoltre, possono anche eseguire un compito in maniera più intuitiva, rendendosi utile all'utente. Dunque l'intelligenza in questo contesto non va intesa come cognizione, ma come utilizzo delle informazioni in maniera appropriata (Hefley & Murray, 1993).

La *Association for Computing Machinery* definisce l'interazione uomo-macchina come "una disciplina che riguarda la progettazione, la valutazione e l'implementazione di sistemi computazionali, che dovranno essere usati dall'uomo" (Hewett et al., 1992). Un ruolo importante nell'interazione uomo-macchina è svolto dalle *interfacce acustiche intelligenti* (IAI). Una IAI traduce l'informazione acustica dall'utente alla macchina, e viceversa, allo scopo di consentire una interazione omogenea tra le parti. Una IAI deve essere in grado di adattarsi all'utente, di acquisire ed elaborare le informazioni ricevute dall'utente, di capire le richieste dell'utente, e di restituire all'utente una risposta che soddisfi le sue richieste, sotto forma di linguaggio naturale o di messaggio multimediale.

Le IAI sono ampiamente utilizzate in diversi campi applicativi, la maggior parte dei quali focalizzati al trattamento delle informazioni vocali (Comminiello, 2011). Nel settore multimediale è possibile pensare ad applicazioni quali l'interazione vocale in tempo reale, l'analisi automatica del parlato, la trascrizione automatica, il riconoscimento e la classificazione di generi e contesti nelle trasmissioni televisive, il riconoscimento del parlatore, l'intrattenimento ad alta interattività. Nella domotica le IAI possono essere impiegate con diversi obiettivi: lo sviluppo di "stanze intelligenti", in cui parlatori e comandi vocali devono essere riconosciuti, la robotica antropomorfa avanzata, lo sviluppo di sistemi di video/audio-sorveglianza. Inoltre, è possibile sfruttare le IAI per sviluppare sistemi di aiuto per disabili, si pensi a sistemi in grado di fornire a persone non vedenti una ricostruzione accurata dell'ambiente acustico.

3.2. Le interfacce acustiche nelle comunicazioni immersive

Dopo anni di incredibile progresso tecnologico nel campo delle telecomunicazioni, la gente non si accontenta più di parlare con qualcun altro a lunga distanza e in tempo reale. Le esigenze attuali sono quelle di collaborare attraverso la comunicazione in modo più produttivo, avendo al tempo stesso la sensazione di stare vicini condividendo lo stesso ambiente. Questo fenomeno viene anche indicato come *esperienza immersiva*. Le comunicazioni immersive sono ormai diventate realtà supportate dalle moderne tecnologie. La sensazione di immersione acustica di una persona è causata dalla risposta agli stimoli acustici generati in un ambiente (Huang et al., 2011).

Le comunicazioni immersive hanno luogo in ambiente in cui sono presenti più sorgenti acustiche, come rappresentato in Fig. 2 (Comminiello, 2011). Fra queste sorgenti alcune

possono essere interferenti e possono provocare una degradazione della qualità e della intelligibilità dell'informazione vocale di interesse. Dunque, l'acquisizione ad alta qualità del parlato di interesse nelle comunicazioni immersive diventa un problema ben più difficile e ambizioso rispetto alle comunicazioni telefoniche in cui il microfono e posizionato in prossimità della bocca di un utente. Infatti, durante una conferenza è possibile udire segnali interferenti provenienti da altre sorgenti presenti in ciascun ambiente, per cui il livello di rumore percepito da ciascun utente può crescere proporzionalmente al numero di partecipanti alla conferenza. Difatti, quando il numero dei partecipanti è elevato e se le sorgenti interferenti non sono adeguatamente controllate, il rumore percepito può raggiungere un livello tale per cui il parlato di interesse viene celato. Dunque, le sorgenti interferenti rappresentano un problema molto minaccioso per la qualità di una comunicazione vocale immersiva (Huang et al., 2006; Huang et al., 2011).

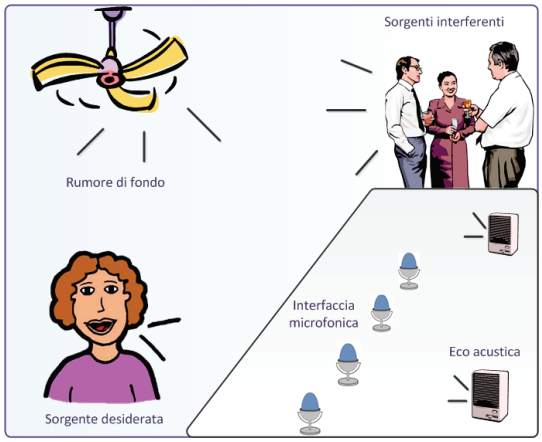


Figura 2. Scenario di comunicazione vocale immersiva in presenza di più sorgenti.

Le comunicazioni immersive offrono grandi opportunità di sviluppo di sistemi di elaborazione di segnali acustici e vocali e implicano l'utilizzo di IAI. La voce è di gran lunga la componente multimediale dominante in un contesto applicativo come quello della conferenza. Una sessione di teleconferenza può proseguire se il collegamento si interrompe, ma non può andare avanti se il collegamento audio viene meno. Perciò, nella progettazione di IAI, oltre a perseguire le capacità multimodali, bisognerebbe non dimenticare mai l'importanza della qualità della voce (includendo anche l'intelligibilità e la naturalezza) e la sinergia intermodale. Inoltre, ci sono grandi possibilità di migliorare questi ultimi due fattori in una teleconferenza immersiva che coinvolge più parti, poiché l'ascolto binaurale è ora possibile e può essere pienamente sfruttato. Questo è un passo necessario per direzionarsi verso le comunicazioni immersive. Se infatti si riceve un'informazione stereofonica, diversificata sui due canali, il nostro sistema uditivo riesce ad estrarre più facilmente il parlato di un singolo interlocutore da un contesto altamente rumoroso.

Una IAI per comunicazioni immersive ha il compito di estrarre da un segnale acustico le informazioni desiderate al fine di svolgere in maniera più efficace e semplificata un successivo lavoro di analisi e/o sintesi dei segnali audio. Questa caratteristica delle IAI è anche nota come *machine listening*. Tuttavia, allo stesso tempo, una IAI deve essere in grado di

riprodurre l'informazione acustica desiderata considerando il fatto che un utente vorrebbe ascoltare il segnale vocale esattamente come se fosse nel campo sonoro originale. Questa caratteristica, invece, è nota come riproduzione spaziale del suono. A tale scopo, una IAI deve replicare quattro attributi fondamentali della comunicazione "faccia-a-faccia":

1. uno scambio bidirezionale delle informazioni;
2. libertà di movimento tramite l'utilizzo di microfoni dislocati nell'ambiente;
3. qualità elevata dei segnali vocali acquisiti a distanza non ravvicinata;
4. realismo spaziale della riproduzione del campo acustico.

Questi requisiti implicano che siano utilizzati schiere di microfoni e di altoparlanti, per cui l'intera struttura di comunicazione vocale deve essere conforme a tali specifiche. Tuttavia, in questo lavoro ci occuperemo prettamente dell'aspetto di acquisizione da parte della IAI, e in particolare della parte di elaborazione dei segnali vocali acquisiti.

4. SISTEMA DI BEAMFORMING ADATTATIVO COMBINATO

Il sistema di beamforming adattativo utilizzato è mostrato in Fig. 3, in cui è possibile notare come il sistema sia composto da un'interfaccia microfonica, un percorso fisso di DSB, e un percorso adattativo di cancellazione dei lobi laterali, in una classica configurazione GSC (Griffiths & Jim, 1982).

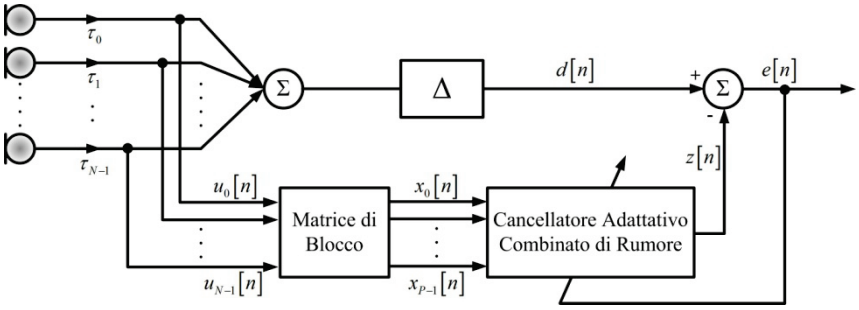


Figura 3. Sistema di beamforming adattativo combinato.

Si consideri un'interfaccia microfonica composta da N sensori. Il segnale $u_i[n]$ acquisito dall'i-esimo microfono, con $i = 0,\ldots, N-1$, è una replica ritardata del segnale obiettivo $s[n]$ convoluto con la risposta impulsiva acustica $\mathbf{a}_i$ che intercorre tra l'i-esimo microfono e la sorgente desiderata con aggiunta di rumore di fondo $v_i[n]$. Il DSB allinea spazialmente i segnali microfonici in base alla direzione della sorgente desiderata, generando un riferimento vocale desiderato $d[n]$:

$$\begin{aligned} d[n] &= \sum_{i=0}^{N-1} u_i[n] \\ &= \sum_{i=0}^{N-1}\sum_{m=0}^{M-1} a_i[m]\, s[n-m-\tau_i] + v[n] \end{aligned} \tag{1}$$

dove si suppone che ogni risposta impulsive acustica fra la sorgente desiderata e ciascun microfono abbia la stessa lunghezza M. τ_i rappresenta il ritardo relativo all'i-esimo microfono. Il ramo adattativo riceve in ingresso i segnali microfonici $u_i[n]$ generando i riferimenti

rumorosi $x_p[n]$, con $p = 0, \ldots, N-2$, per mezzo di una matrice di blocco che lascia passare tutte le componenti al di fuori della direzione della sorgente desiderata (Brandstein & Ward, 2001). Questi segnali vengono quindi filtrati dal *cancellatore adattativo combinato di rumore* che rimuove la correlazione fra la componente di rumore presente nel riferimento vocale e i riferimenti rumorosi generati nel ramo adattativo, generando così il segnale di uscita $e[n]$ del beamformer contenente una sensibile riduzione del rumore. La novità del sistema è rappresentata proprio da quest'ultimo blocco che verrà descritto in dettaglio nel prossimo paragrafo.

5. CANCELLATORE ADATTATIVO COMBINATO DI RUMORE

5.1. L'approccio combinato

Il blocco di cancellazione adattativa del rumore ricopre un ruolo fondamentale nel processo di beamforming. Un ANC classico può essere visto come un sistema MISO (*multiple-input single-output*) composto da un banco di filtri adattativi, ciascuno relativo a un segnale di riferimento del rumore $x_p[n]$. La cancellazione dell'uscita di ciascun filtro dal segnale di riferimento vocale $d[n]$, generato dal DSB, produce un segnale di stima del rumore $z[n]$, che rappresenta la somma dei contributi sottratti al riferimento vocale, e un segnale di uscita del beamformer $e[n]$, che rappresenta il segnale di errore del processo adattativo.

La prerogativa del sistema proposto consiste nell'utilizzo di un ANC combinato al posto dell'ANC classico, sfruttando le proprietà della combinazione adattativa di filtri (Arenas-García et al., 2006). In particolare, la struttura proposta può essere composta da due o più sistemi MISO differenti, ciascuno dei quali offre le proprie capacità a servizio dell'intero sistema. L'utilizzo di questa architettura combinata, infatti, permette di utilizzare sistemi MISO basati sullo stesso algoritmo adattativo, avendo tuttavia un settaggio differenziato dei parametri dei filtri adattativi. La combinazione adattativa dei sistemi MISO permette di utilizzare ad ogni iterazione sempre la configurazione che risulta migliore in termini di prestazioni di filtraggio. A tal fine, e considerando l'obiettivo del lavoro che è quello di progettare un sistema robusto alle sorgenti interferenti non stazionarie, differenzieremo i sistemi MISO in modo da ottenere una riduzione del rumore robusta a condizioni avverse dell'ambiente di comunicazione.

Al fine di eliminare il contributo di rumore proveniente da una sorgente interferente non stazionaria, il sistema di beamforming deve essere abile a tener traccia della sorgente rumorosa nell'ambiente. Dunque il blocco adattativo ANC deve godere di abili proprietà di tracciamento. Una metodo per aumentare le prestazioni di filtraggio nel transitorio, ovvero quando si verifica una non stazionarietà, consiste nel combinare adattativamente due filtri aventi lo stesso algoritmo di aggiornamento ma valori differenti del *passo di adattamento*. È stato infatti dimostrato che una combinazione adattativa, con vincolo convesso, di un filtro veloce, ossia che utilizza un passo di adattamento elevato, con un filtro lento, ossia che utilizza valore piccolo del passo di adattamento, dà luogo ad una velocità di convergenza maggiore, un disallineamento residuo del filtro minore e ad una migliore capacità di tracciamento rispetto ai singoli filtri (Martinez-Ramón et al., 2002; Arenas-García, 2006). Tuttavia, un altro metodo per migliorare ulteriormente le capacità di tracciamento in condizioni non stazionarie si ottiene combinando adattativamente due filtri aventi differenti algoritmi di aggiornamento, uno basato sulla regola gradiente e un altro basato sull'Hessiana (Silva & Nascimento, 2008). Questo tipo di combinazione di filtri sfrutta la rapida velocità di convergenza fornita dal filtro basato sull'Hessiana, e le ottime prestazioni a regime fornite dal

filtro basato sul gradiente, che risultano migliori del filtro basato sull'Hessiana in condizioni non stazionarie (Nascimento et al., 2010, Silva & Nascimento, 2008). In confronto alla combinazione di filtri aventi diverso passo di adattamento, le cui prestazioni in termini di eccesso di errore quadratico medio (EMSE, *excess mean square error*) sono al più pari a quelle del filtro migliore, la combinazione di filtri con differenti regole di aggiornamento forniscono prestazioni che possono superare quelle dei singoli filtri corrispondenti in termini di EMSE (Nascimento et al., 2010).

Il blocco ANC combinato dunque dovrebbe idealmente utilizzare sia una combinazione di sistemi MISO con diversificazione del passo di adattamento, sia un'altra con diversificazione della regola di aggiornamento. Tuttavia, al fine di non sovraccaricare computazionalmente il sistema di riduzione del rumore, in questo lavoro proponiamo un blocco ANC combinato basato sulla combinazione adattativa con vincolo convesso di due sistemi MISO di filtri aventi differenti regole di apprendimento. Un modo molto semplice di diversificare le regole di apprendimento riducendo al minimo la complessità del sistema è quello di utilizzare un algoritmo di filtraggio adattativo basato sull'APA (*affine projection algorithm*) (Ozeki & Umeda, 1994; Uncini, 2010). Questa famiglia di algoritmi basati sulla proiezione affine è caratterizzata da una velocità di convergenza più alta rispetto ai classici algoritmi *least mean square* (LMS), e una complessità computazionale gestibile (Sayed, 2008; Uncini, 2010), motivo per cui l'APA è stato spesso utilizzato in applicazioni di beamforming adattativo (Zheng & Goubran, 2000; Comminiello et al., 2010). Una caratteristica importante dell'APA è che esso è caratterizzato da un *ordine di proiezione*; quando questo valore è maggiore di 1 l'APA è a tutti gli effetti un algoritmo basato sull'Hessiana, mentre quando questo valore è pari ad 1 l'APA si riduce ad un LMS normalizzato (NLMS), per cui il suo algoritmo di aggiornamento segue la regola del gradiente stocastico del primo ordine. Dunque, tutto quello che bisogna fare è considerare un sistema MISO i cui filtri vengono adattati con un APA di ordine 1 e un sistema MISO i cui filtri vengono adattati con un APA di ordine maggiore di 1.

Inoltre, per far fronte al mancato utilizzo della differenziazione del passo di adattamento, possiamo utilizzare per ciascun filtro di entrambi i sistemi MISO un passo di adattamento variabile (VSS, *variable step size*) in modo da ottenere ugualmente un miglioramento della velocità di convergenza nel transitorio (Harris et al, 1986; Shin et al., 2004).

5.2. Schema di combinazione adattativa di sistemi MISO

Lo schema di combinazione adattativa di sistemi MISO proposto in questo lavoro è rappresentato in Fig. 4 e consiste in due sistemi MISO adattativi combinati convessamente filtro per filtro. Ciascun sistema MISO riceve gli stessi segnali di ingresso $x_p[n]$, , con $p = 0, \ldots, N-2$, che altro non sono che i segnali di riferimento del rumore prodotti dalla matrice di blocco, com'è possibile osservare in Fig. 3 e in Fig. 4. Sia $j = 1, 2$ l'indice relativo ai due sistemi MISO, è possibile definire la matrice dei dati in ingresso al p-esimo filtro di ciascun sistema MISO come:

$$\mathbf{X}_{n,p}^{(j)} \in \mathbb{R}^{K_j \times M} = \begin{bmatrix} \mathbf{x}_{n,p}^T \\ \mathbf{x}_{n-1,p}^T \\ \vdots \\ \mathbf{x}_{n-K_j+1,p}^T \end{bmatrix} = \begin{bmatrix} x_p[n] & x_p[n-1] & \cdots & x_p[n-M+1] \\ x_p[n-1] & x_p[n-2] & \cdots & x_p[n-M] \\ \vdots & \vdots & \ddots & \vdots \\ x_p[n-K_j+1] & x_p[n-K_j] & \cdots & x_p[n-M-K_j+2] \end{bmatrix} \tag{2}$$

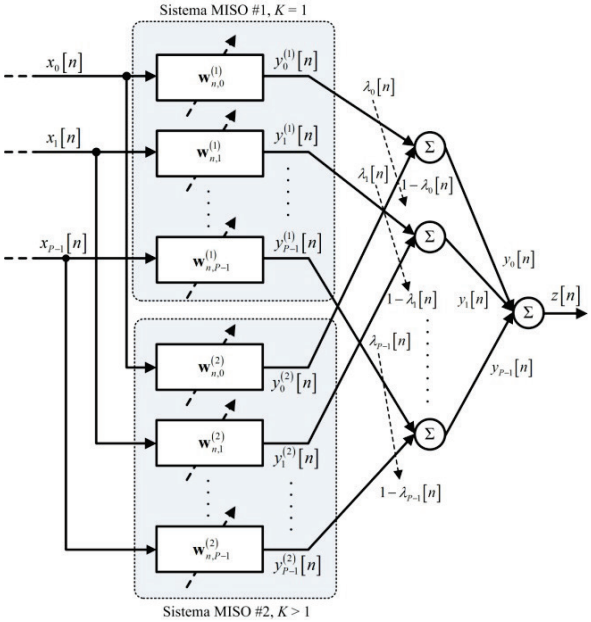


Figura 4. Schema del cancellatore adattativo combinato di rumore

dove K_j rappresenta l'ordine di proiezione per tutti i filtri del j-esimo sistema MISO. All'n-esimo istante temporale, il vettore dei coefficienti del p-esimo filtro appartenente al j-esimo sistema MISO può essere espresso come:

$$\mathbf{w}_{n,p}^{(j)} \in \mathbb{R}^M = \left[w_p^{(j)}[n] \quad w_p^{(j)}[n-1] \quad \cdots \quad w_p^{(j)}[n-M+1] \right]^T \tag{3}$$

Tutti i filtri di ciascun sistema MISO, descritti dall'equazione (3), contengono lo stesso numero di coefficienti, M, e vengono adattati tramite l'algoritmo APA con passo di adattamento variabile (VSS-APA, *variable step size affine projection algorithm*), la cui regola di aggiornamento può essere scritta come:

$$\mathbf{w}_{n,p}^{(j)} = \mathbf{w}_{n-1,p}^{(j)} + \mu_p^{(j)}[n] \mathbf{X}_{n,p}^{(j),T} \left(\delta_j \mathbf{I} + \mathbf{X}_{n,p}^{(j)} \mathbf{X}_{n,p}^{(j),T} \right)^{-1} \mathbf{e}_n^{(j)} \tag{4}$$

dove $\mathbf{e}_n^{(j)} \in \mathbb{R}^{K_j}$ è il vettore di errore del j-esimo sistema MISO, contenente gli ultimi K_j campioni del j-esimo segnale di errore, che viene ricavato come:

$$\mathbf{e}_n^{(j)} = \mathbf{d}_n^{(j)} - \sum_{p=0}^{P-1} \mathbf{y}_{n,p}^{(j)} \tag{5}$$

in cui $\mathbf{d}_n^{(j)} \in \mathbb{R}^{K_j}$ è il vettore contenente gli ultimi K_j campioni del segnale di riferimento vocale desiderato definito in (1), e

$$\mathbf{y}_{n,p}^{(j)} \in \mathbb{R}^{K_j} = \mathbf{X}_{n,p}^{(j)} \mathbf{w}_{n-1,p}^{(j)} \tag{6}$$

è il vettore contenente le K_j proiezioni del segnale di uscita relativo al p-esimo filtro del j-esimo sistema MISO. Il parametro δ_j in (4) rappresenta il *fattore di regolarizzazione* dell'algoritmo ed è comune a tutti i filtri di ciascun sistema MISO. Inoltre, sempre in (4), il parametro $\mu_j^{(p)}[n]$ rappresenta il *passo di adattamento variabile* (VSS) relativo al p-esimo filtro del j-esimo sistema MISO. L'utilizzo di un passo di adattamento variabile consente di avere una risposta maggiore nel transitorio, ossia un valore del passo di adattamento più elevato, e una precisione maggiore a regime condizionata da un valore del passo di adattamento più piccolo. Il passo di adattamento variabile utilizzato viene calcolato ad ogni iterazione e deriva da un processo di minimizzazione della deviazione media quadratica (Paleologu et al., 2008):

$$\mu_j^{(p)}[n] = \left| 1 - \frac{\sqrt{\left|\hat{\sigma}_x^2[n] - \hat{\sigma}_y^2[n]\right|}}{\hat{\sigma}_e^2[n] + \xi} \right| \tag{7}$$

dove ξ è una piccola costante positiva che serve ad evitare divisioni per zero. Il parametro generico $\hat{\sigma}_\alpha^2[n]$, dove $\alpha = \{x, y, e\}$, rappresenta la stima di potenza della corrispondente sequenza generica $\alpha_p[n]$, e può essere calcolato come:

$$\hat{\sigma}_\alpha^2[n] = \beta \hat{\sigma}_\alpha^2[n-1] + (1-\beta)\alpha^2[n] \tag{8}$$

in cui β è un *fattore di smoothing*.

Utilizzando l'equazione di aggiornamento (4) è possibile differenziare i due sistemi MISO considerati semplicemente scegliendo due valori diversi per quanto riguarda l'ordine di proiezione K_j. In particolare, poniamo $K_1 = 1$, come descritto nel sottoparagrafo 4.1, e $K_2 = 4$ (è comunque sufficiente scegliere un valore che sia maggiore di 1).

Dunque, le uscite dei filtri del ciascun sistema MISO, descritte dall'equazione (6), vengono combinate convessamente (Arenas-García, 2006) associando ad un filtro del primo sistema MISO il corrispondente filtro del secondo sistema, come è possibile notare in Fig. 4. Queste combinazioni genereranno perciò $P - 1$ uscite, ciascuna relativa ad un riferimento di rumore:

$$y_p[n] = \lambda_p[n] y_p^{(1)}[n] + \left(1 - \lambda_p[n]\right) y_p^{(2)}[n] \tag{9}$$

dove $\lambda_p[n]$ è il p-esimo *coefficiente di mixing*, che viene aggiornato ad ogni iterazione in modo adattativo tramite l'aggiornamento di un parametro ausiliare $a_p[n]$, legato a $\lambda_p[n]$ tramite la seguente funzione sigmoidale:

$$\lambda_p[n] = \frac{1}{1 + e^{-a_p[n]}} \tag{10}$$

che ha il compito di vincolare i valori di $\lambda_p[n]$ nell'intervallo [0, 1] (Arenas-García, 2006). Dunque, la regola di aggiornamento di $a_p[n]$ può essere scritta come (Azpicueta-Ruiz et al., 2008):

$$a_p[n+1] = a_p[n] + \frac{\mu_a}{r_p[n]} e[n]\left(y_p^{(1)}[n] - y_p^{(2)}[n]\right)\lambda_p[n]\left(1 - \lambda_p[n]\right) \tag{11}$$

dove μ_a è un valore fisso del passo di adattamento con cui vengono aggiornati i parametri di mixing.

Una volta effettuate le combinazioni adattative fra i filtri, è possibile ottenere il segnale di uscita complessivo del cancellatore adattativo combinato di rumore, indicato con $z[n]$, attraverso la somma dei singoli contributi derivanti dalle combinazioni dei filtri, com'è possibile osservare in Fig. 4:

$$z[n] = \sum_{p=0}^{P-1} y_p[n] \tag{12}$$

da cui è possibile ottenere l'uscita complessiva dell'intero beamformer: $e[n] = d[n] - z[n]$.

6. RISULTATI SPERIMENTALI

In questa sezione andremo ad effettuare una valutazione del sistema di beamforming combinato presentato considerando diversi scenari di lavoro possibili.

6.1. Il set-up sperimentale

Lo scenario applicativo è quello di una teleconferenza immersiva, che avviene dunque in viva voce con microfoni dislocati, in cui l'acquisizione delle informazioni desiderate viene disturbata dalla presenza di sorgenti interferenti. L'obiettivo è quello di eliminare le informazioni interferenti allo scopo di inviare all'interlocutore remoto soltanto le informazioni desiderate. In particolare, viene considerato l'ambiente di un solo capo della comunicazione, tuttavia il sistema proposto di elaborazione dei segnali è applicabile indistintamente in tutti gli ambienti coinvolti nella comunicazione.

Lo scenario sperimentale è quello di un laboratorio di ampie dimensioni (7 × 5 × 4 m circa) allestito con diversi materiali, come tende, tavoli, ecc, al fine di ricreare un tipico ambiente di lavoro. Il set-up sperimentale è composto da una sorgente vocale desiderata posta nel campo vicino dell'interfaccia microfonica, due sorgenti interferenti poste lateralmente e rivolte anch'essa verso la schiera microfonica, e infine una sorgente in campo diffuso che riprodurrà rumore ambientale di fondo di varia natura, allo scopo di ricreare un'ampia varietà di condizioni di lavoro. La sorgente desiderata e la sorgente di rumore ambientale sono costituite da sistemi professionali di diffusione acustica, mentre le due sorgenti interferenti sono costituite da due parlatori in modo da avere due sorgenti non stazionarie, e quindi libere di muoversi, all'interno dell'ambiente. L'interfaccia microfonica utilizzata è una classica schiera lineare uniforme composta da 8 microfoni aventi distanza l'uno dall'altro pari a 4 cm.

Per quanto riguarda la parte di riproduzione del segnale desiderato e del rumore di fondo sono stati utilizzati due diffusori attivi Event Tuned Reference 6 (TR6). Per la parte di acquisizione invece sono stati utilizzati 8 microfoni omnidirezionali a condensatore AKG C 562 CM collegati ad un preamplificatore Behringer ADA8000 Ultragain Pro-8 Digital a sua volta interfacciato ad una RME MADI ADI-648.

I segnali vocali desiderati riprodotti dal diffusore fanno parte del database CLIPS, rilasciato nel 2004 e relativo ad un progetto finanziato dal MIUR. In particolare, è stata utilizzata la parte del database relativa al segnale ortofonico realizzata da parlatori professionisti in camera anecoica. Il rumore di fondo è stato realizzato invece utilizzando i segnali del CD NOISE-ROM-0 prodotto nell'ambito del progetto Europeo SAM nel 1990; in particolare è stato utilizzato un rumore rosa. La potenza di uscita dei segnali e la sensibilità dei microfo-

ni sono stati opportunamente calibrati tramite riproduzione di una sequenza binaria pseudo-random.

Dato il set-up sperimentale descritto, abbiamo considerato due possibili scenari, uno stazionario e uno non stazionario. Nello *scenario stazionario*, tutte le sorgenti assumono la stessa posizione per l'intera durata dell'esperimento. La sorgente desiderata è posta di fronte all'interfaccia microfonica a 1 m circa dal centro della schiera. Le due sorgenti interferenti invece sono posizionate rispettivamente a circa 1,9 m e 2,8 m dal centro della schiera microfonica: entrambi sono parlatori di sesso maschile posizionati il primo alla destra e il secondo alla sinistra della sorgente desiderata, come è possibile osservare in Fig. 5 (a).

Nello *scenario non stazionario*, invece, la sorgente desiderata e la sorgente rumorosa di fondo non subiscono variazioni rispetto al caso precedente, mentre ciò che cambia è la posizione delle due sorgenti interferenti. Entrambe le sorgenti infatti assumono tre posizioni diverse durante l'intero esperimento, com'è possibile notare in Fig. 5. In particolare, nei primi 5 secondi le posizioni dei parlatori sono le stesse dello scenario stazionario; al secondo 5 avviene il primo cambiamento: la sorgente interferente #1 passa dalla posizione #1 alla posizione #2, mentre la sorgente #2 rimane ferma nella posizione #1. Al secondo 10 invece,

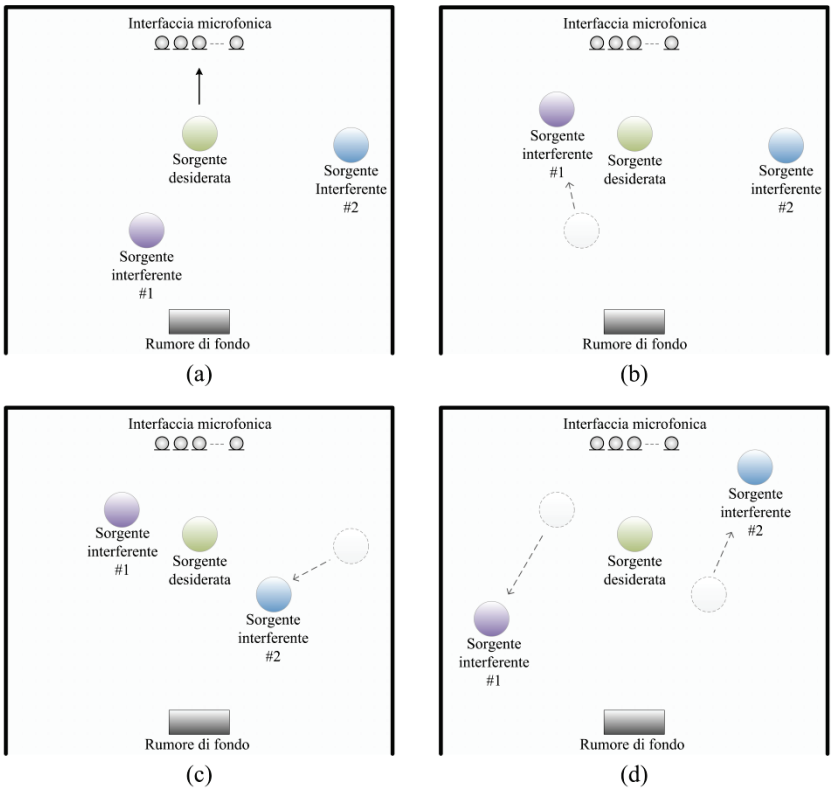


Figura 5. Scenario non stazionario. (a) 0-5 secondi: posizionamento iniziale, coincidente con lo scenario stazionario; (b) 5-10 secondi: spostamento della sorgente interferente #1; (c) 10-15 secondi: spostamento della sorgente interferente #2; (d) 15-20 secondi:spostamento simultaneo di entrambe le sorgenti interferenti.

la sorgente #2 si sposta nella posizione #2 mentre la sorgente #1 rimane ferma, e, infine, al secondo 15 entrambe le sorgenti si spostano contemporaneamente nella corrispettiva posizione #3.

In entrambi gli scenari la durata complessiva dell'esperimento è di 20 secondi.

6.2. Valutazione qualitativa del sistema proposto

Il miglioramento qualitativo del segnale vocale elaborato e la riduzione del rumore apportata da un sistema di beamforming vengono solitamente associate al miglioramento del *rapporto segnale-rumore* (SNR, *signal to noise ratio*), definito come (Brandstein & Ward, 2001; Uncini, 2010):

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\mathrm{E}\left\{r_{\mathrm{IN}}^2[n]\right\}}{\mathrm{E}\left\{r_{\mathrm{OUT}}^2[n]\right\} - \mathrm{E}\left\{r_{\mathrm{IN}}^2[n]\right\}}\right) \tag{13}$$

dove $r_{\mathrm{IN}}[n]$ è il generico segnale di ingresso del sistema ed $r_{\mathrm{OUT}}[n]$ rappresenta il segnale elaborato. L'operatore $\mathrm{E}\{\cdot\}$ indica il valore atteso. Nel caso in cui volessimo misurare il rapporto segnale-rumore di ingresso al nostro sistema, $\mathrm{SNR}_{\mathrm{IN}}$, il segnale $r_{\mathrm{IN}}[n]$ sarebbe nient'altro che il segnale desiderato $s[n]$ emesso dal parlatore, mentre il segnale $r_{\mathrm{OUT}}[n]$ sarebbe il segnale $u_i[n]$ acquisito dall'i-esimo microfono. Analogamente, per ottenere un valore SNR descrittivo dell'uscita del beamformer, $\mathrm{SNR}_{\mathrm{OUT}}$, il segnale acquisito dall'interfaccia diventerebbe $r_{\mathrm{IN}}[n]$ mentre il segnale $r_{\mathrm{OUT}}[n]$ rappresenterebbe il segnale $e[n]$ in uscita dal beamformer. Utilizzando le misure di SNR di ingresso e di uscita è possibile definire il *guadagno d'array* o *direttività* (Uncini, 2010), come il miglioramento del rapporto segnale-rumore tra l'ingresso e l'uscita del beamformer:

$$\mathrm{G} = \frac{\mathrm{SNR}_{\mathrm{OUT}}}{\mathrm{SNR}_{\mathrm{IN}}} \tag{14}$$

Il guadagno d'array dunque è il parametro prestazionale utilizzato per valutare la quantità di miglioramento qualitativo in dB ottenuta attraverso l'elaborazione dei segnali vocali effettuata dal sistema di beamforming.

La valutazione del sistema proposto viene fatta per entrambi gli scenari descritti nel precedente paragrafo e i risultati sono illustrati in Fig. 6. In particolare, abbiamo confrontato il sistema GSC proposto, caratterizzato da un cancellatore adattativo combinato di rumore i cui sistemi MISO sono differenziati in base all'ordine di proiezione, con due sistemi di beamforming GSC convenzionali ciascuno dei quali incorpora uno dei due sistemi MISO utilizzati nel GSC combinato. Dai risultati è possibile notare, come prevedibile, che nel caso stazionario (vedi Fig. 6 (a)) il miglioramento qualitativo apportato dal GSC combinato è lieve perché i sistemi GSC convenzionali considerati, e in particolare quello con ordine di proiezione unitario, risultano essere adeguati a quel tipo di scenario in cui la configurazione delle sorgenti non cambia nel tempo. Tuttavia, come è possibile osservare in Fig. 6 (b), i sistemi GSC convenzionali subiscono le non stazionarietà delle sorgenti rumorose nello secondo scenario e le loro prestazioni risultano degradate soprattutto a regime. In questo caso è possibile osservare che invece il sistema GSC combinato sfrutta le potenzialità dello schema di combinazione dei sistemi MISO adattativi e ottiene un netto miglioramento rispetto ai sistemi convenzionali, risultando così robusto alle non stazionarietà. Dai risultati ottenuti, dunque, si evince che il sistema GSC combinato risulta efficace sia in scenari stazionari semplici che in condizioni ambientali più avverse.

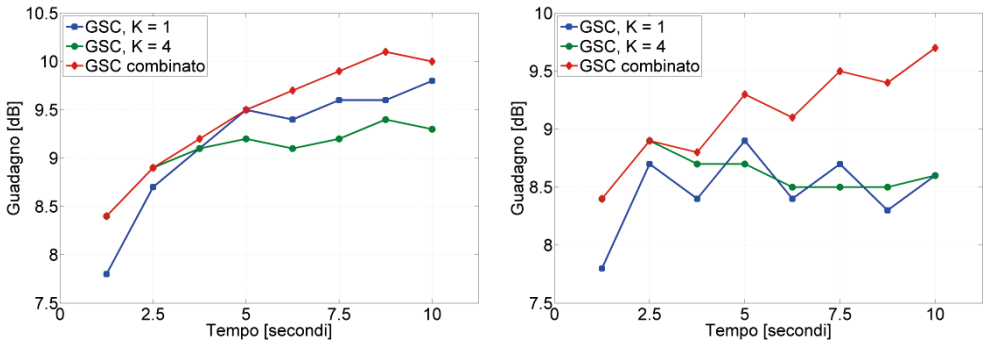


Figura 6. Valutazione del sistema combinato proposto nel caso di scenario stazionario (a) e nel caso di scenario non stazionario (b).

Ad ogni modo, c'è da considerare il fatto che i risultati ottenuti vogliono evidenziare non un miglioramento assoluto del sistema utilizzato ma il miglioramento apportato dall'architettura combinata. È evidente, infatti, che risultati migliori, in termini di riduzione del rumore, si possono ottenere utilizzando interfacce acustiche intelligenti che hanno la stessa architettura combinata ma che utilizzano ad esempio una configurazione microfonica diversa, algoritmi di filtraggio adattativo più performanti, un VAD (*voice activity detector*), o dei *post-filter*, a seconda di quelle che sono le esigenze di scenario.

7. CONCLUSIONI

In questo lavoro è stato presentata una interfaccia acustica intelligente in grado di ridurre il rumore interferente nelle comunicazioni immersive che avvengono in condizioni ambientali non stazionarie. In particolare, l'interfaccia acustica intelligente proposta si basa su uno schema di filtraggio adattativo combinato che consente all'interfaccia di adeguarsi a quelle che sono le condizioni dell'ambiente di comunicazioni e garantendo all'utente una certa qualità della comunicazione vocale. La bontà del sistema di beamforming è stata valutata in due scenari diversi, uno in cui le sorgenti sono tutte stazionarie, e un altro in cui le sorgenti rumorose sono in movimento causando una degradazione della qualità di acquisizione del contributo vocale desiderato. I risultati hanno evidenziato come l'interfaccia acustica intelligente proposta risulti robusta alle condizioni avverse dell'ambiente, consentendo di mantenere elevato il livello qualitativo della conversazione vocale. Inoltre, tale architettura apre la strada a nuove interfacce acustiche intelligenti consentendo così di ottenere un ulteriore miglioramento del parlato nelle comunicazioni immersive.

BIBLIOGRAFIA

Arenas-García, J., Figueiras-Vidal, A.R. & Sayed, A.H. (2006), Mean-square performance of a convex combination of two adaptive filters, IEEE Transactions on Signal Processing, Vol. 54, no. 3, 1078-1090.

Azpicueta-Ruiz, L.A., Figueiras-Vidal, A.R. & Arenas-García, J. (2008), A normalized adaptation scheme for the convex combination of two adaptive filters, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Las Vegas, NV, March 30-April 4, 3301-3304.

Benesty, J., Chen, J. & Huang, Y. (2010), Microphone array signal processing, Berlin, Heidelberg: Springer Verlag.

Blauert, J.P. (1997), Spatial hearing: the psychophysics of human sound localization, Cambridge, MA: MIT Press, revised ed.

Brandstein, M. & Ward, D. (Eds.) (2001), Microphone arrays: signal processing techniques and applications. New York, NY:Springer.

Comminiello, D. (2011), Adaptive algorithms for intelligent acoustic interfaces, PhD Thesis, Sapienza Univ. of Rome, Italy.

Comminiello, D., Scarpiniti, M., Parisi, R. & Uncini, A. (2010), A novel affine projection algorithm for superdirective microphone array beamfomring, in Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS '10), Paris, France, May 30-June 2, 2127-2130.

Comminiello, D., Uncini, A., Cirillo, A., Barone, A. & Falcone, M. (2011), Un sistema a costo minimo per il miglioramento qualitativo del parlato, in Atti del VII Convegno Nazionale AISV, Contesto comunicativo e variabilità nella produzione e percezione della lingua, Lecce, 26-28 Gennaio 2011, (Gili Fivela, B., Stella, A., Garrapa, L., Grimaldi, M. eds.) Roma: Bulzoni Editore, 345-356.

Flanagan, J.L., Johnson, R., Zahn, J.D. & Elko, G.W. (1985), Computer-steered microphone arrays for sound transduction in large rooms, Journal of the Acoustical Society of America, Vol. 75, November 1985, 1508-1518.

Gannot, S., Burshtein, D. & Weinstein, E. (2001), Signal enhancement using beamforming and nonstationarity with applications to speech, IEEE Transactions on Signal Processing, Vol. 49, no. 8, 1614-1626.

Gannot, S. & Cohen, I. (2004), Speech enhancement based on the general transfer function GSC and postfilteing, IEEE Transactions on Speech and Audio Processing, Vol. 12, no. 6, 561-571.

Griffiths, L. & Jim, C. (1982), An alternative approach to linearly constrained adaptive beamforming, IEEE Transactions on Antennas and Propagation, Vol. 30, no. 1, 27-34.

Harris, R.W., Chabries, D.M. & Bishop, F.A. (1986), A variable step size adaptive filter algorithm, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 34, 309-316.

Hefley, W.E. & Murray, D. (1993), Intelligent user interfaces, in Proceedings of the 1st International Conference on Intelligent User Interfaces (IUI '93), New York, NY, 3-10.

Hewett, T.T., Baeker, R., Card, S., Carey, T., Gase, J., Mantei, M., Perlman, G., Strong, G. & Verplank, W. (1992), ACM SIGCHI curricula for human-computer interaction, The Association for Computing Machinery, Inc., New York, NY.

Huang, Y., Benesty, J. & Chen, J. (2006), Acoustic MIMO signal processing, Berlin:Springer-Verlag.

Huang, Y., Chen, J. & Benesty, J. (2011), Immersive audio schemes, IEEE Signal Processing Magazine, Vol. 28, n. 1, January 2011, 20-32.

Li, J. & Stoica, P. (2006), Robust adaptive beamforming, Hoboken, NJ:John Wiley & Sons, Inc.

Lotter, T. & Vary, P. (2006), Dual-channel speech enhancement by superdirective beamforming, EURASIP Journal on Applied Signal Processing, Vol. 2006, no. 1, 1-14.

Martinez-Ramón, M., Arenas-García, J., Navia-Vázquez, A. & Figueiras-Vidal, A.R. (2002), An adaptive combination of adaptive filters for plant identification, in Proceedings of the IEEE International Conference on Digital Signal Processing (DSP ’02), Santorini, Greece, 1195-1198.

Nascimento, V.H., Silva, M.T.M., Azpicueta-Ruiz, L.A. & Arenas-García, J. (2010), On the tracking performance of combination of least mean squares and recursive least squares adaptive filters, in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’10), Dallas, TX, 3710-3713.

Ozeki, K. & Umeda, T. (1984), An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties, Electronics & Communications in Japan, Vol. 67-A, 19-27.

Paleologu, C., Ciochina, S. & Benesty, J. (2008), Variable step-size NLMS algorithm for under-modeling acoustic echo cancellation, IEEE Signal Processing Letters, Vol. 15, 5-8.

Sayed, A.H. (2008), Adaptive filters, Hoboken, NJ: John Wiley & Sons, Inc.

Shin, H.C., Sayed, A.H. & Song, W.-J. (2004), Variable step-size NLMS and affine projection algorithms, IEEE Signal Processing Letters, Vol. 11, no. 2, February, 132-135.

Silva, M.T.M. & Nascimento, V.H. (2008), Improving the tracking capability of adaptive filters via convex combination, IEEE Transactions on Signal Processing, Vol. 56, no. 7, 3137-3149.

Stern, R.M., Gouvea, E., Chanwoo, K., Kumar, K. & Park, H.-M. (2008), Binaural and multiple-microphone signal processing motivated by auditory perception, in Proceedings of Hands-Free Speech Communication and Microphone Arrays (HSCMA 2008), Trento, Italy, May 6-8, 98-103.

Uncini, A. (2010), Elaborazione adattativa dei segnali, Roma: Aracne Editrice S.R.L.

Zheng, Y.R. & Goubran, R.A. (2000), Adaptive beamforming using affine projection algorithms, in Proceedings of the 5th International Conference on Signal Processing (WCCC-ICSP '00), Beijing, China, August 21-25, 1929-1932.

FONETICA II

LA PROMINENZA IN ITALIANO: DEMARCAZIONE PIU' CHE CULMINAZIONE

Gloria Gagliardi
Università di Firenze
gloria.gagliardi@unifi.it

Edoardo Lombardi Vallauri
Università Roma Tre
lombardi@uniroma3.it

Fabio Tamburini
Università di Bologna
fabio.tamburini@unibo.it

1. INTRODUZIONE

L'espressione della Struttura Informativa (SI) negli enunciati linguistici è in gran parte affidata a pattern acustici. Il livello più strettamente associato ad essi è quello a cui prevalentemente ci si riferisce con i termini di "Tema-Rema", "Topic-Focus" o "Topic-Comment", per i quali adottiamo le definizioni proposte da Cresti (1992; 2000) e Lombardi Vallauri (2001; 2009), fondate su quali parti dell'enunciato ne veicolano la forza illocutiva. In questo quadro, il Focus è *la parte dell'enunciato che veicola la forza illocutiva e realizza lo scopo informativo dell'enunciato stesso*. Il Topic invece è *la parte dell'enunciato che non ha forza illocutiva e la cui funzione è di permettere la comprensione del Focus all'interno del discorso*.

Queste definizioni coincidono nella sostanza con i concetti di Topic e Focus (Tema-Rema, Topic-Comment) adoperati da molta della letteratura sui correlati acustici della SI (Halliday, 1989; Ladd, 1978; 1996; Pierrehumbert, 1987; Selkirk, 1984 e, più vicino alla nostra analisi, Avesani, 2000; Avesani, Vayra, 2004; Avesani, *et al*. 2007; Breen, *et al*. 2010; D'Imperio, 2002b; Féry, Krifka, 2008; Frascarelli, 2000; 2004; Frascarelli, Hinterhölzl, 2007, ecc.).

Nel presente studio si sono esaminati enunciati appartenenti a due corpora di italiano parlato, individuando le categorie di Topic e di Focus secondo due principali criteri:

- Primo, la valutazione (in base alla percezione dei pattern acustici e all'applicazione di test di negazione) di quali parti dell'enunciato veicolano forza illocutiva e sono quindi responsabili dell'atto linguistico che viene compiuto; cioè del fatto che l'enunciato sia un'asserzione, una domanda, una richiesta, un comando o qualsiasi altro atto pragmaticamente rilevante (v. Cresti, 2000 per una lista di oltre 80 atti illocutivi).

- Secondo, l'esame del contesto precedente per stabilire quale informazione sia da considerare attiva (Chafe, 1987; 1992) nel momento in cui viene prodotto l'enunciato, cioè Data, e perciò meno probabilmente in Focus; e quale informazione si possa considerare inattiva, cioè Nuova, e quindi più probabilmente in Focus.

Si sono presi in esame solo i tre tipi di SI più frequenti negli enunciati dei corpora, cioè Focus Esteso (a tutto l'enunciato), Topic-Focus e Focus-Appendice[1] (cioè costruzioni con un Focus Ristretto a sinistra dell'enunciato).

Alcuni lavori sull'argomento studiano direttamente le relazioni fra la SI e i fenomeni fonetici, mentre altri introducono un livello intermedio di natura fonologica (ad es. Ladd, 1996; Pierrehumbert, 1987 e tutti gli studi che adottano il sistema di trascrizione ToBI (Beckman, *et al*. 2005). In questa seconda prospettiva le categorie fonologiche sono fatte derivare dai parametri acustici considerando soprattutto l'intonazione, cioè i profili di F0.

[1] Con il termine "Appendice" intendiamo un Topic collocato a Destra del Focus.

La maggior parte degli studi sui correlati prosodici della SI (in particolare del Focus) per l'italiano sono stati condotti entro il paradigma della fonologia Autosegmentale Metrica (AM). Al momento attuale il panorama delle varietà diatopiche esaminate non può dirsi completo, e la quasi totalità dei lavori ha riguardato solo il parlato letto (non il parlato spontaneo o semi-spontaneo). La Tabella 1 sintetizza gli andamenti tonali degli enunciati assertivi individuati dagli studiosi per le varietà che questo contributo condivide; considerando il margine di variabilità riscontrato nei comportamenti dei locutori nella realizzazione degli accenti (Gili Fivela, 2006; Avesani e Vayra, 2004) tali pitch-accent sono da intendersi solo come pattern tipici.

	Focus Esteso	Focus Ristretto	Focus Contrastivo
Roma (Frascarelli, 2004)	H+L*	H*, H*+L	H*, H*+L
Firenze (Avesani, Vayra, 2004)	H+L*	H+L*	L+H*, (L+H)* H+H*
Napoli (D'Imperio, 2002b)	H+L*	L+H*	L+H*

Tabella 1: Profili intonativi tipici di enunciati assertivi nell'ambito degli studi AM.

Come mostra la tabella, l'accento nucleare varia sia in dipendenza della portata del Focus, sia secondo la dimensione diatopica: in particolare, mentre per la varietà Fiorentina ad essere marcata intonativamente è la contrastività, per le varietà Romana e Napoletana è la portata del focus ad essere associata a pitch-accent diversi. Non è ancora chiaro se tali difformità siano interamente imputabili alla variazione diatopica oppure siano legate alle caratteristiche di trascrizione di ToBI. La notazione, da un lato, sembra non riuscire a rendere conto di differenze melodiche chiaramente percepite dai parlanti: ad esempio, nonostante i locutori siano in grado di identificare la provenienza geografica di un parlante solo sulla base dell'intonazione, il broad focus delle assertive è rappresentato mediante lo stesso pitch-accent (Marotta, 2008). Dall'altro lato, sembra esistere un problema di agreement nella descrizione degli accenti. Se l'accordo può dirsi infatti consistente nell'identificazione degli edge tone e dei pitch-accent, è basso nella classificazione dei pitch-accent (Pitrelli et al., 1994; Syrdal & McGorg, 2000). A questo proposito risultano spesso problematici non solo l'allineamento (D'Imperio, 2002a; Gili Fivela, 2002), ma anche l'identificazione dei target tonali, in particolar modo nei plateau, in cui un unico massimo o minimo non possono essere agevolmente identificati (D'Imperio, 2002a). Le informazioni su *scaling* e *slope* sono sottostimate, sebbene potenzialmente distintive (Gili Fivela 2002).

Come suggerito in alcuni studi classici (ad esempio Ladd,1996) e confermato in ricerche più recenti (Breen, et al. 2010; Lee, Yu, 2010), un item in focus potrebbe coinvolgere una combinazione complessa di tratti acustici differenti, vale a dire durata, pitch e intensità, e non può pertanto essere analizzato solo mediante il profilo intonativo. Per queste ragioni proveremo a indagare la correlazione tra elementi focalizzati e caratteristiche fonetiche considerando il concetto di prominenza prosodica come un insieme complesso di tratti acustici, combinati in modo articolato. L'identificazione automatica dei livelli di prominenza è indubbiamente un task complesso che richiede un'analisi più approfondita.

2. IDENTIFICAZIONE AUTOMATICA DELLA PROMINENZA

Facendo riferimento agli studi di (Couper-Kuhlen, 1986; Jensen, 2004; Kohler, 2006; Mertens, 1991; Terken, 1991), è possibile definire il concetto di prominenza prosodica come *un fenomeno percettivo, di natura continua, che consente di enfatizzare alcune unità linguistiche di tipo segmentale rispetto al contesto che le circonda, ed è supportato da una complessa interazione di parametri di tipo prosodico e fonetico/acustico*.

Dei numerosissimi lavori in questo settore sembra opportuno fare riferimento primariamente al lavoro di (Kohler, 2005), per la chiarezza, la lucidità e il rigore metodologico con cui descrive i fenomeni coinvolti. Dai lavori di Kohler emergono chiaramente due attori precisi, a livello linguistico-prosodico, in grado di supportare il fenomeno della prominenza frasale (o *sentence accent*): i *pitch accent* e i *force accent*. Il primo (*pitch accent*) risulta coincidere pressoché totalmente col concetto omonimo introdotto da Bolinger (1958) ed essenzialmente legato a variazioni, o a specifiche configurazioni, nel profilo della frequenza fondamentale (F0), mentre il secondo (*force accent*) risulta essere un fenomeno completamente indipendente dalla componente intonativa degli enunciati e intimamente legato a fenomeni acustici di altro tipo, per esempio l'intensità e la durata delle unità segmentali.

I due fenomeni sembrano giocare entrambi un ruolo preminente nel supportare la prominenza percepita a livello di enunciato, in linea con ciò che sostengono alcuni studiosi (si veda ad esempio il lavoro di Ladd, 1996), ma anziché in un'ottica antagonistica o gerarchica in un'ottica di interazione e rinforzo reciproco.

Una delle sfide più rilevanti nell'identificazione del livello di prominenza sillabica riguarda la determinazione dell'influenza che i vari parametri esercitano sulla percezione della prominenza, in particolare le escursioni della frequenza fondamentale F0, la durata delle unità sillabiche, misure di intensità e anche le aspettative dell'ascoltatore. A livello acustico, numerosi studi (Sluijter, van Heuven, 1996; 1997; Anastakos et al. 1995; Bagshaw, 1994; Heldner, 2003; Streefkerk, 1996) suggeriscono, anche in una prospettiva interlinguistica, una dipendenza tra i *force accent* e parametri come la durata e l'enfasi spettrale (*spectral emphasis*, *spectral tilt* o *spectral balance*), mentre i *pitch accent* sarebbero supportati prevalentemente da movimenti o specifiche configurazioni nel profilo di F0 e dall'intensità globale all'interno dell'unità segmentale di riferimento. Uno degli autori ha condotto alcuni esperimenti che hanno suffragato l'esistenza di tali relazioni in riferimento ad alcune lingue (Tamburini, 2003; 2005; 2006).

Queste considerazioni puramente qualitative si possono trasformare in legami quantitativi definendo una funzione che sia in grado di assegnare livelli continui di prominenza ai nuclei sillabici utilizzando unicamente informazioni di tipo acustico:

$$\begin{aligned} Prom^i = \quad & W_{FA} \cdot \left[SpEmph^i_{SPLH-SPL} \cdot dur^i \right] + \\ & W_{PA} \cdot \left[en^i_{ov} \cdot \left(A^i_{event}(at_M, at_m) \cdot D^i_{event}(at_M, at_m) \right) \right] \end{aligned}$$

dove $SpEmph_{\text{SPLH-SPL}}$ riguarda una misura di enfasi spettrale, *dur* è la durata del nucleo sillabico, en_{ov} è l'energia globale del nucleo e A_{event}, D_{event} sono parametri derivati dal modello intonativo TILT (Taylor, 2000) calcolati in funzione della tipologia di allineamento scelto per i massimi – at_{M} – e i minimi – at_{m} – presenti nel profilo. Tutti questi parametri sono riferiti al generico nucleo sillabico *i* all'interno dell'enunciato. Si veda la Tabella 2 per alcuni dettagli sul calcolo di questi parametri.

La struttura della funzione *Prom*, sebbene sembri scelta arbitrariamente, riflette in realtà le relazioni tra i parametri che abbiamo descritto e, in particolare, la somma dei due contri-

buti esprime matematicamente la visione di rinforzo reciproco che attribuiamo alle due tipologie accentuali considerate.

Parametro	**Descrizione**
Durata del Nucleo (*dur*)	Durata temporale del nucleo sillabico normalizzata considerando la media e la varianza delle durate dei nuclei all'interno dell'enunciato (*z-score*), e calcolata, all'interno di questo lavoro, utilizzando le segmentazioni manuali disponibili per i corpora considerati.
Enfasi Spettrale ($SpEmph_{SPLH\text{-}SPL}$)	Parametro SPLH-SPL (Fant, *et al*. 2000), normalizzato come nel caso precedente (*z-score*).
Configurazioni del Pitch	Rappresentazione del profilo intonativo dell'enunciato utilizzando il modello TILT (Taylor, 2000) a partire da un profilo del pitch calcolato utilizzando il programma *ESPS get_f0* (Talkin, 1995).
Intensità Globale (en_{ov})	Energia RMS calcolata nella banda di frequenza 50-5000 Hz, normalizzata anch'essa considerando la media e la varianza delle intensità all'interno dell'enunciato (*z-score*).

Tabella 2: Parametri acustici utilizzati nell'algoritmo automatico per l'identificazione della prominenza prosodica.

Il nucleo della funzione *Prom* contiene nove parametri. Cinque di essi possono essere visti come parametri acustici in grado di supportare il fenomeno della prominenza da un punto di vista cross-linguistico ($SpEmph_{SPLH\text{-}SPL}$, *dur*, en_{ov} , A_{event} e D_{event}), mentre gli altri quattro, rappresentati dal vettore $\mathbf{W} = (W_{FA}, W_{PA}, at_M, at_m)$, possono essere visti come specifici di una determinata lingua. Nel nostro modello, W_{FA} e W_{PA} pesano il contributo delle due tipologie di accenti, mentre at_M e at_m modellizzano le differenti possibilità di allineamento tra i *pitch accents* e i nuclei sillabici nelle varie lingue (si veda la Figura 1).
Tutti i parametri coinvolti nel calcolo della funzione *Prom* sono normalizzati all'interno dell'enunciato, quindi i contributi dei differenti locutori e dei differenti intervalli numerici dovrebbero essere stati fattorizzati. In tutti gli esperimenti che presenteremo è stato utilizzato l'insieme di parametri $\mathbf{W}$ = (1.0, 1.0, 2, 2), che risulta essere il più opportuno per l'italiano (Tamburini, 2009).

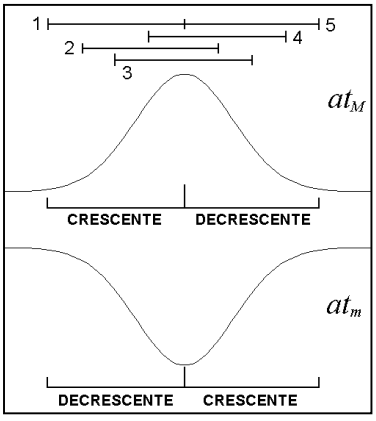


Figura 1: Parametri per l'allineamento tra i pitch accents e i nuclei sillabici.

3. ESPERIMENTI

I due esperimenti che presenteremo sono volti all'identificazione di regolarità tra la posizione e il livello della Prominenza Principale, identificata utilizzando l'algoritmo automatico presentato nella sezione precedente, e la struttura informativa dell'enunciato così come è stata classificata da un annotatore esperto.

Il primo esperimento presenta uno studio pilota su un corpus piuttosto limitato dell'italiano parlato a Roma. Il secondo esperimento ha lo scopo di verificare i risultati ottenuti nel primo sulla stessa varietà, ma considerando un corpus differente, e di estendere l'analisi a due ulteriori varietà, l'italiano parlato a Firenze e a Napoli. L'annotatore ha identificato manualmente l'unità di Focus e le altre unità, se presenti, di Topic e Appendice. Ha inoltre determinato l'estensione del Focus e la sua possibile contrastività.

Negli esperimenti considereremo solo tre tipologie di enunciati che possono essere classificati: (a) TOPIC | FOCUS; (b) FOCUS ESTESO; (c) FOCUS | APPENDICE, FOCUS RISTRETTO, FOCUS CONTRASTIVO. Gli enunciati contenenti riprogrammazioni, esitazioni o disfluenze sono stati esclusi dallo studio, almeno in questa prima fase.

(a) TOPIC | FOCUS							
Varietà-Corpus	*Prominenza Principale sulla…*						*Nessuna Prom.*
	UsT	*UsF*	*UsA*	*sIT*	*sIF*	*sIA*	*Principale*
Roma–Bonvino	18	1	-	0	1	-	3
Roma–CLIPS	12	3	-	1	0	-	3
Firenze–CLIPS	24	1	-	0	1	-	7
Napoli–CLIPS	8	0	-	2	1	-	2
(b) FOCUS ESTESO							
Varietà-Corpus	*Prominenza Principale sulla…*						*Nessuna Prom.*
	UsT	*UsF*	*UsA*	*sIT*	*sIF*	*sIA*	*Principale*
Roma–Bonvino	-	4	-	-	0	-	4
Roma–CLIPS	-	4	-	-	6	-	8
Firenze–CLIPS	-	3	-	-	3	-	2
Napoli–CLIPS	-	4	-	-	7	-	6
(c) FOCUS | APPENDICE, FOCUS RISTR., FOCUS CONTRAST.							
Varietà-Corpus	*Prominenza Principale sulla…*						*Nessuna Prom.*
	UsT	*UsF*	*UsA*	*sIT*	*sIF*	*sIA*	*Principale*
Roma–Bonvino	-	14	0	-	2	0	0
Roma–CLIPS	-	22	1	-	2	0	2
Firenze–CLIPS	-	14	1	-	1	0	2
Napoli–CLIPS	-	25	0	-	6	0	0

Tabella 3: Numero di enunciati divisi per Varietà-Corpus e configurazioni (es. UsT=Ultima sill. del Topic, sIF=sill. Interna del Focus). Alcune combinazioni non sono ammissibili; in questi casi è stato inserto il simbolo '-' nella casella corrispondente.

3.1 Esperimento 1

I dati sono stati estratti dal corpus "Bonvino", una sezione di *Ar.Co.Dip.* (Bonvino, 2005). Il corpus è formato da 12 conversazioni tra locutori provenienti da Roma, omogenei a livello sociale, età, titolo di studio e origine geografica. Da tre delle dodici conversazioni sono stati selezionati 47 enunciati, estratti gli oscillogrammi, ed è stata prodotta la trascrizione

fonetica allineata al fine di identificare i nuclei sillabici necessari alla procedura automatica per l'identificazione della prominenza.

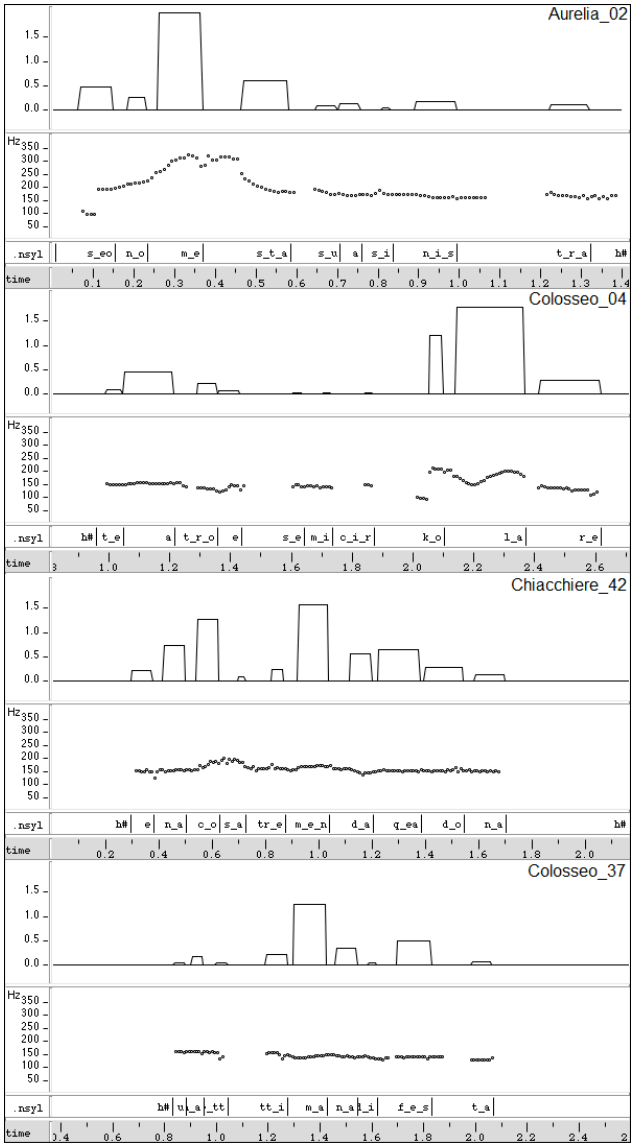


Figura 2: Profili della funzione di prominenza – *Prom* – e profili del pitch per alcuni enunciati considerati in questo studio. Aurelia_02: "*Secondo me* $_{T}$ | *stava sulla sinistra* $_{F}$". Colosseo_04: "*Il teatro è semicircolare* $_{F}$". Chiacchiere_42: "*E' una cosa tremenda* $_{F}$ | *quella donna* $_{A}$". Colosseo_37: "*Una settimana* $_{F}$ | *di festa* $_{A}$".

3.2 Esperimento 2

I dati per il secondo esperimento sono stati estratti dal subcorpus dialogico del corpus CLIPS (in particolare dalle sezioni riguardanti i map-task e i test delle differenze), corpus che risulta stratificato rispetto alla dimensione diatopica e diafasica (Albano Leoni, 2003). La scelta è quindi caduta su testi relativi alla varietà romana, allo scopo di replicare l'esperimento precedente utilizzando dati diversi, e sulle varietà di italiano parlate a Napoli e Firenze, entrambe particolarmente studiate nell'ambito della fonologia Autosegmentale-Metrica. Sono stati selezionati un totale di 184 enunciati: 64 per la varietà romana, 59 per quella fiorentina e 61 per quella napoletana.

I risultati di entrambi gli esperimenti, visualizzati nella Tabella 3, mostrano regolarità notevoli tra la posizione della Prominenza Principale e la Struttura Informativa degli enunciati. Innanzitutto è possibile notare che, considerando le diverse strutture informative, non emergono differenze rilevanti tra le varietà di italiano considerate nello studio: la distribuzione delle Prominenze Principali sembra seguire schemi molto simili in tutte le coppie varietà-corpus. Inoltre, la posizione della Prominenza Principale tende a collocarsi al confine tra le due unità informative per quanto riguarda le strutture TOPIC | FOCUS e FOCUS | APPENDICE, mentre, nel caso di FOCUS ESTESO, il quadro generale sembra essere più complesso anche se è possibile riscontrare una marginale tendenza della Prominenza Principale a collocarsi alla fine dell'enunciato. La Figura 2 mostra queste regolarità in riferimento a tre enunciati, presi come esempi di tali comportamenti: Aurelia_02 (TOPIC | FOCUS), Colosseo_04 (FOCUS ESTESO) and Chiacchiere_42 (FOCUS | APPENDICE), tutti estratti dal corpus Bonvino.

E' importante notare che un numero rilevante di Prominenze Principali tra quelle considerate (per esempio 14 enunciati tra i 47 estratti dal corpus "Bonvino") sono supportate principalmente o unicamente da *force accents*, come mostrato dall'enunciato Colosseo_37 nella Figura 2. In questi casi nessun fenomeno di tipo intonativo (*pitch accent*) ha contribuito a supportare quelle prominenze.

Queste regolarità si sono dimostrate altamente rilevanti anche attraverso un test statistico di Fisher.

4. UN'INTERPRETAZIONE "TOPOLOGICA": DEMARCAZIONE PRIMA CHE CULMINAZIONE

I risultati che abbiamo ottenuto non sono certo assoluti. La convergenza fra percezione e misurazione rivela forti (anche molto forti) tendenze, ma non è mai totale. Risultati del tutto netti, in cui i pattern prosodici associati al Topic e al Focus sono perfettamente coerenti, quando si lavora su corpora di parlato spontaneo si ottengono forse solo con procedure *ex post*, cioè se prima si fa la misurazione e poi sulla base di essa si procede all'etichettatura; cioè, se a tutti gli enunciati che alla misurazione presentano lo stesso pattern si dà la stessa etichetta (ad es. Topic-Focus; o Focus Esteso; ecc.). Vale a dire, solo se si adotta un procedimento circolare. Ma se l'etichettatura si fa prima su base percettiva, poi la misurazione è destinata a produrre sempre qualche sorpresa.

Comunque, dai risultati a cui ha condotto l'esperimento che ora si illustrerà, è stato possibile trarre alcune conclusioni interessanti.

Come si è visto in Tabella 3, l'accostamento fra la valutazione percettiva sugli enunciati del corpus e la loro misurazione automatica mediante l'algoritmo che adottiamo ha portato ai seguenti risultati:

Topic-Focus
- La maggioranza degli enunciati hanno la Prominenza principale all'estremità destra del Topic.
- Una minoranza sembrano non distinguere fra le due unità, su cui cadono Prominenze paragonabili.

Focus Ristretto (a sinistra)
- È sempre marcato dalla Prominenza principale dell'enunciato, all'estremità destra del Focus.

Focus Esteso
- Circa metà degli enunciati hanno la Prominenza principale all'estrema destra.
- L'altra metà non hanno una Prominenza principale, ma varie prominenze equivalenti.

Cioè, sembra che ad essere segnalati stabilmente dalla Prominenza principale siano solo i costituenti che si trovano alla sinistra dell'enunciato (Topic, o Focus Ristretto), e più precisamente l'estremità destra di tali costituenti. Questo ammette la seguente possibile spiegazione: la funzione primaria della Prominenza principale potrebbe essere *demarcativa*, piuttosto che culminativa. In altre parole, il suo primo, immediato effetto potrebbe essere quello di tracciare il confine tra due unità informative, piuttosto che quello di "descrivere" in modo riconoscibile ciascuna di esse.

Questo non significa che diversi tipi di Topic e di Focus non possano essere caratterizzati da diversi e specifici contorni intonativi, che determinino diversi tipi di illocuzione e di funzione pragmatica. Ma la *mera presenza e posizione* della Prominenza Principale (quale risulta dalle misurazioni effettuate) è già sufficiente a segnalare se l'enunciato contiene un confine tra unità informative, e dove esso si trova. E una volta che la Prominenza Principale segnala un confine fra due unità, per riconoscere di quali unità si tratti è sufficiente che il contorno intonativo di quella che si trova a destra segnali se si tratta di un Focus o di un'Appendice.

Gli indizi minimi necessari per rendere riconoscibili da parte del destinatario i confini tra unità unformative sono dunque quelli mostrati nella Tabella 4 (PP = Prominenza Principale):

unità di SI	***inizio segnalato da:***	***fine segnalata da:***
Topic	inizio dell'enunciato/del contorno intonativo	PP sull'ultima sillaba accentata del Topic
Focus a Destra (dopo un Topic)	PP sull'ultima sillaba accentata del Topic	fine dell'enunciato/del contorno intonativo
Focus Esteso	inizio dell'enunciato/del contorno intonativo	fine dell'enunciato/del contorno intonativo
Focus Ristretto (a Sinistra)	inizio dell'enunciato/del contorno intonativo	PP sull'ultima sillaba accentata del Focus, e inizio di un contorno piatto di Appendice
Appendice	PP sull'ultima sillaba accentata del Focus, e inizio di un contorno piatto di Appendice	fine dell'enunciato

Tabella 4: Indizi minimi per il riconoscimento delle unità informative

Questo fornirebbe una spiegazione piuttosto semplice delle questioni seguenti:

- Perché i Topic sono segnalati più energicamente dei Focus Estesi e dei Focus a Destra che seguono un Topic, benché la rilevanza comunicativa dei Focus sia maggiore di quella dei Topic: la ragione può essere che i Topic, a differenza dei Focus Estesi e a Destra, sono seguiti da un'altra unità di informazione all'interno dello stesso enunciato, e perciò il confine fra le due unità deve essere segnalato.
- Perché anche i Focus Ristretti (a Sinistra) sono segnalati energicamente: la ragione è la stessa, e cioè che anche i Focus a Sinistra sono seguiti da un confine tra unità informative entro l'enunciato.

Dunque, per spiegare come la Prominenza Principale consente (almeno in alcune varietà italiane) il riconoscimento delle unità informative, proponiamo di partire da una spiegazione di natura *squisitamente strutturale*, e più precisamente di natura "topologica"; cioè una spiegazione basata solo sulla *presenza e posizione*, non su aspetti qualitativi della Prominenza e dei contorni intonativi:

Ipotesi topologica sulla prominenza principale

"Ciò che è segnalato dalla Prominenza principale è il confine tra unità informative"

In termini essenziali, l'unica differenza *qualitativa* necessaria per il riconoscimento della Struttura Informativa di un enunciato è quella fra la marcatura di un Topic e quella di un Focus (ristretto) a Sinistra, perché entrambi sono seguiti da un'altra unità. Tale differenza può essere assicurata dai diversi contorni intonativi delle unità che seguono (rispettivamente, dopo un Topic si avrà un Focus a Destra, e dopo un Focus a Sinistra si avrà un'Appendice); oppure (anche con qualche ridondanza) da specifici contorni intonativi che caratterizzino rispettivamente il Topic e il Focus a Sinistra.

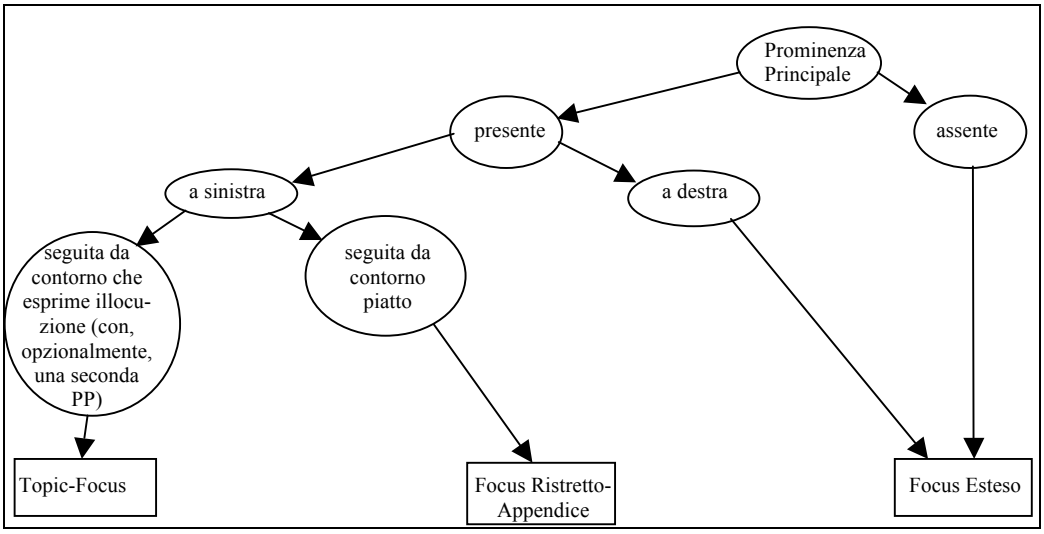
L'assenza di una Prominenza principale, o il suo trovarsi sull'ultima sillaba accentata dell'enunciato, segnalano entrambi un Focus Esteso (non preceduto da un Topic), i cui confini a rigore non richiedono una Prominenza principale che li segnali, poiché coincidono con i confini dell'intero enunciato.

I passi attraverso i quali il destinatario può "computare" la Struttura Informativa di un enunciato orale sono evidenziati nello Schema 1.

In questa interpretazione, i parlanti obbediscono in misura (non) sorprendente alla *Legge del Minimo Sforzo*. Gli unici elementi strettamente necessari sono (a) una Prominenza principale per ogni enunciato, e (b) la differenza tra il contorno "illocutivo" del Focus e il contorno privo di illocuzione di un'Appendice. Ora, poiché i diversi contorni di Focus sono comunque necessari per esprimere le diverse illocuzioni degli enunciati (cioè i diversi atti linguistici), il costo specifico richiesto per segnalare la Struttura Informativa risulta molto basso. Segnalare ciascuna unità informativa con una prominenza culminativa costerebbe più sforzo che segnalare semplicemente i confini, perché:

- distinguere il Topic dal Focus richiederebbe la produzione di due diverse prominenze (una per ciascuna unità) invece di una sola (al confine);
- distinguere il Focus Esteso dal Focus Ristretto richiederebbe due prominenze riconoscibilmente diverse, perché anche i Focus Estesi dovrebbero avere una prominenza "dedicata". Invece, secondo questa interpretazione il linguaggio preferisce funzionare in modo più economico, e cioè marcare solo... l'elemento marcato: cioè il Focus Ristretto.

Questa situazione è ben rappresentata nel corpus, come mostra la Tabella 5. Ma la situazione è più complessa di così, come vedremo nella sezione 4.1.



Schema 1: Passi essenziali per il riconoscimento delle unità di Struttura Informativa.

	Enunciati che corrispondono alla descrizione	Enunciati che non corrispondono alla descrizione
Roma – Bonvino	40 (85.10%)	7 (14.90%)
Roma – Clips	46 (71.88%)	18 (28.12%)
Firenze – Clips	42 (71.19%)	17 (29.81%)
Napoli – Clips	43 (70.49%)	18 (29.50%)
TOTALE	170 (73.59%)	61 (26.41%)

Tabella 5: Risultati previsti e non previsti per le realizzazioni acustiche della SI nel corpus.

4.1. Non alternative discrete, ma un continuum

Come si vede nella Tabella 3, nel corpus una minoranza degli enunciati che sono percepiti come Topic-Focus non hanno una Prominenza principale; e una minoranza di quelli valutati come Focus Estesi hanno una Prominenza principale in posizione interna, non diversa da strutture Topic-Focus.

In altre parole, enunciati acusticamente misurabili come Focus Estesi possono essere percepiti come Topic-Focus, e viceversa. Questo si può spiegare: Topic-Focus e Focus Esteso non sono strutture separate e reciprocamente esclusive, ma piuttosto i due estremi di un *continuum*. Il centro del continuum è occupato da quegli enunciati in cui le unità informative non sono segnalate in maniera netta, e la distinzione fra le due possibili SI rimane sottospecificata o addirittura non specificata.

Insomma, *il parlante non è obbligato a decidere* tra Topic-Focus e Focus Esteso. Alme-

no, non prosodicamente; poiché la disambiguazione può sempre essere affidata a fattori contestuali e pragmatici. Questo è ancora più vero se si considera che il locutore e il destinatario possono valutare in maniera diversa gli indizi prosodici, e il locutore è sempre consapevole di tale possibilità. Pertanto, si regola sempre sulla previsione che la percezione della SI può essere soggetta a un certo grado di indeterminatezza.

Più radicalmente, non c'è ragione di pensare che un contenuto debba necessariamente essere focalizzato o al 100% o allo 0%. Al contrario, ogni contenuto può essere focalizzato in una varietà illimitata di gradi (Daneš, 1967, 1974; Firbas, 1966, 1987, 1989; Sgall 1975; Sgall *et al.* 1973), o anche in una misura che può semplicemente rimanere sottospecificata.

Perciò non c'è da stupirsi se la Prominenza principale non è sempre chiaramente riconoscibile. E' buona norma aspettarsi che alcuni enunciati abbiano uno statuto intermedio tra Topic-Focus e Focus Esteso. E lo statuto informativo di alcune porzioni di informazione, tipicamente "a metà" dell'enunciato, può rimanere incerto.

Insomma, Topic versus Focus non si presenta come un'opposizione di bianco e di nero, ma come una scala di toni di grigio. Questo è ciò che si osserva in enunciati come quelli in Figura 3.

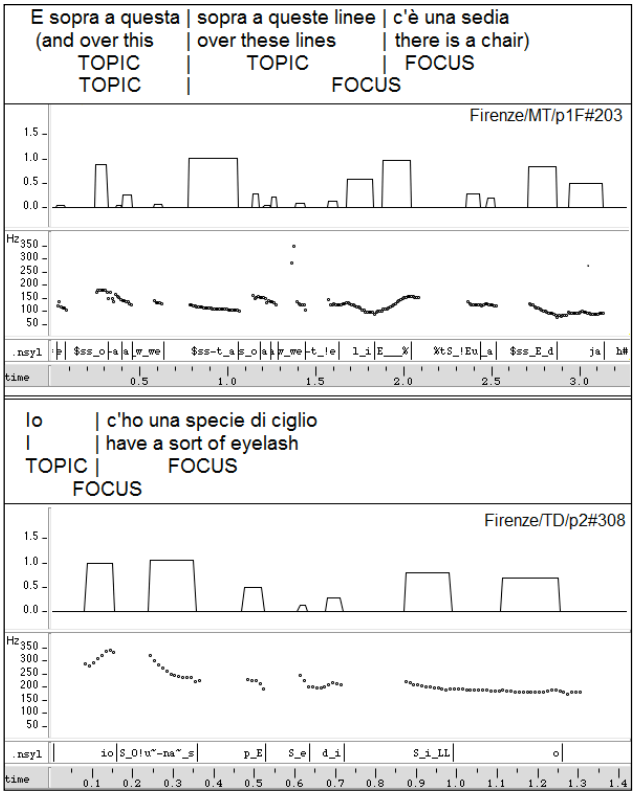


Figura 3: Enunciati sottospecificati per la distinzione fra Topic-Focus e Focus Esteso.

L'assenza di una distinzione netta fra Topic-Focus e Focus Esteso corrisponde al fatto che si tratta di strutture spesso possibili negli stessi contesti, e che spesso non influenzano il seguito del discorso in modo significativamente diverso. Inoltre, si può fare un'osservazione più generale: il fatto che le categorie della SI rimangano sottospecificate negli scambi comunicativi reali non è un problema per gli esiti della comunicazione, come non lo è il fatto che questo ovviamente accada per molti altri aspetti dell'interpretazione semantico-pragmatica degli enunciati.

Per esempio, se dico "Tito ha fermato la macchina", il destinatario può produrre ogni tipo di arricchimento libero per interpretare il mio enunciato, arrivando a interpretazioni anche molto diverse: che Tito è il conducente della macchina, oppure un vigile che ha intimato l'Alt, o un elefante che gli ha attraversato la strada davanti, e così via. Anche informazione di natura meno pragmatica può restare sottospecificata o del tutto inespressa. Per esempio, in molte lingue il tempo verbale può restare non espresso, permettendo diverse interpretazioni (spesso non del tutto disambiguate dal contesto) delle coordinate temporali dell'evento descritto dall'enunciato. Ancora più comunemente, l'identità dei partecipanti a un evento può mancare di espressione esplicita nelle lingue in cui il Soggetto esplicito non è la regola e il Verbo non ha marche morfologiche per la persona. L'esempio giapponese che segue contiene entrambe le ambiguità di cui abbiamo appena parlato.

Tokyo-e ikimasu
Tokyo-a andare
"Io/tu/lui/lei/noi/voi/loro vado/vai/andrò/andremo ecc. a Tokyo"

Ebbene, se nel nostro corpus consideriamo come coerenti con il modello anche tutti i casi in cui la SI rimane sottospecificata tra Topic-Focus e Focus Esteso, otteniamo le nuove percentuali mostrate in Tabella 6:

	Enunciati che corrispondono alla descrizione	Enunciati che non corrispondono alla descrizione
Roma – Bonvino	43 (91.49%)	4 (8.51%)
Roma – Clips	55 (85.94%)	9 (14.06%)
Firenze – Clips	53 (89.83%)	6 (10.17%)
Napoli – Clips	53 (86.89%)	8 (13.11%)
TOTALE	170 (87.88%)	28 (12.12%)

Tabella 6: Risultati previsti e non previsti per le realizzazioni acustiche della SI nel corpus (incluso il continuum tra Topic-Focus e Focus Esteso).

Questo significa che quasi il 90% degli enunciati presentano uno dei seguenti tipi di convergenza fra valutazione percettiva e risultati della misurazione:

- strutture valutate come Topic-Focus, con la Prominenza Principale all'estremità destra del Topic;
- strutture valutate come Focus-Appendice, con la Prominenza Principale all'estremità destra del Focus;

- strutture valutate come Focus Estesi, o senza Prominenza Principale, o con la Prominenza Principale all'estremità destra;
- strutture valutate o come Topic-Focus o come Focus Esteso, senza una Prominenza Principale evidente.

Solo nel 10% dei casi, le misurazioni automatiche danno risultati in cui la Prominenza principale è in posizioni diverse da quelle previste. Questi possono probabilmente considerarsi "rumore" residuo nella procedura: l'esistenza di una minoranza di casi con pattern diversi è da attendersi, perché (a) ragionevolmente devono esserci stati errori umani nella prima fase (determinazione della posizione delle Unità Informative negli enunciati, mediante valutazione soggettiva dei parametri acustici e del contesto), (b) una parte dei dati devono necessariamente risentire dei tipici "difetti" dell'oralità, come produzioni imperfette, cambi di programmazione ed esecuzione, ecc., e (c) l'efficienza dell'algoritmo automatico nell'assegnare livelli di prominenza alle sillabe non può essere, e non è, del 100%.

5. CONCLUSIONI

Gli esperimenti descritti e la loro possibile interpretazione data qui sopra consentono le seguenti provvisorie conclusioni riguardo alle varietà italiane esaminate:

1. Si può evidenziare un livello astratto e meramente strutturale della Prominenza, topologico e non qualitativo, in cui la sua *mera collocazione* ha la funzione di demarcare il confine tra le unità informative, prima che quella di di produrre una culminazione su ciascuna di esse, e una "caratterizzazione" di ciascuna.

   Questo aspetto della Prominenza potrebbe già bastare a spiegare i processi attraverso cui i parlanti interpretano la Struttura Informativa degli enunciati nel discorso. Altri tratti, come gli specifici contorni intonativi delle diverse unità informative, in questa funzione potrebbero dunque rappresentare una certa dose di ridondanza.
2. Gli enunciati reali non segnalano sempre in maniera chiara la distribuzione di Topic e Focus. Sul piano acustico, molti rimangono sottospecificati da questo punto di vista. Ciò è vero in modo particolare per la distinzione tra Topic-Focus e Focus Esteso, che spesso non ha effetti rilevanti sulla progressione del dinamismo comunicativo nel discorso successivo.
3. La coerenza di questi risultati con la legge del minimo sforzo, e l'alta pecentuale di coincidenza fra valutazioni percettive e misurazione automatica, sembrano convalidare in maniera molto soddisfacente l'algoritmo adoperato per questa analisi.

6. BIBLIOGRAFIA

Albano Leoni, F. (2003), Tre progetti per l'italiano parlato, in Atti del XXXIV Congresso SLI, Firenze, 675–683.

Avesani, C. (2000), Costruzioni marcate e non marcate in italiano. Il ruolo dell'intonazione., in Atti delle X giornate di studio del GFS, Il parlante e la sua lingua (D. Locchi, A. Giannini & M. Pettorino, edtors), Napoli, 1–14.

Avesani, C. & Vayra, M. (2004), Focus ristretto e focus contrastivo in italiano, in Il Parlato Italiano, Atti del Convegno Nazionale (F. Albano Leoni, F. Cutugno, M. Pettorino & R. Savy, editors), Napoli, 1–20.

Avesani, C., Vayra, M., Zmarich, C., Paggiaro, R. & Sperandio, D. (2007), Le basi articolatorie della prominenza accentuale in italiano, in Atti del III convegno AISV (V. Giordani, V. Bruseghini & P. Cosi, editors), Trento, 1–22.

Bagshaw, P. (1994), Automatic prosodic analysis for computer-aided pronunciation teaching. PhD thesis, University of Edinburgh, UK.

Beckman, M.E., Hirshberg, J. & Shattuck-Hufnagel, S. (2005), The original ToBI system and the evolution of the ToBI framework, in Prosodic models and transcription: Towards prosodic typology (S. Jun, editor), Oxford: Oxford University Press, 9–54.

Bolinger, D. (1958), A theory of pitch-accent in English, Word, 14, 109–149.

Bonvino, E. (2005), Le sujet postverbal. Une étude sur l'italien parlè, Paris: Ophrys.

Breen, M., Fedorenko, E., Wagner, M. & Gibson, E. (2010), Acoustic correlates of information structure, Language and Cognitive Processes, 25, 1044–1098.

Chafe, W. (1987), Cognitive Constraints on Information Flow, in Coherence and Grounding in Discourse (R.S. Tomlin, editor), Benjamins, 21–51.

Chafe, W. (1992), Information Flow in Speaking and Writing, in The Linguistics of Literacy (P. Downing, S.D. Lima & M. Noonan, editors), Benjamins, 17–29.

Couper-Kuhlen, E. (1986), English prosody, London: Edward Arnold.

Cresti, E. (1992), Le unità d'informazione e la teoria degli atti linguistici, in Atti del XXIV Congresso SLI (G. Gobber, editor), Bulzoni, 501–529.

Cresti, E. (2000), Corpus di italiano parlato, Firenze: Accademia della Crusca.

Daneš, F. (1967), Order of Elements and Sentence Intonation, in Studies to Honor Roman Jakobson, The Hague-Paris: Mouton, 499–512.

Daneš, F. (1974), Functional Sentence Perspective and the Organization of the Text, in Papers on Functional Sentence Persepctive (F. Daneš, editor), Prague: Academia /The Hague: Mouton, 106–128.

D'Imperio, M. (2002a), Language-specific and universal constraints on tonal alignment: the nature of targets and anchors, in Proceedings of Speech Prosody 2002, Aix-en-Provence, France, 101-106.

D'Imperio, M. (2002b), Italian Intonation: An overview and some questions, Probus, 14, 37–69.

Fant, G., Kruckenberg, A. & Liljencrants, J. (2000), Acoustic-phonetic Analysis of Prominence in Swedish, in Intonation (A. Botinis, editor), Kluwer, 55–86.

Féry, C. & Krifka, M. (2008), Information structure. Notional distinctions, ways of expression, in Unity and diversity of languages (P. van Sterkenburg, editor), Benjamins, 123-136.

Firbas, J. (1966), On Defining the Theme in Functional Sentence Analysis, Travaux Linguistiques de Prague, 1, 267–280.

Firbas, J. (1987), On the Delimitation of the Theme in Functional Sentence Perspective, in Functionalism in linguistics (R. Dirven & V. Fried, editors), Amsterdam-Philadelphia: Benjamins, 137-156.

Firbas, J. (1989), Degrees of communicative dynamism and degrees of prosodic prominence (weight), Brno Studies In English, 18, 21–66.

Frascarelli, M. (2000), The Syntax-Phonology Interface in Focus and Topic Constructions in Italian, Studies in Natural Language and Linguistic Theory, 50, Kluwer.

Frascarelli, M. (2004), L'interpretazione del Focus e la portata degli operatori sintattici, in Il Parlato Italiano, Atti del Convegno Nazionale (F. Albano Leoni, F. Cutugno, M. Pettorino, R. Savy, editors), B06, Napoli.

Frascarelli, M. & Hinterhölzl, R. (2007), Types of Topics in German and Italian, in On Information Structure, Meaning and Form (S. Winkler & K. Schwabe, editors), Benjamins, 87–116.

Gili Fivela, B. (2006), Tonal alignment in two Pisa Italian peak accents, in Proceedings of Speech Prosody 2002, Aix-en-Provence, France, 339–342.

Halliday, M.A.K. (1989), Spoken and Written Language, Oxford: Oxford University Press.

Heldner, M. (2003), On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish, Journal of Phonetics, 31, 39–62.

Jensen, C. (2004), Stress and Accent, Phd thesis, University of Copenhagen.

Kohler, K.J. (2005), Form and Function of Non-Pitch Accents, in Prosodic Patterns of German Spontaneous Speech, AIPUK, 35a, 97–123.

Kohler, K.J. (2006), What is emphasis and how is it coded? In Proceedings of Speech Prosody 2006, Dresden, Germany, 748–751.

Ladd, D.R. (1978), The Structure of Intonational Meaning, Bloomington: Indiana University Press.

Ladd, D.R. (1996), Intonational Phonology, Cambridge: Cambridge University Press.

Lee, Y. & Xu, Y. (2010), Phonetic Realization of Contrastive Focus in Korean, in Proceeding of Speech Prosody 2010, Chicago, paper 033.

Lombardi Vallauri, E. (2001), La teoria come separatrice di fatti di livello diverso, L'esempio della struttura informativa dell'enunciato, in Atti del XXXIII Congresso SLI, Napoli, 151–173.

Lombardi Vallauri, E. (2009), La struttura informativa, Forma e funzione negli enunciati linguistici, Carocci.

Marotta, G. (2008), Phonology or non phonology? That is the question (in intonation), Estudios de Fonética Experimental, Universitat Autònoma de Barcelona, XVII,177–206.

Mertens, P. (1991), Local prominence of acoustic and psychoacoustic functions and perceived stress in French, in Proceedings of ICPhS'91, Aix-en-Provence, 218–221.

Pierrehumbert, J. (1987), The Phonology and Phonetics of English Intonation (Ph.D. thesis 1980), Indiana University Linguistics Club.

Pitrelli, J.F., Beckman, M.E., Hirschberg, J. (1994), Evaluation of Prosodic Transcription Labelling Reliability in the ToBI Framework, in Proceedings of ICSLP'94, Yokohama, 123–126.

Selkirk, E. (1984), Phonology and Syntax: The Relation between Sound and Structure, Cambridge, MA: MIT Press.

Sgall, P. (1975), Conditions of the Use of Sentences and a Semantic Representation of Topic and Focus, in Formal Semantics of Natural Language (E. Keenan, editor), Cambridge: Cambridge University Press, 297–312.

Sgall, P., Hajicová, E. & Benesová, E. (1973), Topic, Focus and Generative Semantics, Kronberg Taunus: Scriptor.

Sluijter, A., & van Heuven, V. (1996), Spectral balance as an acoustic correlate of linguistic stress, Journal of the Acoustical Society of America, 100, 2471–2485.

Streefkerk, B. (1996), Prominent accent and pitch movements, Institue of Phoneic Sciences Proceedings, University of Amsterdam, 20,111–119.

Syrdal, A. & McGorg, J. (2000), Inter-transcriber reliability of ToBi prosodic labelling, in Proceedings of ICSLP2000, Bejing, 235-238.

Tamburini, F. (2005), Automatic Prominence Identification and Prosodic Typology, in Proceedings of InterSpeech 2005, Lisbon, 1813–1816.

Tamburini, F. (2006), Reliable Prominence Identification in English Spontaneous Speech, in Proceedings of Speech Prosody 2006, Dresden, Germany, PS1–9–19.

Tamburini, F. (2009), Prominenza frasale e tipologia prosodica: un approccio acustico, in Atti del XL Congresso SLI, Vercelli, 437–455.

Taylor, P.A. (2000), Analysis and Synthesis of Intonation using the Tilt Model, Journal of the Acoustical Society of America, 107, 1697–1714.

Talkin, D. (1995), A robust algorithm for pitch tracking (rapt), in Speech coding and synthesis (W. Kleijn & K. Paliwal, editors), New York: Elsevier, 495–518.

Terken, J. (1991), Fundamental Frequency and perceived prominence parameters, Journal of the Acoustical Society of America, 87, 1768–1776.

PERCEZIONE LINGUISTICA E DISCRIMINAZIONE DI ACCENTI INTONATIVI

Barbara Gili Fivela
Università del Salento & CRIL, Lecce
barbara.gili@unisalento.it

1. SOMMARIO

L'articolo descrive un test di percezione categoriale, effettuato in relazione a due accenti tonali individuati nell'italiano di Pisa. L'obiettivo principale è verificare se il disegno sperimentale e, in particolare, la scelta dei passi di manipolazione possa aver un forte peso sui risultati relativi alla percezione categoriale; inoltre, un secondo obiettivo è capire se, in mancanza della possibilità di fare riferimento a soglie percettive come la Just Noticeble Difference per indagini sulla sincronizzazione degli eventi intonativi, si possano fornire delle indicazioni procedurali per la scelta del passo di manipolazione da usare nel test di discriminazione, notoriamente più problematico rispetto al test di identificazione.

Sulla base dei risultati discussi nell'articolo, si mette in evidenza l'importanza della scelta del passo di manipolazione per i test di percezione categoriale e si propone un'indicazione procedurale. In particolare si suggerisce che i risultati migliori nel test di discriminazione si possano ottenere quando la differenza tra gli stimoli delle coppie da discriminare corrisponde a circa 2/3 dell'intervallo richiesto per il cambio di categoria nell'identificazione. Si tratta di un criterio da verificare con ulteriori indagini sperimentali, ma utile nel tentativo di minimizzare i problemi connessi ai test di percezione categoriale e di discriminazione che riguardino eventi intonativi.

2. INTRODUZIONE

Per lungo tempo la percezione di eventi fonologici è stata considerata categoriale e verificata grazie a un metodo sperimentale proposto inizialmente per la percezione di consonanti. Lieberman et al. (1957) descrivono il primo esperimento volto a osservare la percezione dei fonemi consonantici /b/, /d/, /g/, basandosi sul presupposto che, dato un continuum di variazione fonetica, i soggetti siano in grado di associare gli stimoli che compongono il continuum alle diverse categorie linguistiche rilevati, passando repentinamente dal riconoscimento dell'una a quello dell'altra. Di fatto l'assunto è che essi riportino gli stimoli acusticamente diversi all'una o all'altra categoria, poiché la variabilità acustica interna alle categorie si suppone non abbia un forte impatto sul loro riconoscimento; al contrario, anche una minima variazione al confine tra categorie linguistiche diverse produce un chiaro effetto sulla percezione (v. *Quantal Theory of speech* - Stevens 1972, 1989; Steven e Kaiser 2010).

Secondo la proposta di Lieberman et al. (1957) e gli studi successivi, la percezione categoriale è confermata quando gli stimoli siano sia identificati che discriminati. Ci sono diversi tipi di test di identificazione e discriminazione, ma, in generale, nel primo i soggetti ascoltano stimoli che rappresentano un continuum di variazione tra possibili categorie fonologiche e sono invitati ad assegnare gli stimoli a una delle categorie previste nell'esperimento; nel test di discriminazione, invece, i soggetti ascoltano stimoli adiacenti nel continuum utilizzato per l'identificazione e stabiliscono se essi siano uguali o diversi. La presenza di un brusco cambiamento nell'associazione degli stimoli all'una o all'altra categoria nel test di identificazione (curva ad S nel grafico delle risposte) e il riscontro di un picco di discriminazione in corrispondenza degli stimoli al confine tra le categorie (quelli rispetto ai quali si realizza il repentino cambiamento nelle risposte dei soggetti) permettono di verificare se esista una percezione di tipo categoriale. In sostanza, ci si aspetta che gli stimoli che appartengono a una categoria fonologica non siano discrimi-

nati tra loro, anche se acusticamente differenti, mentre quelli che si trovano al confine tra categorie diverse, ossia stimoli che non sono stati identificati come chiaramente appartenenti all'una o all'altra categoria, dovrebbero essere facilmente discriminati tra loro (v. Quantal Theory, citata in precedenza).

Il paradigma sperimentale tradizionale è quindi coerente con l'idea che le unità fonologiche siano discrete. Tuttavia, la percezione categoriale non è stata ugualmente verificata per tutti gli eventi linguistici. Ad esempio, le vocali, a differenza delle consonanti, non risultano essere categorialmente: l'identificazione delle vocali è caratterizzata da un cambiamento meno repentino nell'associazione degli stimoli alle categorie e i risultati del compito di discriminazione mostrano plateau piuttosto che picchi, indicando la presenza di un confine non netto e discreto tra le categorie. Ad esempio, nel caso della percezione delle vocali inglesi / /, / /, /æ/, Fry et al. (1962) suggeriscono che risultati simili a quelli descritti siano dovuti alla mancanza della discontinuità articolatoria che si riscontra nelle consonanti. La percezione categoriale potrebbe quindi dipendere dalla presenza di discontinuità articolatorie (una sorta di produzione categoriale), in linea con l'idea che la percezione implichi una sorta di *auditory-to-articulatory mapping* e sia quindi mediata dal sistema motorio (v. *Motor Theory of Speech Perception,* Lieberman et al., 1957; Lieberman & Mattingley, 1985).

Da questo punto di vista, le inchieste relative a eventi intonativi risultano quindi particolarmente interessanti e ci si può aspettare diano risultati simili a quelli ottenuti per le vocali. Infatti, il correlato principale delle variazioni intonative è la modulazione della frequenza fondamentale della voce (F0) che avviene attraverso cambiamenti continui e graduali, piuttosto che discontinui e bruschi: qualsiasi incremento o decremento di frequenza avviene in modo continuo, seppur repentino, mentre l'unica modificazione che possa veramente dirsi discontinua è quella che corrisponde al passaggio da assenza a presenza di vibrazione delle pliche vocali, o viceversa, come avviene nel caso del passaggio da foni sordi a sonori. In questi casi, però, le modificazioni sono di tipo micro-prosodico e non veicolano informazioni intonative; il riferimento ad elementi di discontinuità può essere quindi solo di tipo secondario (v. oltre).

Se i risultati ottenuti per gli eventi intonativi fossero sempre simili a quelli ottenuti per le vocali – e quindi si riscontrasse sempre un plateau piuttosto che un picco di discriminazione – l'osservazione obbligata e più semplice sarebbe che alcune categorie linguistiche, per esempio le vocali e gli eventi intonativi, siano caratterizzate da confini non discreti, diversamente da quanto avviene per altre categorie fonologiche, ad esempio le consonanti. Tuttavia, gli studi nei quali si è applicato all'intonazione il paradigma sperimentale della percezione categoriale hanno riportato risultati non sempre coerenti, rendendo più complicato valutare l'adeguatezza del metodo e, soprattutto, le conseguenze dei risultati della sua applicazione. Infatti, negli esperimenti nei quali sono state studiate le categorie intonative talvolta è stata riscontrata la mancanza di un picco di discriminazione nelle risposte dei soggetti, talvolta è stato individuato un plateau di discriminazione piuttosto che un picco e, in altri casi, è stato individuato l'atteso picco di discriminazione. Per esempio, Savino e Grice (2011) hanno studiato il ruolo dell'altezza tonale nella percezione di due categorie pragmatiche individuate nella varietà di italiano parlata a Bari, corrispondenti alle domande per richiesta di informazioni e ad atti linguistici che mettono in discussione le informazioni date e condivise. Entrambe le funzioni sono svolte dallo stesso accento tonale (L+H*) che viene però realizzato a due diverse altezze tonali medie. Per testare la percezione dei due accenti, gli autori hanno realizzato un test di identificazione e uno di discriminazione e hanno trovato che i soggetti interpretano categorialmente gli stimoli nel compito di identificazione (un compito motivato semanticamente), ma sembrano essere inaffidabili quando gli viene chiesto di discriminare le coppie di stimoli. D'altra parte, la percezione categoriale viene individuata da Vanrell (2006) in uno studio sulle domande aperte e polari in catalano maiorchino, veicolate entrambe da un innalzamento finale di F0: in questo studio, i soggetti sono stati in grado di identificare e discriminare chiaramente due andamenti distinti in un continuum di variazione di altezza tonale. Una sorta di risultato intermedio è stato riportato, invece, per il te-

desco da Schneider et al. (2006) che hanno studiato la percezione di un continuum di variazione di altezza tonale in relazione all'interpretazione dichiarativa e interrogativa. Gli autori non hanno trovato evidenza a favore della percezione categoriale, visto che i loro risultati vanno nella direzione della categorialità solo per l'identificazione (benché il passaggio nelle risposte a favore dell'una o dell'altra categoria avvenga in 5 passi); per la discriminazione, infatti, ottengono due picchi collegati da un plateau.

Un primo aspetto da chiarire è che per spiegare l'incoerenza tra i risultati non sembrano essere sufficienti le specifiche caratteristiche fonetiche degli andamenti intonativi analizzati e le funzioni linguistiche che essi svolgono. Infatti, le incongruenze nei risultati riguardano anche andamenti molto simili tra loro, benché usati in diverse lingue. Remijsen e van Heuven (1999) e Schneider et al. (2006), per esempio, hanno preso in esame un andamento intonativo che differenzia affermazioni e domande, rispettivamente in olandese e tedesco, e che consiste in un innalzamento tonale sulla sillaba finale di un sintagma prosodico (in termini autosegmentali, L% si trova nelle affermative e H% nelle domande). Gli andamenti considerati sono quindi molto simili sia foneticamente, ad esempio rispetto alla forma della curva di F0 (ascendente) e alla sincronizzazione rispetto alla sillaba, sia funzionalmente, perché distinguono affermazioni e domande. Nonostante questo, i due studi riportano risultati diversi sulla percezione: solo in olandese gli andamenti sono percepiti in modo categoriale, mentre in tedesco questo non avviene, vista la presenza di un plateau di discriminazione piuttosto che di un picco.

In realtà, i problemi legati all'uso del paradigma sperimentale della percezione categoriale per studiare l'intonazione sono stati evidenziati sin dalle sue prime applicazioni per questo scopo (Massaro, 1998), in particolare con riferimento al test di discriminazione. I motivi principali discussi nella letteratura sull'argomento per rendere conto della mancanza di picchi di discriminazione nei test sull'intonazione sono stati la somiglianza tra la percezione e la produzione di intonazione e di suoni vocalici, data la mancanza di discontinuità (Fry et al, 1962;. Schneider et al., 2006), e la mancanza di adeguati eventi acustici di riferimento che facilitino la percezione del cambiamento categoriale (Niebuhr e Kohler, 2004). In particolare, nell'ultimo caso si tratta di discontinuità nel segnale acustico che non riguardano direttamente la variazione di F0, ma piuttosto gli eventi linguistici ai quali la variazione di F0 è correlata e rispetto ai quali è valutata (infatti l'allineamento degli eventi tonali rispetto ai fenomeni segmentali è di importanza primaria per la distinzione degli eventi intonativi). Per esempio, Niebuhr e Kohler (2004) ottengono diversi risultati sul tedesco, nello studio di un andamento con picco di F0 e di uno con valle: i risultati del test di discriminazione che riguardano l'andamento con picco di F0 corrispondono a un picco di discriminazione, mentre quelli che riguardano la valle intonativa non presentano alcun picco di discriminazione. Gli autori sostengono che la minore capacità di discriminazione riscontrata per le valli tonali potrebbe essere dovuta al fatto che le valli sono distinte da più di un correlato e sono allineate a periodi interni alla vocale, piuttosto che a punti articolatoriamente ben definiti nel tempo, come l'onset della vocale stessa. Un aspetto importante è che comunque la mancanza di percezione categoriale in questo caso è attribuita dagli autori a proprietà percettive e non è presa come prova dell'assenza di categorie fonologiche.

In questo nostro contributo, l'attenzione è rivolta a un altro problema relativo ai test di percezione categoriale, che sembra nuovamente avere più effetti sul compito di discriminazione che su quello di identificazione: si tratta di un aspetto che riguarda il disegno sperimentale e, in particolare, la scelta dei passi di manipolazione. Per via del riferimento alle proprietà percettive, infatti, ci si può chiedere se la mancanza di un picco di discriminazione nei risultati possa anche essere dovuta alla scelta di un passo di manipolazione sbagliato nella creazione del continuum di variazione e, soprattutto, alla strategia adottata per creare le coppie di stimoli da discriminare; in particolare, gli stimoli nelle coppie potrebbero essere troppo simili per essere distinguibili, ad esempio se la soglia corrispondente alla minima differenza percepibile (Just Noticeable Difference - JND; Gescheider, 1976) non viene superata.

Come House (1997) afferma nel suo lavoro sulla sensibilità alle informazioni temporali che riguardano l'intonazione rispetto ai cambiamenti di categoria, "the perception of tonal movement in contour tones within a given tonal system does not readily lend itself to comparison with psychophysical thresholds". Di fatto, e non a caso, ancora oggi in letteratura non esiste alcun lavoro, di cui l'autore di questo contributo sia a conoscenza, in cui siano riportate indicazioni circa la JND per la percezione di movimenti intonativi in relazione a differenze nella loro sincronizzazione temporale. Inoltre, se anche ci fossero indicazioni pertinenti in letteratura, si dovrebbero considerare distintamente a seconda che riguardino andamenti ascendenti o discendenti, e a seconda delle loro diverse caratteristiche di allineamento rispetto alla sillaba o ad altra unità di riferimento. Per esempio, House (1987) prende in considerazione i passi di manipolazione necessari per effettuare un cambiamento completo di categoria in test percettivi e conclude che, nella maggior parte dei lavori analizzati, la sensibilità alle differenze di sincronizzazione è di circa 50 ms. Tuttavia questa osservazione sulla sensibilità alle differenze di sincronizzazione non ci dice nulla sulle soglie di percezione relative ad andamenti ascendenti e discendenti. Infatti, come osserva House, visto che 50 ms sono un intervallo di tempo abbastanza lungo per "substantially alter the pitch pattern in the vowel, it seems reasonable to assume a more complex perceptual mechanism involving higher order cognitive processing and short term memory such as the precategorical acoustic storage". Concludendo, "Perceptual timing sensitivity would thus be conditioned by the spectral environment" e, quindi, sembra che solo l'evidenza sperimentale possa mostrare qual è la soglia per la percezione di un dato andamento intonativo.

Alcuni sperimentatori hanno dimostrato che incrementando la differenza tra gli stimoli, per esempio aumentando la durata del passo di allineamento o accoppiando nel test di discriminazione stimoli non adiacenti nel continuum di variazione usato per il test di identificazione, si può facilitare l'emergere di un picco di discriminazione nei risultati (Kohler, 1987). Tuttavia, questa soluzione non sembra essere sempre sufficiente. Per esempio, in uno studio sulla variazione dell'altezza tonale e accoppiando sia stimoli non adiacenti che stimoli adiacenti nel continuum di variazione tra H*+L e H+L* nell'italiano di Pisa, Gili Fivela (2008) ha riscontrato un aumento complessivo nella capacità di discriminazione, piuttosto che l'emergere di un picco di discriminazione. La differenza tra i risultati riportati in letteratura è probabilmente anche dovuta al fatto che le manipolazioni eseguite (ad esempio quelle di scaling nel caso di H*+L e H+L* nell'italiano di Pisa) non permettono sempre si superare effettivamente un confine di categoria o un confine per il quale ci sia percezione categoriale. Tuttavia, sembra che ad oggi solo l'evidenza sperimentale possa mostrare se un picco di discriminazione non emerge a causa delle opzioni di manipolazioni scelte o perché non vi è effettivamente alcun picco di discriminazione per una manipolazione tonale specifica. In ogni caso, è chiaro che la scelta della dimensione del passo di manipolazione deve essere effettuata molto attentamente, in modo che sia sufficiente ampia da favorire l'emergere di un picco di discriminazione (se presente), ma abbastanza limitata da non causare un aumento generalizzato della discriminazione tra gli stimoli che possa addirittura mascherare la presenza di un picco significativo. Inoltre, è chiara anche la necessità di individuare sperimentalmente quali possano essere le caratteristiche ideali dei passi di manipolazione da utilizzare per effettuare indagini su eventi intonativi specifici. In questo studio, ad esempio, l'attenzione è stata rivolta a due accenti tonali individuati nell'italiano parlato a Pisa.

Di fatto, il riferimento a differenze acustiche (e articolatorie) non sembra poter spiegare la complessità del fenomeno. Lo studio dei contrasti intonativi offre uno scenario composito, in cui le caratteristiche acustiche, così come la presenza di funzioni linguistiche contrastive, possono non essere i soli fattori rilevanti.

3. LA PERCEZIONE DI DUE ACCENTI NELL'ITALIANO DI PISA

Due accenti tonali individuati nell'italiano di Pisa, fonologicamente analizzati come H* e H*+L, presentano le caratteristiche adatte per uno studio sulle categorie intonative e sui loro confini precettivi. I due accenti sono stati ampiamente osservati e analizzati in dati di produzione e mostrano una forma fonetica simile, adatta alla creazione di un continuum di manipolazione quando entrambi gli accenti siano seguiti da un confine prosodico intermedio basso (in termini autosegmentali-metrici, un accento di sintagma intermedio). Entrambi gli accenti sono caratterizzati da un innalzamento di F0 che inizia verso l'attacco della sillaba accentata. Nell'accento H*, l'innalzamento raggiunge un picco nei pressi della fine della sillaba tonica, mentre nell'accento H*+L il picco di F0 è raggiunto prima, all'interno della sillaba. Inoltre, nelle dichiarative H* può essere seguito da un confine prosodico intermedio quando occupi la posizione nucleare all'interno di un sintagma che non sia in posizione finale di enunciato (Gili Fivela, 1999). Invece, l'accento H*+L, come accento contrastivo, secondo alcuni in italiano è sempre seguito una barriera prosodica (Frascarelli 2000), situazione che sembra verosimile anche per l'italiano di Pisa (Gili Fivela, 1999, 2008). Rispetto ai valori di F0 dell'accento e di durata della sillaba tonica, H* presenta valori di F0 mediamente maggiori e valori di durata della sillaba tonica mediamente inferiori rispetto a H*+L (Gili Fivela, 2004, 2005, 2006; Prieto et al, 2005)[1]. Quindi, entrambi gli accenti si possono trovare in posizione finale di sintagma, ma generalmente (v. funzioni specificate nel seguito) solo H*+L si trova in posizione finale di enunciato; inoltre, sono solitamente seguiti da diversi andamenti post-focali, visto che normalmente H*+L è seguito da materiale prodotto in una gamma compressa di frequenze. Per questo motivo, quando gli accenti sono stati studiati percettivamente (Gili Fivela 2004), sono stati posizionati in sintagmi iniziali all'interno dell'enunciato, senza fornire il materiale seguente (che avrebbe dato indicazioni utili, ma che non erano oggetto di indagine, per l'individuazione della categoria). Questa scelta ha permesso di studiare gli accenti in contesti fonetici simili (entrambi seguiti da un confine tonale basso), anche se questo può aver favorito una sfumatura di interpretazione di tipo "conclusivo" vs "di continuazione" oltre alla usuale funzione svolta dagli accenti. Non a caso, le domande alle quali i soggetti hanno risposto sono state formulate tenendo conto di questi aspetti (vedi 4.2.1).

Per quanto riguarda la loro funzione e il loro significato, nell'italiano parlato a Pisa H * seguito da un accento di sintagma intermedio basso è particolarmente adatto per indicare la continuazione, il focus ristretto o, nel caso di contrasto tra più elementi dati, sembra essere adatto per una forma di contrasto sintagmatico interno all'enunciato piuttosto che per un contrasto paradigmatico (ad esempio, "X mangia, Y dorme" piuttosto che "X mangia" che implica che la predicazione non vale per Y, Z, ecc.; l'accento può quindi essere utilizzato in topic contrastivi – Gili Fivela 1999). L'accento fonologicamente analizzato come H*+L seguito da un tono di confine basso esprime un accento contrastivo con chiaro valore di correzione; la correzione è perentoria e conclusiva e quindi l'accento si può trovare anche in posizione finale di enunciato (inoltre, in alcune situazioni dialogiche l'accento può essere pragmaticamente interpretato come una richiesta di conferma di informazione, grazie al fatto che l'ascoltatore effettua una sorta di im-

[1] Altre caratteristiche fonetiche sono le seguenti: il picco è allineato a circa il 50% e al 100% della durata della sillaba (aperta) rispettivamente in H*+L e in H*; la differenza è in media di circa 80 ms, considerando la latenza dall'attacco sillabico. Nei due accenti, il leading tone basso è allineato a 1 ms vs 25 ms dall'onset della sillaba e, quando sia presente un confine intermedio dopo l'accento, il secondo target basso è allineato 10 ms prima dell'attacco della sillaba posttonica in H*+L e 133 ms dopo l'attacco nell'accento H*. Per quanto riguarda la durata della sillaba tonica, la sillaba è più lunga del 7-10% nel caso dell'accento H*+L (Gili Fivela, 2004, 2005, 2006a; Prieto et al., 2005). In particolare, per gli stimoli presi in esame in questo studio, la differenza di durata della sillaba è di 18 ms e quella di altezza del picco di 23 Hz.

plicatura conversazionale griceana, motivata intonativamente – per una discussione più approfondita, v. Gili Fivela, 2008). In un primo tentativo di differenziare i loro significati, è stato sostenuto che H* può essere utilizzato per attirare l'attenzione sulle unità di informazione all'interno delle quali viene realizzato - anche per (re)introdurle - e per aggiungere informazioni ad esse correlate (quindi anche per la continuazione); invece H*+L può essere utilizzato per opporre un elemento alle conoscenze condivise (Gili Fivela, 2008). Così sia H* che H*+L possono esprimere una forma di focus ristretto di tipo contrastivo, e in questo caso i fattori principali che orientano la scelta dell'accento da parte del parlante sono l'atteggiamento del parlante e/o la posizione dell'accento nell'enunciato in relazione alla struttura dell'informazione.

Gili Fivela (2005, 2008) ha studiato la percezione di questi due accenti per mezzo del paradigma sperimentale di percezione categoriale. Una serie di esperimenti di identificazione e di discriminazione è stata eseguita al fine di verificare la percezione categoriale dei due accenti, analizzando il ruolo delle caratteristiche di altezza tonale e di allineamento nella loro percezione. Ciò significa che gli esperimenti sono stati eseguiti manipolando separatamente le caratteristiche di altezza e allineamento tonale. I risultati hanno mostrato che non si realizzava una percezione categoriale per H* e H*+L, tuttavia l'allineamento sembrava giocare un ruolo significativo rispetto all'identificazione, influenzato dai valori di altezza tonale: i valori di identificazione tendevano a spostarsi verso la categoria che mostrava caratteristiche medie di altezza tonale simili a quelle dello stimolo. Tuttavia i soggetti sembravano interpretare categorialmente gli stimoli agli estremi del continuum, cioè gli stimoli che mostravano correlati coerenti sia in termini di allineamento e altezza tonale. D'altra parte, i risultati mettevano in evidenza che i soggetti non discriminavano gli stimoli a seconda delle caratteristiche di allineamento, ma piuttosto percepivano chiaramente le differenze di altezza tonale. In ogni caso, nei risultati non era visibile un picco di discriminazione, quanto una buona capacità di discriminazione generalizzata. Quindi, le indagini hanno mostrato che, anche se non sembra esserci capacità di discriminazione nel caso di stimoli adiacenti nel continuum di allineamento presentati a coppie, i soggetti percepiscono le differenze tra stimoli foneticamente molto simili, tanto da giudicarli come non ugualmente adatti a rappresentare la categoria alla quale erano associati (v. i risultati di identificazione e v. anche Gili Fivela, 2012, in cui si descrive un esperimento di valutazione degli stimoli come rappresentanti di una data categoria). D'altra parte, i test appena menzionati sono stati organizzati per studiare separatamente il ruolo dell'altezza e dell'allineamento nella percezione degli accenti e le scelte effettuate nel disegno sperimentale a causa di questo obiettivo possono aver mascherato l'esistenza della percezione categoriale. Infatti, la maggior parte degli stimoli aveva necessariamente caratteristiche ambigue in termini di allineamento e/o di scaling, visto che i correlati sono stati manipolati separatamente.

Nello studio descritto in questo articolo, i due accenti saranno nuovamente analizzati al fine di verificare se possano essere percepiti in modo categoriale e quale sia il disegno sperimentale migliore al fine di verificare la percezione categoriale. In particolare, si cercherà di capire fino a che punto i risultati dei test di discriminazione possano dipendere dal passo di manipolazione degli stimoli e se il passo di manipolazione ideale si possa inferire dalle categorie prese in esame (es. dal numero di passi di manipolazione necessari per il cambio di categoria nell'identificazione e, quindi, dalla durata dell'intervallo necessario per passare da una categoria all'altra; v. considerazioni di House, 1997).

4. VERIFICA DELLA PERCEZIONE CATEGORIALE

4.1. Stimoli

Lo stimolo di base è rappresentato dall'enunciato "No. Ho detto velava velocemente", prodotto da un locutore pisano in un contesto che induceva la realizzazione di focalizzazione contrastivo-correttiva sulla parola bersaglio 'velava'. Si tratta di uno stimolo già utilizzato per le precedenti indagini percettive

relative alla varietà di Pisa (Gili Fivela, 2005, 2008). In questo studio, lo stimolo, dopo esser stato privato della negazione e dell'avverbio (ridotto quindi a "ho detto velava"),è stato manipolato al fine di ottenere un passaggio graduale da un accento H*+L a un H* seguito da un tono di confine prosodico intermedio basso (L-). L'allineamento del picco e della discesa successiva nell'accento H*+L (realizzato sulla penultima sillaba della parola finale nella frase "ho detto velava") è stata spostato in avanti in 8 passi di 15 ms; per ogni passo di manipolazione di allineamento, l'altezza tonale di tutti i target è stata incrementata di 4.5 Hz, per ottenere una variazione coerente con quella di allineamento – Figura 1. La durata della sillaba non è stata modificata e il numero totale di stimoli ottenuto è 9.

La scelta del passo di manipolazione di allineamento è di importanza cruciale ed è stata effettuata al fine di creare stimoli che permettessero di creare un continuum di variazione graduale da una categoria all'altra (i valori di F0 sono stati manipolati coerentemente con il passo di allineamento scelto). In particolare è stato scelto un passo di 15 ms, calcolando il passo più grande che poteva essere utilizzato per ottenere un continuum che fosse composto da almeno 5 stimoli, per poter effettivamente osservare l'esistenza di un cambiamento graduale o repentino nelle risposte dei soggetti. Poiché, secondo misure effettuate in precedenza, la distanza media tra i due picchi è di 80 ms (vedi n.1), un passo di 15 ms garantisce la possibilità di spostarsi in almeno 5 passi dalla posizione media del picco in H*+L alla sua posizione media in H* (in questo esperimento, tre passi sono stati aggiunti al continuum affinché la manipolazione riguardasse l'intera durata della sillaba tonica dello stimolo base, originariamente associata all'accento H*+L dovuto all'interpretazione contrastivo-correttiva).

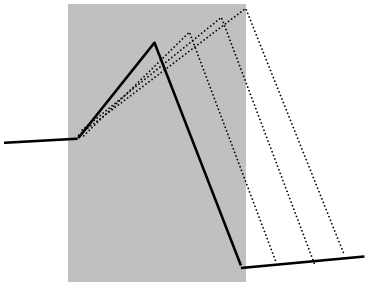


Figura 1. Schema esemplificativo delle manipolazioni effettuate per il passaggio da un accento H*+L (usato per la correzione, linea continua) a un accento H* (seguito da L-, utilizzato per la prosecuzione / (re) introduzione, linea tratteggiata). Il rettangolo grigio rappresenta la sillaba accentata. Per chiarezza, sono rappresentati solo alcuni passi di manipolazione.

La dimensione del passo scelto rientra nella gamma molto ampia di passi usati nella letteratura sulla percezione dell'intonazione e sull'identificazione di categorie. Per esempio, Bruce (1977) utilizza un passo 10 ms, Pierrehumbert e Steele (1989) utilizzano un passo 20 ms, mentre in altri esperimenti ci si fa riferimento a passi di manipolazione più grandi, ad esempio 25 ms in House (1990) e 33 ms in D'Imperio e House (1997). D'altra parte, la dimensione del passo può sembrare bassa rispetto a quella solitamente usata in test di discriminazione. Ad esempio, Kohler (1987) usa un passo di manipolazione di 30 ms ed ottiene un picco di discriminazione per mezzo di un test nel quale accoppia stimoli non adiacenti che si trovano a due passi di distanza, e con coppie, quindi, in cui l'allineamento del picco differisce di 60 ms (strategia simile è stata scelta da Frota, 2012 o Dilley 2005). La dimensione del passo utile per ottenere un picco discriminazione, quindi, in questo caso è ancora maggiore dell'intervallo di tempo identificato da House (1997) come corrispondente alla sensibilità media alle informazioni temporali relative a cambiamenti di categoria (circa 50 ms). A meno che non si voglia equiparare questo intervallo con

quello necessario per ottenere la minima differenza percepibile (JND), e questo non sembra essere il caso (vedi la discussione sui risultati di House nella sezione 2), ci si potrebbe aspettare di trovare risultati a favore della discriminazione (se presente) anche sfruttando passi corrispondenti a intervalli di dimensioni inferiori a quella che consente un cambio di categoria. Si consideri inoltre che passi di entità elevata non sono sempre funzionali all'emergere del picco, anche se facilitano la discriminazione. Queste considerazioni saranno quindi considerate nel disegno del test di discriminazione.

D'altro canto, nel caso in questione, il passaggio di categoria riguarda configurazioni che differiscono anche in termini di altezza tonale. La percezione dell'altezza tonale risulta molto accurata, con studi di tipo psicoacustico che indicano in 1 Hz la JND per toni puri sotto i 250 Hz a 40 dB SL (Gelfand, 1981). Per quanto riguarda la percezione di differenze di altezza nel caso di eventi tonali linguistici, gli studi in letteratura mostrano che meno di 10 Hz sono sufficienti per la percezione di toni alti e bassi in Kammu del nord (Svantesson e House 1996) e 15 Hz è funzionale alla percezione nel catalano di Majorca. Non a caso, un passo di 15 Hz è già stato usato spesso nel caso di identificazione e discriminazione di categorie (Vanrell, 2006; Gili Fivela, 2008; Grice and Savino, 2011).

In ogni caso, in questa indagine la precedenza sarà data alla determinazione dei passi in termini di allineamento, variando coerentemente le caratteristiche di altezza tonale.

4.2.Identificazione

4.2.1 Metodo

Dodici soggetti (25-35 anni), parlanti della varietà di italiano di Pisa, senza problemi di udito e ignari dello scopo dell'esperimento, hanno partecipato al test di identificazione ricevendo un pagamento simbolico per la loro partecipazione. Gli stimoli sono stati presentati attraverso cuffie Sennheiser m@b40 di qualità professionale in una stanza insonorizzata presso il Laboratorio di Linguistica della Scuola Normale Superiore tramite il software Praat (Boersma & Weenick, 2007). I soggetti hanno ascoltato i 9 stimoli per cinque volte, in due blocchi in cui l'ordine degli stimoli è stato randomizzato. Tra i blocchi è stata proposta una pausa e ogni soggetto ha scelto dopo quanto tempo riprendere l'esperimento.

Ai soggetti è stato chiesto di identificare gli stimoli in un test a risposta forzata. La domanda alla quale dovevano rispondere era "può correggere un enunciato precedente in modo perentorio e conclusivo?". In un esperimento pilota, i soggetti erano stati invitati a concentrarsi solo sulla funzione (la domanda era "può correggere un enunciato precedente?") e non sono stati in grado di eseguire il compito, mentre la domanda composita indicata precedentemente ha permesso ai soggetti di identificare diversi tipi di accento (sia nell'esperimento pilota che nei test di identificazione eseguiti finora). Questo aspetto si considera collegato alla funzione svolta dai due accenti e al fatto che entrambi possono esprimere un tipo di focalizzazione ristretta e contrasto, come descritto nel paragrafo 3 (v. discussione in Gili Fivela 2008). I soggetti hanno risposto con il mouse, cliccando in corrispondenza di una delle due caselle visualizzate sullo schermo del PC (una includeva un'etichetta in cui era scritto "sì, è una correzione perentoria e definitiva" e l'altra l'etichetta "no, non è una correzione perentoria e definitiva"). Prima di ogni test è stata effettuata una prova composta da 18 target estratti casualmente dalla lista di stimoli dell'esperimento; se il compito risultava chiaro ai soggetti, l'esperimento poteva avere inizio e durava circa dieci minuti in totale (principalmente a seconda della lunghezza della pausa).

I risultati dell'identificazione sono riportati come percentuale di risposte e analizzati statisticamente mediante ANOVA eseguite sulle medie delle risposte date dai soggetti per ogni item, con il passo di manipolazione (9 livelli) come variabile indipendente.

4.2.2 Risultati

La Figura 2 mostra la media delle risposte dei soggetti a favore dell'interpretazione contrastivo-correttiva.

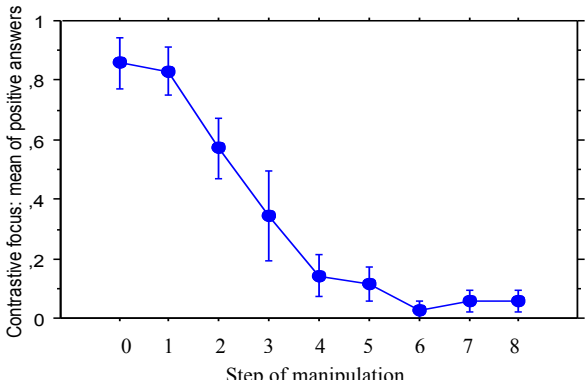


Figura 2. Risultati del test di identificazione: media delle risposte a favore dell'interpretazione contrastivo-correttiva per ogni passo di manipolazione (1 ± SE).

Come mostra il grafico, il cambiamento di categoria nelle risposte dei soggetti si realizza in tre passi ed è quindi piuttosto repentino. Un test ANOVA a una via, effettuato sulla media delle risposte dei soggetti per ogni item con il passo di manipolazione come variabile indipendente, mette in evidenza che il passo di manipolazione è un fattore significativo nell'influenzare il numero di risposte positive a favore dell'interpretazione contrastivo-correttiva [$F(8,99) = 47.280$; $p<0,0001$]. Inoltre, il post-hoc di Fisher mostra che gli stimoli per i passi di manipolazione "0" e "1" corrispondono a punteggi simili tra loro e diversi dagli altri; quelli per i passi "2" e "3" differiscono tra loro e dagli altri, mentre quelli dati per i passi da "4" a "8" sono simili tra loro.

4.3 Discriminazione

4.3.1 Metodo

Gli stimoli utilizzati per il compito di identificazione sono stati poi usati per un test di discriminazione di tipo AX. Le coppie erano di tipo AA – per gli stimoli di controllo - e AX, dove X era sempre caratterizzato da un picco di F0 più alto e più ritardato, al fine di favorire la possibilità di discriminazione[2]. Infatti, diversi lavori in letteratura hanno dimostrato che la discriminazione è facilitata quando il secondo stimolo è caratterizzato da un'altezza tonale maggiore di quella del primo (Ladd & Morton, 1997; Remijsen & van Heuven, 1999; Vanrell, 2006); inoltre, la seconda posizione è stata riservata agli stimoli che hanno un picco di F0 posticipato, in quanto un picco ritardato può rappresentare una strategia alternativa per aumentare l'altezza tonale percepita (Gussenhoven, 2002). L'ordine di presentazione dei due stimoli per ogni trial AX non è stato modificato nelle diverse somministrazioni in modo da ottenere sempre la migliore discriminazione possibile.

[2] Le coppie AX includevano stimoli caratterizzati sempre dalle stesse differenze acustiche e gli stimoli di controllo AA erano composti da stimoli uguali. In linea con gli studi del settore, non era prevista una coppia di stimoli che fossero giudicati con certezza come sicuramente diversi. Tuttavia, come emerso durante la discussione seguita alla presentazione orale di questo lavoro (si ringrazia F. Cutugno per questa osservazione), una verifica della percezione di elementi sicuramente diversi sarebbe auspicabile. In lavori futuri si terrà conto di questo aspetto, ma per i dati discussi in questa sede è sufficiente osservare che gli stessi soggetti hanno effettivamente dimostrato di percepire differenze a seconda degli stimoli e soprattutto delle serie (v. oltre). Quindi, un controllo interno all'esperimento è comunque disponibile.

Sono state create tre serie di coppie di stimoli che sono poi state somministrate separatamente. In un primo blocco, infatti, sono stati accoppiati stimoli adiacenti nel continuum utilizzato per l'identificazione - Serie-1 (cioè all'interno della coppia l'allineamento differiva di 15 ms e la F0 di 4.5 Hz). Tuttavia, per verificare se i risultati del test potessero dipendere dalla differenza di allineamento e F0 tra gli stimoli, sono state create due serie supplementari: nella seconda serie, Series-2, le coppie erano composte da stimoli a due passi di distanza nel continuum usato per l'identificazione (ossia gli stimoli nelle coppie differivano di 30 ms e 9 Hz), mentre nella terza serie, Serie-3, le coppie includevano stimoli che erano a tre passi di distanza nel continuum (45ms e 13.5Hz di differenza).

Il numero di coppie nella Serie-1 era 9 (8 coppie AX e 1 coppia AA, vale a dire circa l'89% dei trial includeva stimoli diversi, mentre circa l'11% corrispondeva a coppie di stimoli uguali, in cui lo stesso stimolo veniva presentato due volte), nella Serie-2 era 8 (7 coppie AX e 1 coppia AA, quindi circa l'87,5% dei trial era "diverso", mentre il 12,5% era "uguale"), nella Serie-3 era 7 (6 coppie AX e 1 coppia AA, vale a dire circa l'86% dei trial era "diverso", mentre circa il 14% era "uguale").

Gli stessi 12 soggetti che hanno partecipato al test di identificazione hanno partecipato anche alla prova la discriminazione - Serie-1 - in una diversa sessione sperimentale che ha avuto luogo circa una settimana dopo il primo test. Gli esperimenti che coinvolgono la Serie-2 e la Serie-3 sono stati eseguiti richiamando gli stessi soggetti alcuni mesi dopo, eseguendo i due test di discriminazione supplementari in due diverse sessioni a tre giorni di distanza l'una dall'altra. Tutti i soggetti erano naïve riguardo allo scopo degli esperimenti e hanno ricevuto un pagamento simbolico per la loro partecipazione. Ogni serie è stata presentata come blocco sperimentale e l'ordine dei blocchi 2 e 3 è stato randomizzato.

Gli stimoli sono stati presentati ai soggetti attraverso cuffie Sennheiser m@b40 di qualità professionale, in una stanza insonorizzata del Laboratorio di Linguistica della Scuola Normale Superiore di Pisa attraverso il software Praat (Boersma & Weenick, 2007). L'intervallo tra gli stimoli in una coppia (ISI) era 350 ms. Per ogni blocco, i soggetti hanno sentito le coppie per cinque volte in ordine casuale. A circa metà blocco è stata proposta loro una pausa la cui durata dipendeva dalla scelta individuale rispetto a quando riprendere l'esperimento. È stato effettuato un test AX (uguale-diverso), chiedendo ai soggetti di giudicare se gli stimoli nelle coppie fossero uguali o diversi; i soggetti rispondevano cliccando con il mouse sulla casella appropriata visualizzata sullo schermo. Prima dell'esperimento vero e proprio, i soggetti hanno effettuato un addestramento corrispondente a 18 trial. L'esperimento durava circa dieci minuti per ogni blocco (a seconda della durata della pausa).

I risultati sono riportati come percentuale di risposte e come valori di *d'*, una misura della sensibilità percettiva derivata dalla *Signal Detection Theory* (MacMillan & Creelman, 1991-2005) che offre valori indipendenti dal rapporto tra gli stimoli "uguali" e "diversi". Per ogni soggetto, è stata calcolata 1) la media delle risposte corrette (hit rate) sulla base del numero delle risposte "diverso" date agli stimoli effettivamente differenti e 2) la media dei falsi allarme (false allarm rate) considerando le risposte sbagliate date agli stimoli di controllo e quindi le risposte "diverso" date agli stimoli uguali. Queste misure sono state quindi utilizzate per calcolare il *d'* per ogni condizione.

I dati sono stati analizzati statisticamente mediante analisi ANOVA a una via, eseguite sulle medie delle risposte date dai soggetti a tutti i trial, separatamente per ogni serie. I fattori considerati sono "coppia" (coppia "0,1", "1,2", e così via, con 8 livelli per la Serie-1; coppia "0,2", "1,3", ecc., corrispondente a 7 livelli per la Serie-2 e, infine, coppia "0,3", "1,4", ecc., per un totale di 6 livelli per la Serie-3).

4.3.2 Risultati

I risultati del test di discriminazione per le tre serie sono riportati in Figura 3. Per ogni coppia, i grafici a sinistra mostrano la media delle risposte dei soggetti a favore di "diverso", mentre i grafici a destra mostrano la media dei valori di *d'*. Dall'alto in basso, i grafici riportano i risultati per la Serie-1, la Serie-2 e la Serie-3.

I dati riportati in Figura 3 mostrano che per la Serie-1 (in cui gli stimoli sono adiacenti nel continuum di manipolazione e quindi differiscono di 15 ms in allineamento e 4.5 Hz in altezza tonale) si ottengono valori bassi di discriminazione e nessun picco. Coerentemente, l'ANOVA a una via, effettuata sulle risposte medie dei soggetti e sui valori di *d'* con "coppia" come fattore indipendente, conferma che non vi è alcuna differenza significativa nelle risposte e, quindi, anche l'assenza di un picco di discriminazione (per la percentuale di discriminazione: [F(7,88)=0,162;p>0,05], per i valori *d'*: [F(7,88)=0,275;p>0,05]).

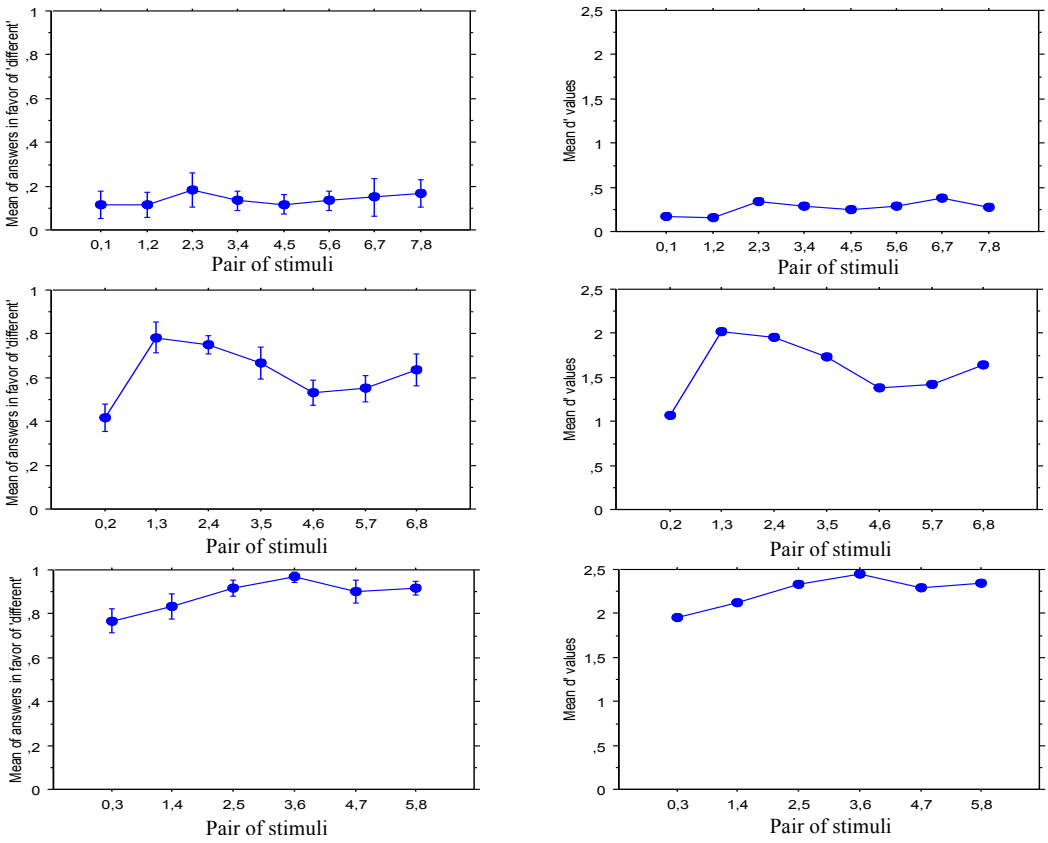


Figura 3. Risultati di Discriminazione per la Serie-1 (grafici in alto), la Serie-2 (grafici centrali) e la Serie-3 (grafici in basso): media delle risposte a favore di "diverso" per ogni coppia di stimoli - grafici di sinistra – e media dei valori dell'indice *d'* di sensibilità alla differenza tra gli stimoli – grafici di destra (± 1 SE).

D'altra parte, valori di discriminazione progressivamente più alti sono individuati per gli stimoli della Series-2 (in cui i membri delle coppie differiscono per 30 ms di allineamento e 9 Hz di altezza tonale) e della Serie-3 (in cui i membri differiscono per 45 ms e 13.5 Hz). Tuttavia, a causa dei valori di discriminazione globalmente elevati, nessun picco di discriminazione è chiaramente visibile per la Serie-3. Per questa serie, infatti, l'intervallo di variazione nel punteggio ottenuto dalle diverse coppie è piuttosto limitato, analogamente a quanto osservato per la Serie-1 (anche se corrisponde a valori globalmente superiori rispetto a quelli riscontrati per la Serie-1). L'ANOVA a una via effettuata sulle risposte medie dei soggetti e sui valori di *d'* con "coppia" come fattore indipendente mostra che esiste una differenza significa-

tiva nella percentuale delle risposte dei soggetti, benché non si riscontri differenza statistica per la sensibilità misurata in termini di *d′* (per la percentuale di discriminazione: [F(5,66) =2,541, p=0,036], per i valori *d′*:[F(5,66)=2,019, p=0,08]). In particolare, il post-hoc di Fisher mostra che i punteggi di discriminazione per la coppia "3,6" sono significativamente più elevati rispetto a quelli assegnati alle coppie "0,3", e "1,4", e che la coppia "0,3" ha ricevuto punteggi di discriminazione inferiori a ogni altra coppia tranne che a "1,4". Quindi, è molto difficile identificare un picco di discriminazione ("3,6" non differisce da "2,5", "4,7", "5,8"), anche se risulta che la coppia più discriminata ("3,6") include uno dei due stimoli corrispondenti al passaggio di categoria nei risultati di identificazione (passo "3"), benché se solo quando sia confrontato con un estremo del continuum (passo "6" e non "0").

Infine, anche i risultati per la Serie-2 (30 ms di differenza in allineamento e 9 Hz di altezza tonale) indicano solo la presenza di una regione di discriminazione che coinvolge le coppie che includono lo stimolo "3" e lo stimolo "2" (in particolare le coppie "1,3", "2,4" e, in misura minore, "3,5"). In particolare, l'ANOVA a una via effettuata sulle risposte medie dei soggetti e sui valori di *d′* con "coppia" come fattore indipendente mostra che vi è differenza significativa sia nei punteggi corrispondenti alle risposte soggetti [F(6,77)=4,112, p=0.0012], sia nella sensibilità misurata dai valori dell'indice *d′* [F(6,77)=3,907, p=0,0019]. In entrambi i casi, il post-hoc di Fisher mostra che i soggetti hanno discriminato le coppie "1,3" e "2,4" più di "4,6", "5,7" e "0,2"; quest'ultima è significativamente meno discriminata delle altre.

4.4 Discussione

I test di percezione categoriale hanno confermato che, dato un continuum di variazione di allineamento e altezza tonale del picco (e della fase decrescente successiva), gli accenti H*+L e H* possono essere identificati. Nell'esperimento descritto in questo articolo, gli stimoli differivano di 15 ms in allineamento e, coerentemente, di 4,5 Hz in altezza tonale per ogni passo di manipolazione. Sulla base dei risultati di identificazione, il cambio di categoria è realizzato in due passi intermedi (stimoli "2" e "3"), ossia in tre passi di manipolazione che corrispondono a 45ms e 13.5 Hz di differenza, rispettivamente per allineamento e altezza tonale.

Per quanto riguarda la discriminazione, i risultati hanno mostrato che nel caso di coppie di stimoli adiacenti nel continuum di manipolazione, ossia di coppie in cui gli stimoli differivano di 15 ms e 4.5 Hz (cfr. risultati per la Serie-1), i soggetti non sono stati in grado di discriminare. D'altra parte, quando sono stati presentati stimoli che mostravano maggiori differenze di allineamento e altezza tonale, i soggetti sono riusciti e discriminare gli stimoli nelle coppie, anche se questo non è necessariamente coinciso con l'emergere di un chiaro picco di discriminazione nei risultati. Per la Serie-3 (45 ms e 13.5 Hz di differenza in allineamento e altezza tonale), a causa della percentuale di discriminazione globalmente elevata, non è chiaramente visibile alcun picco di discriminazione (l'intervallo di variazione nella discriminazione delle coppie è piuttosto limitato, anche se corrisponde a valori globalmente elevati). Infatti, come è stato confermato dalle statistiche sui valori di *d′*, non vi è alcuna differenza significativa nella sensibilità alle differenze interne alle coppie. Tuttavia, la coppia di stimoli che sembra essere discriminata meglio, almeno considerando le percentuali di discriminazione e alcuni risultati statistici, include uno degli stimoli corrispondenti al confine tra categorie nei risultati di identificazione (stimolo "3"), anche se solo quando esso sia confrontato con un estremo del continuum (stimolo "6", e non stimolo "0"). Questo potrebbe suggerire che lo stimolo "3" appartenga alla stessa categoria alla quale appartiene lo stimolo "0" (o sia più simile allo stimolo "0"), piuttosto che alla categoria alla quale appartiene lo stimolo "6". In modo analogo, l'altro stimolo corrispondente al passaggio di categoria in base ai risultati di identificazione, lo stimolo "2", viene discriminato piuttosto bene se abbinato con lo stimolo "5", come se i due stimoli non facessero parte della stessa categoria. Tuttavia, la discriminazione della coppia "2,5" non è significativamente diversa da altre, in particolare dalla discriminazione delle coppie "4,7" o "5,8", che dovrebbero invece includere i membri appartenenti ad una stessa categoria. Quindi, non sembra possibile soste-

nere che un picco di discriminazione è stato individuato (in corrispondenza della coppia "3,6"), perché la differenza nei risultati non è sempre significativa; inoltre, è evidente che i risultati non indicano che le coppie che coinvolgono stimoli al confine tra le categorie siano discriminate meglio delle altre.

I risultati indicano, invece, una differenza abbastanza chiara nella discriminazione per la Serie-2 (variazione di 30 ms in allineamento e 9 Hz in altezza tonale per gli stimoli appaiati). Nel caso di questo blocco di stimoli, le percentuali di discriminazione e i valori di sensibilità alle differenze secondo il *d'* risultano essere leggermente più bassi e maggiormente differenziati che per la Serie-3; inoltre nel grafico dei risultati è visibile un plateau che riguarda le coppie che coinvolgono gli stimoli al confine tra le due categorie in base al test di identificazione, ossia gli stimoli "2" e "3". Questa serie offre la migliore corrispondenza tra i risultati di identificazione e discriminazione, anche se mette in evidenza alcuni aspetti abbastanza interessanti. Prima di tutto, non è chiaro perché gli stimoli al confine tra categorie debbano essere discriminati solo in alcune delle coppie in cui sono coinvolti, se sono visti come stimoli ambigui e non, piuttosto, come possibili "cattivi esemplari" di una specifica categoria (v. esistenza dei prototipi nel caso delle categorie intonative – Gili Fivela, 2012). Ad esempio, il punteggio più basso per la Serie-2 si ottiene per la coppia "0,2", mentre valori di discriminazione più alti si riscontrano per la coppia "2,4". Questo suggerisce che lo stimolo "2" sia percepito come più simile ai membri di una delle categorie, in particolare a quelli della categoria a cui appartiene lo stimolo "0". In secondo luogo, un risultato sorprendente, correlato a quello appena menzionato, riguarda le differenze osservate a seconda della serie considerata (ossia della differenza acustica tra gli stimoli presentati in coppia). Infatti, dai risultati ottenuti per la Serie-2 risulta che lo stimolo "3" è chiaramente discriminato rispetto stimolo "1" e un po' meno discriminato rispetto allo stimolo "5" (si confrontino i risultati per "1,3" e "3,5"). Tuttavia, prendendo in esame i risultati per la Serie-3, lo stimolo "3" sembra essere più facilmente discriminato rispetto allo stimolo "6" che allo stimolo "0" (si confrontino i risultati per "3,6" e "0,3"). Quindi, secondo la stessa logica, in base ai risultati della Serie-2 lo stimolo "3" sembra essere percepito come più simile agli stimoli della categoria alla quale appartiene lo stimolo "5", mentre nella Serie-3 viene percepito come più simile ai membri dell'altra categoria, che è quella alla quale appartiene lo stimolo "0". In questa fase dell'indagine non è chiaro il perché esistano delle differenze simili a seconda della serie considerata. Tuttavia è evidente che gli stimoli sono percepiti in modo diverso all'interno delle categorie e che gli stimoli al confine tra categorie possono essere considerati più o meno simili ai membri delle categorie indagate, diversamente da quanto previsto in base alla Quantal Theory of speech (v. anche Gili Fivela, 2012).

Rispetto alle questioni metodologiche che erano alla base di questa indagine, è possibile sottolineare che l'identificazione di stimoli che variano in passi di 15 ms di allineamento e 4.5 Hz di frequenza fondamentale sembra essere effettuata facilmente. Questo conferma i risultati discussi in letteratura ed è in linea con lavori precedenti, svolti sull'italiano di Pisa, che hanno mostrato che i soggetti sono in grado di identificare stimoli appartenenti ad un continuum di variazione molto simile quello descritto in questa sede e, inoltre, sono in grado di imitarli, producendo esempi di due categorie distinte (Gili Fivela,, 2008, 2009)[3]. Per quanto riguarda la discriminazione, i risultati migliori si ottengono quando la differenza acustica tra gli stimoli nella coppia è circa 2/3 della differenza necessaria per il cambio di categoria: nel nostro caso, 30 ms e 9 Hz di differenza per gli stimoli nelle coppie, dato un cambio di categoria che avviene in circa 45 ms e 13,5 Hz. Tuttavia, è importante sottolineare che "risultati migliori" o una "discriminazione migliore" in realtà non corrispondono ad un chiaro picco di discriminazione, ma piuttosto ad una regione di discriminazione caratterizzata da valori leggermente superiori rispetto a quelli ricavati per

[3] Gli stimoli usati nei lavori del 2008 e 2009 erano ambigui anche per le caratteristiche di altezza tonale. Questo aspetto rende i risultati di imitazione dei soggetti ancora più indicativi della presenza di una chiara identificazione di due categorie accentuali diverse.

le coppie circostanti. Di fatto, quindi, i risultati sono perfettamente in linea con le osservazioni riscontrate in letteratura, e confermano che l'identificazione possa essere un compito più facile rispetto alla discriminazione (almeno nel caso di stimoli molto simili tra loro).

In ogni caso, non sembra opportuno riportare l'assenza di discriminazione osservata per la Serie-1 a problemi connessi alla soglia di differenza appena percepibile (JND). Infatti, nel caso in discussione, una differenza di 15 ms e 4.5 Hz non è stata sufficiente ai soggetti per discriminare, ma è stata abbastanza perché loro "notassero la differenza" in un compito più linguistico come quello di identificazione. In particolare, sembra molto improbabile che si possa essere in grado di sfruttare linguisticamente (nell'identificazione) un'informazione che non può essere percepita (in discriminazione). Di fatto, sembrerebbe eventualmente più naturale il contrario. D'altra parte, una differenza di 45 ms e 13.5 Hz sembra essere troppo elevata, dato che è correlata a un aumento complessivo della capacità di discriminazione, e sembra piuttosto interferire con la comparsa di picchi di discriminazione categoriale in corrispondenza di stimoli al confine tra categorie. Inoltre, la presenza di una regione di discriminazione piuttosto che di un picco, nel migliore dei casi, ossia per la Serie-2, supporta l'idea della presenza di confini di categoria non discreti (*fuzzy*), che nel caso dell'intonazione potrebbero essere dovuti anche ai diversi significati veicolati dagli eventi intonativi (v. Gili Fivela, 2012).

Pertanto, a parte il possibile impatto metodologico dei risultati descritti in questa sede, le osservazioni riportate supportano l'ipotesi di una differenziazione interna alle categorie che possa in qualche modo influire sulla percezione categoriale quando questa sia verificata con metodi tradizionali, come nel caso dei test di identificazione e discriminazione previsti nel paradigma della percezione categoriale.

5. RIFLESSIONI CONCLUSIVE

L'esperimento descritto in questo articolo dimostra che la scelta del passo di manipolazione usato ha un forte impatto sui risultati che si possono ottenere in relazione alla percezione categoriale, soprattutto per il test di discriminazione. Infatti, il compito di identificazione si conferma come un test robusto, probabilmente perché richiede un'elaborazione di tipo linguistico, mentre quello di discriminazione rappresenta la fase più problematica. Nell'esperimento descritto in questa sede, i soggetti si dimostrano in grado di identificare due categorie dato un continuum fonetico di stimoli. Gli stessi soggetti, invece, non discriminano le coppie di stimoli (ad esempio, non discriminano nella Serie-1, composta di stimoli che differiscono di 15 ms e 9 Hz, adiacenti lungo il continuum di variazione) oppure discriminano in modo così evidente che la presenza di un picco positivo nei loro giudizi di discriminazione non è apprezzabile (v. Serie-3, in cui gli stimoli differiscono per 45 ms e 13.5 Hz). Di fatto, la discriminazione migliore si ha quando la differenza di allineamento è circa 2/3 (30ms) rispetto all'intervallo richiesto per il cambio di categoria (45ms). Anche in questo caso, però, non emerge un picco di discriminazione, ma un plateau. È quindi evidente che i risultati non indicano una chiara percezione categoriale, ma suggeriscono l'esistenza di una "regione" di discriminazione migliore rispetto a valori di discriminazione complessivamente elevati. L'uso di test statistici diversi potrebbe forse fornire indicazioni differenti rispetto alla significatività dei dati, ma è comunque chiaro che, rispetto agli accenti considerati, non emerge un picco di discriminazione (anche alla luce del fatto che, nel caso di risultati sperimentali evidentemente a favore della percezione categoriale, in cui il picco di discriminazione è presente, la differenza interna ai dati è molto elevata; v. percezione dei toni di confine in Gili Fivela, 2008).

In ogni caso, non si ritiene che i risultati descritti in questa sede rappresentino una prova a sfavore dell'esistenza di due categorie intonative distinte, visto che dati di produzione, di imitazione e di identificazione riportati in altra sede (Gili Fivela, 2008) mostrano che i parlanti usano i due pattern con chiare differenze funzionali e fonetiche. Piuttosto i risultati sembrano indicare l'esistenza di confini non discreti per le categorie intonative in esame (così come per le vocali, v. Gili Fivela, 2012) e la possibilità di rica-

vare indicazioni procedurali per il disegno dei test di discriminazione (benché si tratti di un'osservazione da verificare con ulteriori indagini). In particolare, la differenza ottimale tra gli stimoli delle coppie da discriminare sembra possa corrispondere a circa 2/3 dell'intervallo richiesto per il cambio di categoria nell'identificazione.

Rispetto alla possibilità di identificare una JND per le differenze di sincronizzazione di eventi tonali, ci associamo pienamente alle considerazioni di House (1997), sottolineando che in casi analoghi a quelli qui studiati, i risultati in percezione non possono essere semplicemente riportati alla presenza di soglie psicofisiche. Basti riflettere su quanto emerso in questi esperimenti e in particolare sul fatto che, benché i soggetti non sembrino discriminare stimoli che differiscono in allineamento per 15 ms e in altezza tonale per 4.5 Hz, percepiscono differenze utili a ottenere risultati significativi nel compito di identificazione, notoriamente caratterizzato dal riferimento a risorse cognitive di ordine superiore. Anche se è noto che i test di discriminazione tradizionali rappresentano compiti squisitamente percettivi, non è verosimile che i soggetti usino risorse cognitive di ordine superiore (nell'identificazione) per elaborare le differenze tra stimoli che non possono percepire come diversi (nella discriminazione, in termini di superamento della differenza minima percepibile, JND).

Infine, sottolineiamo l'importanza di alcuni sviluppi futuri della ricerca descritta in questa sede. In particolare, è importante: 1) effettuare una verifica delle ipotesi alla base di questo studio su stimoli in cui anche la durata sia modificata coerentemente con gli altri parametri; 2) realizzare dei test di discriminazione con stimoli ancor meno diversi tra loro, per verificare se la possibile indicazione procedurale data in questa sede possa essere resa ancora più precisa; 3) controllare che i risultati non varino a seconda del test statistico e delle misure scelte.

BIBLIOGRAFIA

Bruce, G (1977). *Swedish Word Accents in Sentence Perspective*. Gleerup: CWK.

D'Imperio, M. & D. House (1997). Perception of questions and statements in Neapolitan Italian. In *Proceedings of the EuroSpeech*: 251-254.

Frascarelli, Mara (2000). The Syntax-Phonology Interface in Focus and Topic Constructions in Italian. *Studies in Natural Language and Linguistic Theory* 50, Dordrecht, Kluwer Academic Publishers.

Frota S. (2012). A focus intonational morpheme in European Portuguese: Production and perception, in Prieto & Alcibar (eds.) *Prosody and meaning*, Mouton de Gruyter's Trends in Linguistics.

Fry, D.B., A.S. Abramson, P.D. Eimas & A.M. Liberman (1962). The identification and discrimination of synthetic vowels. *Language and Speech* 5: 171-189.

Gili Fivela, B. (1999). The prosody of left-dislocated topicalized constituents in Italian read speech. In Proceedings of the EuroSpeech, Budapest: 1, 531-534.

Gili Fivela, B. (2004). *The phonetics and phonology of intonation: The case of Pisa Italian*. Ph.D. dissertation, Scuola Normale Superiore, Pisa.

Gili Fivela, B. (2005). La percezione degli accenti: il ruolo dell'allineamento e dello 'scaling' dei bersagli tonali. In Proceedings Convegno AISV: 313-326. Torriana (RN), EDK.

Gili Fivela, B. (2006).The coding of target alignment and scaling in pitch accent transcription. *Italian Journal of Linguistics* 18(1): 189-221.

Gili Fivela, B. (2008). *Intonation in Production and Perception: The Case of Pisa Italian*. Alessandria, Edizioni dell'Orso.

Gili Fivela, B. (2009). From production to perception and back: An analysis of two pitch accents. In: S.Fuchs, H.Lœvenbruck, D.Pape & P.Perrier (eds.), *Some Aspects of Speech and the Brain*, 363-405. Peter Lang, Frankfurt am Main.

Gili Fivela, B. (2012). Meanings, shades of meanings and prototypes of intonational categories, in Prieto & Alcibar (eds.) *Prosody and meaning*, Mouton's Trends in Linguistics, 197-237.

Gelfand, S. (1981). *Hearing, an introduction to psychological and physiological acoustics*. NY: Marcel Dekker, Inc.

Gescheider, G. (1976). *Psychophysics: Method and Theory*. Lawrence Erlbaum Ass., Hillsdale, NJ.

House, D. (1990). *Tonal Perception in Speech*. Lund: Lund University Press.

House, D. (1997). Perceptual thresholds and tonal categories. *Phonum* 4: 179-182, Department of Phonetics, University of Umeå.

Kohler, K. (1987). Categorical pitch perception. In *Proceedings ICPhS*, 5: 331-333. Tallin.

Liberman, A.M., K.S.Harris, H.S. Hoffman & B.C. Griffith (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology* 54, 5: 358-368.

Macmillan, N.A. & D.C. Creelman (1991). *Detection Theory:A User's Guide*. NY: CUP.

Massaro, D.W. (1998). Categorical perception: Important phenomenon or lasting myth? In *Proceedings of the 5th ICSLP*, Sydney, Australia, 2275-2279.

Niebuhr, O. & Kohler, K. (2004). Perception and cognitive processing of tonal alignment in German. *Proc. Int. Symp. on Tonal Aspects of Languages: Emphasis on Tone Languages*, Beijing, 155-158.

Pierrehumbert, J. & S. Steele (1989). Categories of tonal alignment in English. *Phonetica 4*6: 181-96.

Prieto, P., M. D'Imperio & B.Gili Fivela (2005). Pitch accent alignment in Romance: primary and secondary association with metrical structure. *Language and Speech*, 48:359-396.

Remijsen, B. & V.J. van Heuven (1999). Gradient and categorical pitch dimensions in Dutch: diagnostic test. In *Proceedings ICPhS*, 1865-1868. San Francisco.

Savino, M. & M. Grice (2011), The perception of negative bias in Bari Italian question, in S.Frota et al.(ed), *Prosodic categories:Production, perception and comprehension*,Springer:Dordrecht,187-206.

Schneider, K., B.Lintfert, G. Dogil & B. Möbius (2006). Phonetic grounding of prosodic categories. In S.Sudhoff et al. (eds.), *Methods in Empirical Prosody Research*, 335-362. Berlin: De Gruyter.

Stevens, K.N. (1972). The quantal nature of speech: evidence from articulatory-acoustic data, in E. David Jr. and P.B. Denes (eds.), *Human Communication: A Unified View*, NY: McGraw-Hill, 51-66.

Stevens, K.N. (1989). On the quantal nature of speech. *J. Phonetics,* 17: 3-45.

Stevens, K. & S. Keyser (2010). Quantal theory, enhancement and overlap. *J. Phonetics,* 38,1: 10-19.

Svantesson J.O. & House D. (1996). Tones and non-tones in Kammu dialects. Proceedings of Fonetik 96, Swedish Phonetics Conference, TMH-QPSR 2/1996, 85-87.

Vanrell, M. (2006). A scaling contrast in Majorcan Catalan interrogatives. In *Speech Prosody*, 807-10. Dresden

DISFONIA FUNZIONALE: CORRELAZIONE CON SINTOMI DEPRESSIVI E D'ANSIA

Chiara Chialva[a], Giulia Bertino[a], Silvia Migliazzi[a], Vera Abbiati[b], Valentina Ciappolino[b], Roberto Pagani[b], Natascia Brondino[b], Edgardo Caverzasi[b], Marco Benazzo[a]

[a] Struttura Complessa di Otorinolaringoiatria, Università di Pavia, Fondazione IRCCS Policlinico San Matteo, Pavia
[b] Centro Interdipartimentale di Ricerca sui Disturbi di Personalità (CIRDIP), Università di Pavia

bruna.chiara@libero.it; giulia.bertino@tin.it; natascia.brondino@libero.it

RIASSUNTO

Obiettivo: la disfonia funzionale rappresenta l'8% di tutti i casi di disfonia. La sua eziologia non è stata ancora completamente chiarita, tuttavia sembrano avere un ruolo chiave fattori psicologici e personologici. L'obiettivo del presente studio è indagare l'eventuale presenza di correlazioni significative tra la disfonia funzionale e i sintomi depressivi e ansiosi.
Materiali e metodi: quarantadue pazienti sono stati reclutati consecutivamente presso l'Ambulatorio di Stroboscopia della Clinica Otorinolaringoiatrica IRCCS Fondazione Policlinico San Matteo di Pavia. Tutti i soggetti sono stati sottoposti ad una visita otorinolaringoiatrica, ad un esame stroboscopico e ad un inquadramento logopedico della voce. Tutti i soggetti hanno compilato tre questionari: *Voice Handicap Index* (VHI), *Beck Depression Inventory* (BDI) e *Zung Self-Rating Anxiety Scale* (SAS). I criteri di inclusione erano: comprensione della lingua italiana parlata e scritta, età>18 anni, assenza di patologie organiche rilevanti in atto, assenza di patologia psichiatrica pre-esistente, assenza di problematiche organiche responsabili della disfonia, assenza di disfonia spasmodica.
Risultati: il grado globale di alterazione della voce correla significativamente con il punteggio totale alla VHI, con le sue sottoscale F, P e con il punteggio totale al BDI e alla SAS. Il punteggio totale VHI correla positivamente con il punteggio BDI e SAS. I soggetti clinicamente depressi presentano un punteggio alla VHI e alle sottoscale F, E e P superiore rispetto ai pazienti non clinicamente depressi. Gli individui con ansia clinicamente rilevante presentano punteggi più elevati alla VHI e alle sottoscale F e P, ma non alla sottoscala E.
Conclusioni: lo studio evidenzia un'elevata prevalenza di sintomi depressivi e ansiosi in un campione di soggetti con disfonia funzionale. Tale aspetto deve essere preso in considerazione, favorendo una maggiore presa in carico del soggetto anche dal punto di vista psicologico.

1. INTRODUZIONE

Con il termine disfonia s'intende un'alterazione qualitativa della voce caratterizzata dalla compresenza di sintomi acustici e non acustici (p.e. alterazione del ritmo d'eloquio, della postura). Nell'ambito della disfonia, si distinguono due principali categorie: la disfonia organica (correlabile a una patologia laringea) e quella funzionale (in assenza di alterazioni laringee). Nel presente studio definiremo, quindi, non organiche tutte le forme di disfonia in cui la laringe sia normale per morfologia e motilità e, all'interno di tale gruppo, prenderemo in esame le disfonie funzionali, definibili come "alterazioni della voce – o meglio della sua produzione (Magnani, 2005) – legate prevalentemente a un uso scorretto del sistema muscolare coinvolto nella vociferazione" (Ruoppolo, 2010). La disfonia funzionale è una dia-

gnosi di esclusione e deve essere fatta solo dopo una valutazione specialistica della laringe da parte di un otorinolaringoiatra o di un foniatra (Wilson, Deary, Scott, and MacKenzie, 1995).

Le disfonie funzionali possono essere classificate in due categorie: ipocinetiche e ipercinetiche.

Le prime presentano, come elemento comune, un deficit di adduzione cordale per ipotonia di uno o più muscoli laringei intrinseci (Fussi e Magnani, 2003; Le Huche e Allali, 1990). Dal punto di vista obiettivo, si può osservare un deficit di adduzione cordale posteriore per ipotonia della muscolatura interaritenoidea, oppure una glottide ovalare per ipotonia dei muscoli vocali (tiroaritenoidei) o, infine, un deficit di adduzione lineare per la concomitante ipotonia degli interaritenoidei e dei tiroaritenoidei e/o per l'insufficiente stabilizzazione dell'adduzione cordale (deficit dei cricoaritenoidei laterali). Nelle forme ipercinetiche si osserva invece un utilizzo della glottide in eccesso di tensione muscolare e frequentemente si rileva un'iperadduzione delle strutture sovraglottiche (completa, laterale oppure anteroposteriore).

Sebbene tale rigorosa classificazione, ancora ampiamente utilizzata nella pratica clinica, sia stata posta in discussione negli anni più recenti da vari autori (Ruoppolo, 2010; Rammage et al., 2000; Le Huche e Allali, 1990), non è disponibile attualmente una classificazione razionale ed universalmente condivisa delle disfonie. Ad esempio secondo il DSM IV (*American Psychiatric Association*, 2004) la disfonia funzionale è da considerarsi un disturbo somatoforme, più specificatamente, disturbo da conversione con sintomi o deficit motori. La caratteristica comune di tali disturbi è la presenza di sintomi fisici che suggeriscono la presenza di una malattia organica ma che non sono pienamente giustificati da una condizione medica generale, o da un altro disturbo mentale o dall'effetto diretto di una sostanza. Questi sintomi causano disagio clinicamente significativo o compromissione nella vita sociale, lavorativa o in un'altra area di funzionamento (*American Psychiatric Association*, 2004).

La disfonia funzionale rappresenta l'8% di tutti i casi di disfonia (Willinger et al., 2005). Inoltre, si stima che la disfonia funzionale interessi dal 10 al 40% degli accessi alle cliniche ORL e foniatriche per problemi di voce (Roy, 2003). In molti studi la disfonia funzionale si correla in modo significativo al tipo di professione svolta, in particolare a quella di insegnante (Herrington-Hall et al, 1988). La maggior parte dei pazienti sono donne di giovane/media età, in buono stato di salute. Circa il 40% dei casi vengono in prima istanza misdiagnosticati e trattati con antibioticoterapia, supponendo la presenza di una laringite.

Inoltre sembrano implicati quali fattori determinanti le alterazioni posturali e respiratorie, l'uso professionale della voce, la gestione delle dinamiche di velocità e di intensità dell'eloquio, la personalità e il profilo psicologico (Patrocinio, Trittola, 2010). I quadri di ipercinesia ed ipocinesia, in particolare, sembrerebbero essere essi stessi correlati con l'assetto personologico dei singoli individui (Ruoppolo, 2010; Schindler e Limarzi, 2002). L'eziologia della disfonia funzionale è ancora poca chiara, ma sembrano giocare un ruolo importante fattori psicosociali che, secondo Morrison (Morrison et al. 1986), in una via finale comune potrebbero determinare un uso anomalo della muscolatura volontaria associata alla fonazione.

Per le ragioni sopradescritte, un certo grado di disadattamento emozionale può essere rilevato, nella pratica clinica, nella disfonia funzionale. In questi pazienti non è raro osservare una risoluzione del disturbo foniatrico, in seguito a rieducazione logopedica, ma è altrettanto comune identificare un significativo tasso di recidiva (>10%) a breve e a lungo termine

(Lauriello et al., 2003). Questo fenomeno può essere ascritto a un'inadeguata valutazione e gestione degli aspetti emotivi del paziente. Inoltre, è stato dimostrato che i pazienti con disfonia funzionale presentano scarse abilità adattative e di controllo emotivo ma alti livelli di ansia, disturbi psico-somatici e introversione (Roy et al., 1997).

Al momento attuale, il ruolo giocato dall'ansia e dalla depressione nella disfonia funzionale non è tuttavia ancora definito in modo chiaro.
L'obiettivo che si pone questo studio è di evidenziare delle correlazioni significative tra la disfonia funzionale e i sintomi depressivi e ansiosi, utilizzando un campione di pazienti affetti da disfonia funzionale.

2. MATERIALI E METODI

Quarantadue pazienti sono stati reclutati consecutivamente da giugno ad ottobre 2011 presso l'Ambulatorio di Stroboscopia della Clinica Otorinolaringoiatrica IRCCS Fondazione Policlinico San Matteo di Pavia. Tutti i soggetti sono stati sottoposti ad una visita otorinolaringoiatrica, ad un esame stroboscopico con valutazione della chiusura glottica, dell'ampiezza, della periodicità e della simmetria dell'onda mucosa. È stato inoltre effettuato un inquadramento logopedico della voce che comprende: valutazione clinica dello stile di eloquio (intensità, fluenza, risonanza, articolazione); valutazione clinica della respirazione; valutazione clinica della postura generale e laringea, tramite palpazione (Ricci Maccarini et al., 2010; Magnani, 2010), misurazione in secondi del Tempo Massimo Fonatorio (TMF) sulla vocale /a/; valutazione acustica della voce tramite spettrogramma e classificazione di Yanagihara (Ricci Maccarini et al 2010), effettuata sulla vocale /a/ sostenuta per almeno quattro secondi senza interruzione di sonorità; valutazione percettiva della voce tramite la scala GIRBAS (Hirano,1981; rivista Dejonckere, 1996), che permette di valutare i seguenti parametri assegnando un punteggio che va da 0 (condizione di eufonia) a 3 (alterazione grave):

- G: grado globale di alterazione della voce;
- I: indice d'instabilità nella voce;
- R: voce rauca, presenza d'irregolarità nella vibrazione delle corde vocali percepita nella fonazione;
- B: voce soffiata, presenza di fuga d'aria udibile nella voce;
- A: voce astenica, debolezza o mancanza di forza nella voce;
- S: voce pressata, indice di sforzo vocale.

I criteri d'inclusione erano: comprensione della lingua italiana parlata e scritta, età>18 anni, assenza di patologie organiche rilevanti in atto (cancro, demenza, infezioni), assenza di patologia psichiatrica pre-esistente, assenza di problematiche organiche responsabili della disfonia, assenza di disfonia spasmodica. Dati anamnestici relativi a tutti pazienti sono stati raccolti durante la prima visita. Un consenso informato scritto è stato ottenuto da ciascun paziente. Lo studio è stato condotto in accordo con la Dichiarazione di Helsinki.

2.1 Questionari

Ogni paziente ha compilato il *Voice Handicap Index* (VHI; Jacobson, Johnson, Grywalski et al, 1997), questionario autosomministrato a 30 item che permette di definire la percezione della gravità della problematica disfonica da parte del paziente, classificandola in disabilità assente, lieve, media, grave o totale. È costituito da tre sottoscale:

- scala F (*Functional subscale*) che valuta l'impatto delle problematiche vocali sulle normali attività quotidiane;
- scala E (*Emotional subscale*), che valuta l'impatto psicologico della disfonia;
- scala P (*Physical subscale*), relativa alla percezione da parte del paziente delle caratteristiche della propria voce.

La severità nonché la presenza di sintomi depressivi sono stati valutati tramite *Beck Depression Inventory* (BDI). Trattasi di un questionario auto-somministrato a 21 item (Beck, 1961) valutante l'intensità della sintomatologia dell'umore. Una soglia di cut-off pari a 18 è stata utilizzata per indicare la presenza di uno stato depressivo clinicamente significativo.

Informazioni relative a sintomi d'ansia sono state raccolte tramite la *Zung Self-Rating Anxiety Scale* (SAS, Zung, 1971), questionario autosomministrato a 20 item. Gli item sono valutati su una scala Likert a 4 punti in base a se un individuo nelle ultime due settimane ha provato ogni specifico sintomo "quasi mai o raramente" (rating=1), "qualche volta" (2), "spesso"(3), o "quasi sempre" (4). Una soglia di cut-off pari a 36 (Zung, 1971) è stata utilizzata per definire la presenza di ansia clinicamente significativa.

2.2 Analisi Statistica

I dati sono presentati come media±deviazione standard o come percentuale, a seconda del caso. Tutti i dati rilevati sono stati analizzati tramite Kolmogorov-Smirnoff per valutare la deviazione da una distribuzione normale e tramite Levene's test per l'omogeneità della varianza. Il coefficiente di correlazione di Pearson o Spearman è stato valutato per stabilire l'esistenza di una correlazione tra variabili. Le differenze relative ai parametri vocali tra pazienti clinicamente significativi per ansia e depressione e soggetti non affetti da patologia psichica sono state analizzate tramite il test del Chi-quadro o il test di Student. Un valore p a due code inferiore a 0.05 è stato considerato statisticamente significativo. Tutte le analisi sono state effettuate tramite SPSS 16.0 (SPSS, Chicago, IL).

3. RISULTATI

L'81% (34) del campione è costituito da donne, il 78.6% (33) presenta una disfonia ipercinetica. La percentuale di professionisti vocali è del 50% (21). L'età media è 48.43±14.83.

La tabella 1 illustra le caratteristiche generali del campione, riportando la media e le deviazioni standard per ogni parametro esaminato nel presente studio. Nella tabella 2, invece, sono presentate tutte le correlazioni statisticamente significative.

Nello specifico, è stata osservata una correlazione negativa tra Tempo Massimo Fonatorio (TMF) e frequenza fondamentale (frequenza di vibrazione delle corde vocali), misurata allo spettrogramma della vocale /a/ sostenuta per almeno quattro secondi senza interruzioni di sonorità (figura 1); in altre parole, si è rilevato che all'aumentare della frequenza corrisponde una diminuzione della durata fonatoria sia nei pazienti di sesso maschile che nei pazienti di sesso femminile.

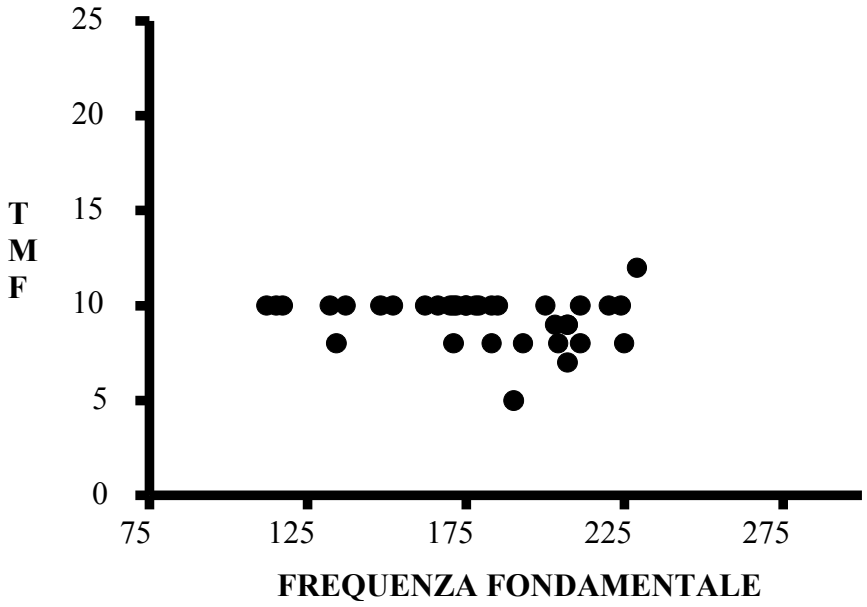


Figura 1: correlazione tra Tempo Massimo Fonatorio, espresso in secondi, e frequenza fondamentale (in Hz), rilevata allo spettrogramma della vocale /a/ sostenuta per almeno quattro secondi senza interruzione di sonorità.
All'aumentare della frequenza fondamentale (frequenza di vibrazione delle corde vocali, espressa in cicli di ondulazione per secondo) corrisponde una diminuzione della durata fonatoria.

Si è osservata, inoltre, una correlazione negativa tra l'età del soggetto e il punteggio ottenuto alla scala E del VHI. Il grado globale di alterazione della voce della scala GIRBAS correla significativamente con il punteggio totale al VHI, con le sottoscale F, P e con il punteggio totale al BDI e alla SAS. La presenza di voce astenica, debolezza o mancanza di forza nella voce valutata tramite la GIRBAS correla significativamente con il punteggio totale alla VHI, con le sottoscale F e P. Il punteggio totale VHI correla positivamente con il punteggio BDI e SAS. La medesima correlazione si riscontra per le sottoscale F e P del VHI. Utilizzando il cut-off di 18 al BDI la percentuale di affetti da depressione clinica è pari al 16.70% (7) del campione (Figura 2).

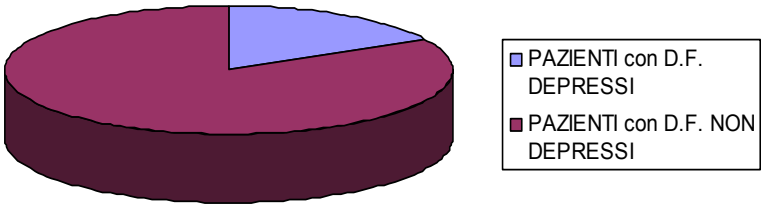


Figura 2: correlazione tra la disfonia funzionale e i sintomi depressivi

Le differenze riguardanti i parametri vocali tra i pazienti con patologia psichica clinicamente significativa e pazienti non clinicamente depressi o ansiosi sono riportate in tabella 3. Da tale confronto emerge che i soggetti depressi presentano un punteggio alla VHI e alle sottoscale F, E e P superiore rispetto ai pazienti non clinicamente depressi.

La soglia di cut-off fornita da Zung per la SAS determina la positività per ansia clinicamente rilevante nel 35.70% (15) del campione (Figura 3). Gli ansiosi presentano punteggi più elevati al punteggio totale del VHI e alle sottoscale F e P, ma non alla sottoscala E. Non sono state rilevate differenze statisticamente significative tra i pazienti professionisti vocali e i pazienti che non usano la voce dal punto di vista professionale.

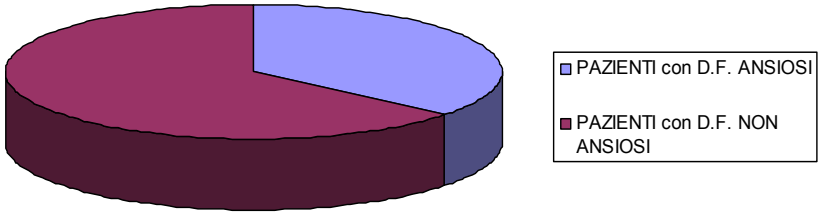


Figura 3: correlazione tra la disfonia funzionale e i sintomi ansiosi

4. DISCUSSIONE

Nel presente studio abbiamo valutato pazienti affetti da disfonia funzionale relativamente alla presenza di sintomi depressivi e ansiosi. A parità di gravità di disfonia, più del 16% del campione presenta una depressione di rilevanza clinica. Tale dato risulta inferiore rispetto allo studio di Willinger et al (Willinger et al, 2005) dove venivano rilevati sintomi depressivi nel 39% del campione. Tuttavia il nostro è uno studio preliminare ancora in svolgimento, con una numerosità campionaria limitata. Tale motivazione potrebbe essere anche alla base della differenza osservata tra i nostri risultati relativi all'ansia (oltre un terzo dei soggetti presenta un disturbo d'ansia) e quelli del sopracitato studio dove la prevalenza era del 20%. Non sono state riscontrate differenze significative in termini di ansia e depressione nelle due forme di disfonia considerate, anche in conseguenza del basso numero di soggetti affetti da disfonia ipocinetica.

La corrispondenza negativa tra il Tempo Massimo Fonatorio e la frequenza fondamentale potrebbe essere riconducibile al fatto che l'utilizzo di una frequenza acuta comporta una maggiore tensione longitudinale della muscolatura vocale, una maggiore fatica fonatoria con conseguente riduzione del TMF.

La correlazione significativa tra i punteggi al VHI e quelli al BDI e alla SAS è in accordo con i dati riportati in letteratura (Elam et al, 2010). Si può quindi ipotizzare che i tre questionari catturino costrutti sovrapponibili, ma tuttavia esistono anche alcune differenze. Infatti, se si considerano le sottoscale del VHI, la correlazione con BDI e SAS rimane significativa solo per la scala F e la P. La mancata correlazione tra la scala E, che valuta più specificatamente l'impatto psicologico del problema vocale, e la sintomatologia ansioso-

depressiva è in disaccordo con quanto riportato da Elam et al. Tali autori hanno considerato, a differenza del nostro studio, un campione di soggetti affetti da reflusso faringolaringeo. È tuttavia possibile ipotizzare che soggetti affetti da disfonia funzionale siano meno consapevoli degli aspetti psicologici sottesi a tale condizione, che verrebbero colti solo da questionari mirati, in quanto tali individui sarebbero maggiormente focalizzati sulla disabilità fisica.

Da sottolineare come dal nostro studio non emergano differenze significative tra i pazienti professionisti vocali e i non professionisti vocali per quanto riguarda né i sintomi depressivi e d'ansia né la valutazione dell'impatto psicologico della disfonia (scala E del VHI).

Alcuni studi hanno suggerito come la gravità percepita dei sintomi ansiosi e depressivi possa influenzare il decorso della disfonia (Dietrich et al, 2008). Nel presente lavoro, è stato selezionato un campione di pazienti disfonici di pari gravità. Come linea di ricerca futura, riteniamo utili studi prospettici con un campione di numerosità maggiore al fine di valutare la causalità o meno dei sintomi psicologici nella genesi della problematica vocale e l'impatto di questi ultimi sul decorso e la prognosi della disfonia.

Il nostro lavoro è uno studio preliminare e come tale ha le limitazioni connesse alla numerosità del campione e alla mancanza di un gruppo di controllo. Tuttavia, è emersa un'elevata prevalenza di sintomi depressivi e ansiosi in un campione di soggetti con disfonia funzionale. Tale aspetto deve essere preso in considerazione sia nella fase diagnostica che nel processo terapeutico, favorendo una maggiore presa in carico del soggetto anche dal punto di vista psicologico con eventuale intervento specialistico, offrendo così al paziente strategie e strumenti personalizzati che non si limitino esclusivamente al rilassamento o agli esercizi vocali.

Tabella 1. Caratteriste generali del campione

Parametri esaminati	**Media±SD o % (n)**
Disfonia ipercinetica	78.60(33)
ipocinetica	21.40(9)
Professionisti vocali	50 (21)
Frequenza fondamentale (espresso in Hertz)	181.33±38.69
Tempo Massimo Fonatorio (espresso in secondi)	9.50±2.11
Scala GIRBAS	
G	1.64±0.53
I	0.19±0.39
R	1.29±0.59
B	1.19±1.4
A	1.14±0.75
S	1.71±0.67
Presenza di diplofonia	38.10(16)
Classe Yanagihara	
Classe 1 (alterazione lieve)	47.60 (20)
Classe 2 (alterazione moderata)	50 (21)
Classe 3 (alterazione grave)	2.40 (1)
Intensità	
normale	21.40 (9)
forte	52.40 (22)
debole	26.20 (11)
Rappresentazioni armoniche (allo spettrogramma)	
normale	4.80 (2)
ampia	66.70 (28)
scarsa	28.60 (12)
Risonanza	
normale	97.60 (41)
rinofonia	2.40 (1)
Articolazione	
normale	85.70 (36)
ipoarticolazione	14.30 (6)
Fluenza	
normale	31 (13)
veloce	64.30 (27)
lenta	4.80 (2)
Accordo Pneumo-fonico	
adeguato	16.70 (7)
uso d'aria residua	81 (34)
frequenti atti respiratori	2.40 (1)
Tipo presa aerea a riposo	
nasale	97.60 (41)
orale	2.40 (1)

Tipo presa aerea in fonazione	
nasale	38.10 (16)
orale	61.90 (26)
Modalità respirazione in fonazione	
nasale	4.80 (2)
orale	95.20 (40)
Postura generale	
corretta	54.80 (23)
postura cranio-cervicale in iperestensione	31 (13)
postura cranio-cervicale in ipoestensione	14.60 (6)
Postura glottica	
adeguata	35.70 (15)
in attrazione superiore	64.30 (27)
Muscolatura del collo	
adeguato tono muscolare	33.30 (14)
marcata contrattura mm. sovraioidei	66.70 (28)
Postura delle spalle	
corretta	42.90 (18)
innalzate	42.90 (18)
infraruotate (posizione di chiusura)	14.30 (6)
VHI Totale	38.02±21.16
VHI Scala F	9.12±7.75
VHI Scala E	8.52±7.61
VHI Scala P	20.38±8.49
BDI	10±7.91
SAS	33.62±7.79

Tabella 2. Correlazioni statisticamente significative

Confronto tra i parametri esaminati	**Coefficiente di correlazione**	**P value**
TMF vs frequenza fondamentale	r=-0,456	p=0,005
Età del soggetto vs scala E (VHI)	r=-0,313	p=0,040
Parametro G della scala GIRBAS vs punteggio totale VHI	r=0,351	p=0,023
Parametro G della scala GIRBAS vs scala F (VHI)	r=0,430	p=0,004
Parametro G della scala GIRBAS vs scala P (VHI)	r=0,370	p=0,016
Parametro G della scala GIRBAS vs punteggio totale BDI	r=0,310	p=0,040
Parametro G della scala GIRBAS vs punteggio totale SAS	r=0,408	p=0,009
Parametro A della scala GIRBAS vs punteggio totale VHI	r=0,386	p=0,011
Parametro A della scala GIRBAS vs scala F (VHI)	r=0,362	p=0,019
Parametro a della scala GIRBAS vs scala P (VHI)	r=0,362	p=0,018
Punteggio totale VHI vs punteggio totale BDI	r=0,394	p=0,011
Punteggio totale VHI vs punteggio totale SAS	r=0,369	p=0,019
Scala F (VHI) vs punteggio totale BDI	r=0,418	p=0,007
Scala F (VHI) vs punteggio totale SAS	r=0,384	p=0,014
Scala P (VHI) vs punteggio totale BDI	r=0,368	p=0,018
Scala P (VHI) vs punteggio totale SAS	r=0,441	p=0,004

Tabella 3. Confronto tra pazienti con patologia psichica clinicamente significativa e pazienti non clinicamente depressi o ansiosi tramite Chi Square test e Student's t-test

Parametri esaminati in p.ti con patologia psichica vs p.ti non depressi o ansiosi	**Media±SD p.ti con patologia psichica vs Media±SD p.ti non depressi o ansiosi**	**P value**
Punteggio tot. VHI soggetti depressi vs punt. tot. VHI soggetti non clinicamente depressi	59.86±26.12 vs 33.68±17.62	p<0.01
Scala F (VHI) soggetti depressi vs scala F (VHI) soggetti non clinicamente depressi	18.57±9.89 vs 7.38±5.76	p<0.01
Scala E (VHI) soggetti depressi vs scala E (VHI) soggetti non clinicamente depressi	14.29±10.73 vs 7.24±6.48	p<0.01
Scala P (VHI) soggetti depressi vs scala P (VHI) soggetti non clinicamente depressi	27±9.02 vs 19.06±7.98	p<0.01
Punteggio tot. VHI soggetti ansiosi vs punt. tot. VHI soggetti non clinicamente ansiosi	48.60±23.73 vs 32.52±17.97	p<0.01
Scala F (VHI) soggetti ansiosi vs scala F (VHI) soggetti non clinicamente ansiosi	13.57±8.54 vs 6.80±6.04	p<0.01
Scala P (VHI) soggetti ansiosi vs scala P (VHI) soggetti non clinicamente ansiosi	24.13±8.85 vs 18.24±7.98	p<0.01
Scala E (VHI) soggetti ansiosi vs scala E (VHI) soggetti non clinicamente ansiosi	10.60±9.48 vs 7.48±6.26	p=0,240

BIBLIOGRAFIA

Dejonckere, P., Remacle, M., Fresnel-Elbaz, E. et al. (1996), Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements, Revue de Laryngologie-Otologie-Rhinologie, 3, 219-224

Dietrich, M., Verdolini Abbott, K., Gartner-Schmidt, J. et al. (2008), The frequency of perceived stress, anxiety, and depression in patients with common pathologies affecting voice. Journal of Voice, 22(4), 472-488

Elam, J.C., Ishman, S.L., Dunbar, K.B. et al. (2010), The relationship between depressive symptoms and Voice Handicap Index scores in laryngopharyngeal reflux, Laryngoscope, 120 (9), 1900-1903

Fussi, F. e Magnani, S. (2003), Lo spartito logopedico, Torino, Ed. Omega

Herrington-Hall, B.L., Lee, L., Stemple, J.C. et al. (1988), Description of laryngeal pathologies by age, sex, and occupation in a treatment-seeking sample, Journal of Speech and Hearing Disorders, 53(1), 57-64

Hirano, M. (1981), Clinical Examination of voice, New York, Springer-Verlag

Jacobson, B.H., Johnson, A., Grywalski, C. et al. (1997), The Voice Handicap Index (VHI): development and validation, American Journal of Speech-Language Pathology, 6, 66-70

Lauriello, M., Cozza, K., Rossi, A. et al. (2003), Psychological profile of dysfunctional dysphonia, Acta Otorhinolaryngologica Italica, 23(6), 467-473

Le Huche, F., Allali, A. (1990), La voix, Paris, Ed. Masson

Likert, R. (1932), A Technique for the Measurement of Attitudes, Archives of Psychology, 140, 1–55

Magnani, S. (2005), Curare la voce, Milano, Franco Angeli

Magnani, S. (2010), Valutazione posturale e respiratoria, in La voce-Fisiologia, patologia clinica e terapia, a cura di Schindler O., Padova, Piccin Edizioni, 123-132

Morrison, M.D., Nichel, H. e Rammage, L.A. (1986), Diagnostic criteria in functional dysphonia, Laryngoscope, 94, 1-8

Patrocinio, D. e Trittola, A. (2010), L'autovalutazione nei disturbi della voce, in La voce-Fisiologia, patologia clinica e terapia, a cura di Schindler O., Padova, Piccin Edizioni, 267-288

Rammage, L., Morrison, M.D. e Nichol H. (2000), Management of the Voice and Its Disorders, San Diego, Ed. Singular Thomson Learning

Ricci Maccarini, A. et al. (2010), Palpazione e manipolazione laringea, in La voce-Fisiologia, patologia clinica e terapia, a cura di Schindler O., Padova, Piccin Edizioni, 133-142

Roy, N., McGrory, J.J., Tasko, S.M., et al. (1997), Psychological correlates of functional dysphonia: an investigation using the Minnesota Multiphasic Personality Inventory, Journal of Voice, 11(4), 443-451

Roy, N. (2003), Functional dysphonia, Current Opinion in Otolaryngology & Head and Neck Surgery, 11 (3), 144-148

Ruoppolo, G. (2010), Disfonie non organiche, in La voce-Fisiologia, patologia clinica e terapia, a cura di Schindler O., Padova, Piccin Edizioni, 363-374

Schindler, O. e Limarzi M. (2002), Le disfonie disfunzionali, in Le disfonie: fisiopatologia ed aspetti medico-legali, a cura di Casolino D., Atti LXXXIX Congresso Nazionale SIO, Pisa, Ed. Pacini, 201-220

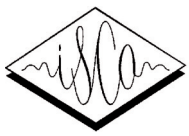
Willinger, U., Völkl-Kernstock, S. e Aschauer, H.N. (2005), Marked depression and anxiety in patients with functional dysphonia, Psychiatry Research, 134(1), 85-91

Wilson, J.A., Deary, I.J., Scott, S. et al. (1995), Functional dysphonia, British Medical Journal, 311 (7012), 1039-1040

Zung, W.W. (1971), A rating instrument for anxiety disorders, Psychosomatics, 12(6), 371-9

Organizzatori

Sponsor



Libro + CD
€ 25,00