AISV 2013

Multimodalità e Multilingualità: la Sfida più Avanzata della Comunicazione Orale

ABSTRACTS OF SUBMITTED PAPERS

9° Convegno Nazionale AISV
21-23 gennaio 2013

Università Ca' Foscari – Venezia

Invited speaker

THE STATISTICAL APPROACH TO SPEECH RECOGNITION AND NATURAL LANGUAGE PROCESSING: ACHIEVEMENTS AND OPEN PROBLEMS

Prof. Hermann Ney

RWTH Aachen University, Aachen - DIGITEO Chair, LIMSI-CNRS, Paris

The last 25 years have seen a dramatic progress in statistical methods for recognizing speech signals and for translating spoken and written language. This lecture gives an overview of the underlying statistical methods. In particular, the lecture will focus on the remarkable fact that, for these tasks and similar tasks like handwriting recognition, the statistical approach makes use of the same four principles: 1) Bayes decision rule for minimum error rate; 2) probabilistic models, e.g. Hidden Markov models or conditional random fields for handling strings of observations (like acoustic vectors for speech recognition and written words for language translation); 3) training criteria and algorithms for estimating the free model parameters from large amounts of data; 4) the generation or search process that generates the recognition or translation result.

Most of these methods had originally been designed for speech recognition. However, it has turned out that, with suitable modifications, the same concepts carry over to language translation and other tasks in natural language processing. This lecture will summarize the achievements and the open problems in this field.

Invited speaker

THE KTH TALKING HEAD IN SPACE - A VEHICLE FOR SITUATED MULTI-PARTY INTERACTION

Prof. Bjorn Granstrom

KTH Royal Institute of Technology, Stockholm

The KTH 3D talking head model has been used for more than ten years. It has been applied mostly in spoken dialogue systems and as a lip reading support for hard of hearing persons. In all cases the 3D model has been displayed on 2D computer displays. Recently we have experimented with the same model displayed on 3D facial masks, like the back-projected solution incorporated in a robotic head. This presentation will describe some of the new opportunities and challenges when moving from flat screens to a 3D rendering in the physical space of the user. While the original ambition with the model was to accurately display segmental articulation, we have recently expanded the capabilities of the head to prosodic and non-verbal signals. Several studies concern aspects of human-robot interaction in a multi-party setting, where e.g. selective gaze control is of great importance. Preliminary results from a comparison of our back projected head with a more conventional mechatronic robot head will also be presented as part of the EU IURO project.

Abstracts

La dimensione 3D del parlato e il problema del calcolo numerico nell'AG500

Massimo Stella [a)], Paolo Bernardini
Dipartimento di Matematica e Fisica "Ennio De Giorgi" - Università del Salento, via per Arnesano, 73100 Lecce (Italy)
Francesco Sigona, Antonio Stella, Mirko Grimaldi, Barbara Gili Fivela
Centro di Ricerca Interdisciplinare sul Linguaggio (CRIL) - Università del Salento, via Pappacoda 12, 73100 Lecce (Italy)

Abstract

Introduzione

Il rapido sviluppo tecnologico dell'ultimo trentennio ha portato numerose innovazioni nello studio della fonetica articolatoria e del parlato, offrendo tecniche sempre più accurate (cfr. Stone, 1997). Una di queste è l'Articulografo Elettromagnetico AG500 (Carstens Medizinelektronik), che permette la misurazione contemporanea della posizione e della velocità degli articolatori tramite la collocazione di sensori su lingua, denti, labbra o muscoli orofacciali in presenza di sei campi elettromagnetici variabili nel tempo.

L'AG500 è in grado di determinare le coordinate spaziali (x, y e z), l'azimuth (φ) e l'elevazione (θ) di un massimo di 12 sensori, ad una frequenza di 200Hz. I campi elettromagnetici sono generati da sei spire incastonate all'interno di un "cubo" di plastica, percorse da correnti alternate con frequenze tra 7500Hz e 13750Hz. Le correnti indotte sui sensori sono digitalizzate a 16 bit e quindi usate dal software per risolvere un sistema sovradeterminato di sei equazioni non lineari nelle incognite (x,y,z,φ,θ) (Kaburagi *et al.*, 2005; Zierdt *et al.*, 1999; Zierdt, 2007).

Tuttavia, le posizioni calcolate sono poco attendibili, come rilevato durante l'acquisizione di dati del parlato, infatti in alcune porzioni del volume di misura il rilevamento della posizione dei sensori è instabile. Soluzioni proposte dalla casa madre, quali l'impiego di sensori di nuova generazione oppure una calibrazione accurata del campo magnetico eseguita dai loro tecnici, non hanno portato miglioramenti. Tale problema costituisce un limite per l'AG500, che a differenza del modello precedente (AG200) permette al soggetto di muovere liberamente il capo nello spazio corrispondente al volume di misura V_m, come evidenziato in Fig.1, favorendo la naturalezza del parlato. Già Yunusova *et al*. (2009) avevano identificato errori sino a 2mm nella misura della posizione di due sensori posti ad una distanza fissa, senza individuare alcuna regione del campo dove *l'instabilità fosse accentuata o sistematica.*

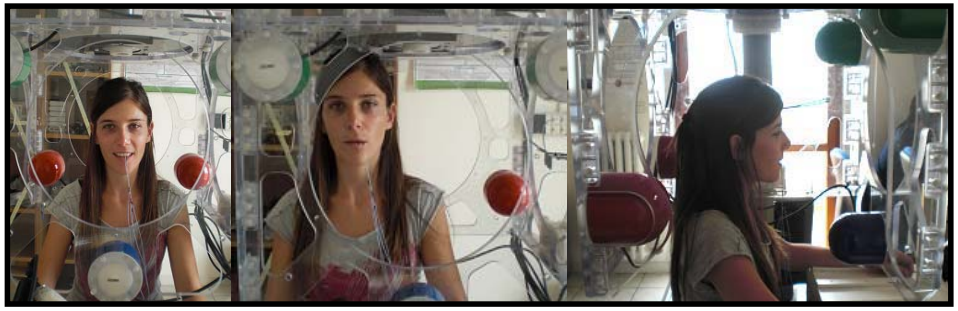


Fig. 1: Soggetto libero di muoversi nel volume di registrazione dell'AG500.

[a)] Email: massimo.stella@inbox.com

Obiettivi

Questo lavoro si propone: 1) di mettere in evidenza possibili regolarità degli errori relativi al calcolo delle posizioni e, soprattutto, 2) di individuare le cause ultime di tali errori, che riducono notevolmente l'affidabilità delle misure articolatorie. Il raggiungimento di questi obiettivi offre le basi sperimentali per un miglioramento efficace dell'AG500.

Metodi e Risultati

Per evidenziare regioni instabili sono state registrate serie di ripetizioni di sillabe [ko] e [ta] in sequenza, pronunciate da un parlante che spostava il capo in punti diversi del cubo per ogni serie. I dati, alcuni dei quali riportati in Fig.2, hanno rilevato errori sistematici in particolari zone di V_m.

Per stabilire l'origine di tali perturbazioni, si è prima cercato d'individuare possibili fonti d'interferenza esterne all'apparecchio, quindi, si è verificata la presenza di concause interne legate all'hardware. Infine, si è analizzato il software, alla ricerca di criticità nell'algoritmo numerico.

L'AG500 è stato installato in un laboratorio senza sorgenti d'interferenza, come lastre metalliche o lampade a fluorescenza, e con temperatura ambiente sotto controllo. Tutti i test sono stati eseguiti dopo aver portato a regime e calibrato la macchina. Si sono effettuate prove di movimento controllato dei sensori, verifiche sulla mutua induzione, analisi di stabilità della posizione calcolata per sensori fermi e prove numeriche con dati fittizi.

Le prove di movimento controllato dei sensori (lungo circonferenze solidali al cubo EMA, spazzate con velocità costante e a quota costante, come in Fig. 3) hanno permesso di escludere la presenza di interferenze esterne direzionali. Infatti ruotando l'intero cubo EMA non si sono notate variazioni né direzionali né d'intensità delle zone d'interferenza. L'errore medio relativo al raggio delle circonferenze risulta ~0.9mm, con punte di 10mm sui valori istantanei.

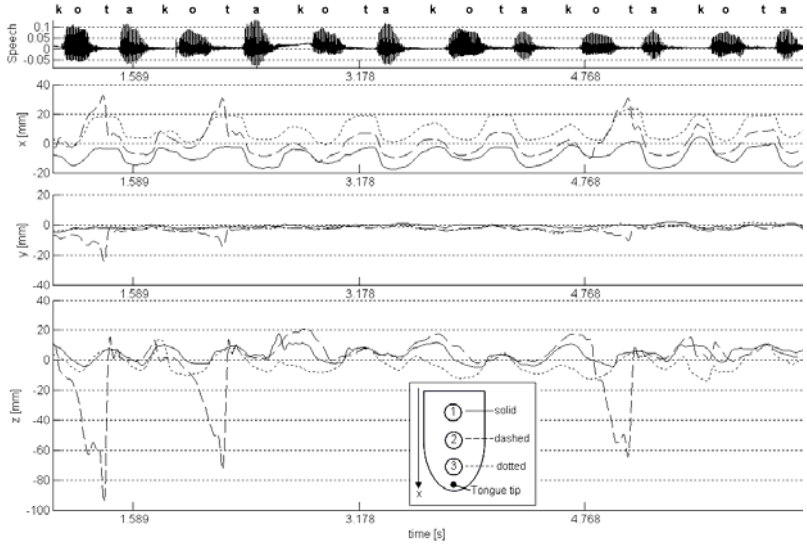


Fig. 2: Coordinate (x,y,z) nel tempo, con evidenti disturbi, per 3 sensori per le sillabe [ko] e [ta].

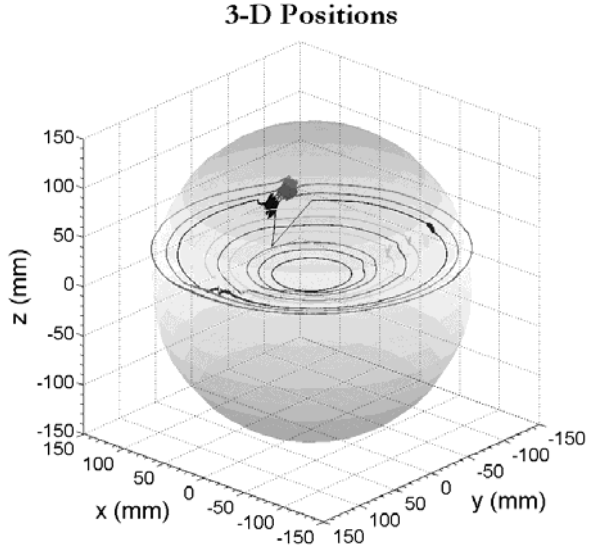
Per le prove sulla mutua induzione si è studiato il comportamento di 10 sensori posti a distanze reciproche superiori a 8mm, come indicato dai costruttori. Nessuna variazione significativa negli errori spaziali è stata riscontrata tra le configurazioni con distanza 1.5cm e 3cm, segno della scarsa influenza della mutua induzione sulle perturbazioni. Nell'articulografo 2D Kaburagi e Honda (1994) avevano trovato risultati analoghi, con errori posizionali di soli 0.3mm per la mutua induzione di sensori distanti 6 mm.

Scartate cause d'interferenza fisiche, tramite gli esperimenti di stabilità si è studiata statisticamente la dispersione delle posizioni attorno ad uno o più valori medi. Per un sensore immobile per circa 160s, entro V_m, si è registrata una deviazione standard della quota *z* di 3.9mm (mediamente, per altri sensori, in posizioni diverse, la deviazione è di ~0.2mm).

L'individuazione definitiva della natura delle perturbazioni è stata possibile grazie a controlli numerici. È stato fornito al programma CalcPos, in dotazione all'AG500, un set di correnti codificate a 16 bit e costanti per 100.000 misurazioni, come se tali dati fossero il risultato di 500s di misure. Il valore degli input è stato scelto in base ai risultati perturbati della prova di stabilità. Eliminando così ogni effetto fisico dovuto al dispositivo di misura, si è studiata soltanto la procedura numerica, dedicata al calcolo delle posizioni. La prova numerica con input costanti ha fornito in output delle quote numeriche non costanti, con una variazione massima di circa 7mm ed una distribuzione bimodale.

Tale risultato rappresenta la conferma definitiva delle ipotesi avanzate in precedenti lavori, tra cui Kroos (2008) e Kroos (2012), in cui s'ipotizzava che gli errori non fossero dovuti ai *dispositivi fisici ma piuttosto al software, basato sull'algoritmo di Newton-Raphson.*

Fig. 3: Risultato di una prova di movimento controllato per 10 sensori. Le perturbazioni sono evidenti.



La convergenza globale di tale procedura numerica alla posizione effettiva del sensore non è sempre garantita, visto la complessità delle equazioni da risolvere. L'algoritmo numerico iterativo

ha limitate capacità esplorative nello spazio delle possibili configurazioni *(x,y,z,φ,θ)* e l'individuazione della vera posizione del sensore rappresenta una sorta di discesa su una superficie con "pozzi" e "valli": i primi sono delle "trappole", che limitano la variabilità dei risultati e li trattengono lontani dal risultato ricercato; le seconde invece sono zone con scarsa variabilità, in cui l'algoritmo può "girovagare" senza mai giungere alla soluzione ottimale.

Conclusioni

In questo studio, dopo aver scartato ogni sorgente fisica di errore, si è per la prima volta dimostrato che le anomalie nella ricostruzione della posizione dei sensori sono sistematiche in alcune regioni, all'interno del volume di registrazione, e mostrano un pattern inequivocabilmente riconducibile alla mancata convergenza del metodo di calcolo.

Bibliografia

Carstens Medizinelektronik (**2009**). "AG500 Manual", available at http://www.ag500.de/manual/ag500/AG500_manual.pdf (date last viewed: 1/31/11).

Kaburagi, T., Honda, M. (**1994**). "Determination of sagittal tongue shape from the positions of points on the tongue surface", J. Acoust. Soc. Am. **96**, 1356-1366.

Kaburagi, T., Wakamiya, K., and Honda, M. (**2005**). "Three-dimensional electromagnetic articulography: A measurement principle," J. Acoust. Soc. Am. **118**, 428-443.

Kroos, C. (**2008**). "Measurement Accuracy in 3D Electromagnetic Articulography (Carstens AG500)," in *Proceedings of the 8th Seminar on Speech Production*, edited by R. Sock, S. Fuchs, and Y. Laprie, (INRIA, Strasbourg, France), pp. 61-64.

Kroos, C. (**2012**). "Evaluation of the measurement precision in three-dimensional Electromagnetic Articulography (Carstens AG500)," J. Phonetics **40**, 453-465.

Stone, M. (**1997**). "Laboratory Techniques for Investigating Speech Articulation", in W. Hardcastle & J. Laver (eds.), *The Handbook of Phonetic Sciences*, pp. 11-32.

Yunusova, Y., Green, J.R., and Mefferd, A. (**2009**). "Accuracy Assessment for AG500, Electromagnetic Articulograph," J. Speech Lang. Hear. Res. **52**, 547-555.

Zierdt, A., Hoole, P., and Tillmann, H.G. (**1999**). "Development of a system for three- dimensional fleshpoint measurement of speech movements," in *Proceedings of the XIVth International Congress on Phonetic Sciences* (San Francisco, CA), edited by J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, and A. C. Baile, vol. 1, pp. 73–76.

Zierdt, A. (**2007**). "EMA and the crux of calibration," in *Proceedings of the XVIth International Congress on Phonetic Sciences* (Saarbrücken, Germany), edited by J. Trouvain, and W. J. Barry, pp. 593-596.

THE PRODUCTION OF SYLLABLES IN STUTTERING ADULTS UNDER NORMAL AND ALTERED AUDITORY FEEDBACK

Claudio Zmarich[1] Daria Balbo[2], Vincenzo Galatà[1,3], Marine Verdurand[4], Solange Rossato[4],
[1]CNR-ISTC, Padova (I), [2]Università di Padova (I), [3]CNR-IRAT, Napoli (I), GIPSA-Lab Grenoble (F) [4]

ABSTRACT

Altered Auditory Feedback (AAF) is a powerful instrument to reduce the frequency of the stuttering episodes, although the rationales for this effect are largely unknown (Bloodstein & Bernstein-Ratner, 2008). The alterations in stutterers' speech do not only concern the motor aspect but also imply the sensory-motor loop (Hickok, Houde & Rong, 2011; Namasivayam & van Lieshout, 2011). It is also known that stutterers' fluent speech is affected by the intrinsic characteristics of the speech units, as the frequency of occurrence and the articulatory complexity of the syllables, among others (Smith, 2010; Howell et al., 2006). The present study describes the influence of the AAF on the production of different types of syllables varying for frequency and complexity in four Italian adult females who use to stutter (AWS). The subjects differ in stuttering severity according to SSI-3 (Riley, 1994), ranging from moderate to very severe.
We wanted to study:

- the influence of the articulatory complexity (Howell et al., 2006) and/or frequency of occurrence in the syllables (Stella & Job, 2000), without AAF, by quantitative-qualitative analyses of disfluencies and errors;
- the effects of AAF on disfluencies and errors in AWS'speech;
- the influence of articulatory complexity, frequency of occurrence and feedback condition on the fluent speech of AWS, in terms of:
    - acoustic duration of words and phones;
    - intrasyllabic coarticulation of CV syllables (C=voiced plosives, V=[i];[u];[a]), according to "locus equation" (Sussman et al. 2010).

Subjects repeated each target syllable nine times, immediately after a recorded voice, under normal auditory feedback (NAF) and AAF. Target syllables were always embedded within the phrase "Say CVt, then CVt, then CVt". AAF combined a delayed auditory feedback of 60 ms with a frequency shift of the original F0 (40% reduction). This combination has proved to be the most effective for inducing fluency in AWS (Antipova et al., 2008).
Results show that both higher articulatory complexity and lower frequency of occurrence of syllables increased the number of errors and stuttering episodes, but only for two subjects, the other two being fluent in all auditory feedback conditions. For the subjects who stuttered, AAF improved fluency, in terms of a reduction in errors and dysfluencies.
As for the influence of the AAF, with respect to NAF, on the intrasyllabic coarticulation degree, we found opposite effects depending on stuttering severity: while severe AWS, which benefitted most from AAF, showed a lower degree of coarticulation, moderate AWS, which benefitted less from AAF, showed a higher degree of coarticulation.
One could try to find an unitary explanation by considering that, from a theoretical point of view, two general strategies promoting fluency could exist: the reduction of the speech rate (speech variations in the temporal dimension) and the reduction of the coarticulation (speech variations in the frequency dimension). They could be independent in principle, but more often they interact in variable ways (Namasivayam & van Lieshout, 2011). Every stuttering subjects could adopt one of them, or a peculiar combination, perhaps according to the degree of severity. In this experiment, these considerations could be exemplified by the results concerning the subject L (the less severe) and the subject A (the most severe). Under AAF condition, L and A seem to improve fluency by

using two opposite strategies: L reduces speech rate without changing CV coarticulation degree, A reduces CV coarticulation degree without changing speech rate.

REFERENCES

Antipova E. A., Purdy S. C., Blakeley M., Williams S., (2008), "Effects of altered auditory feedback (AAF) on stuttering frequency during monologue speech production", Journal of Fluency Disorders, 33, 274-290.

Balbo D., (A.A. 2010/2011), "La produzione delle sillabe nella balbuzie: difficoltà articolatoria vs. frequenza d'occorrenza", tesi di laurea in Logopedia, Padova.

Bloodstein O., Bernstein Ratner N., (2008), A Handbook on Stuttering, Thomson Delmar Learning, New York (NY), 283-304.

Hickok G., Houde J. & Rong F.(2011), Sensorimotor integration in speech processing: Computational basis and neural organization, Neuron, 69, 407-422.

Howell P., Au Yeung A., Yauruss S., Eldridge K., (2006), "Phonetic difficulty and stuttering in English", Clin. Linguist. Phon., 20(9), 703-716.

Namasivayam A.K., van Lieshout P. (2001), Speech motor skill and stuttering, J. Motor Behavior, 43, 477-489

PRAAT: http://www.praat.org

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

Smith A., Sadagopan N., Walsh B., Weber-Fox C. (2010), Increasing phonological complexity reveals heightened instability in inter-articulatory coordination in adults who stutter, Journal of Fluency Disorders, 35 (1), p.1-18.

Stella V., Job R. (2000), "Frequenza sillabica e frequenza di lemmi della lingua italiana scritta", *Giornale Italiano di Psicologia,* 3, 633-639

Sussman H. M., Byrd C. T., Guitar B., (2010), "The integrity of anticipatory coarticulation in fluent and non-fluent tokens of adults who stutter", Clinical linguistics & phonetics, 25, 169-186.

Rappresentazioni uditive e (sotto)specificazione fonologica nella percezione dei contrasti consonantici: uno studio elettrofisiologico.

Roberto Petrosino[a,b], Mirko Grimaldi[a], Sandra Miglietta[a,c] e Andrea Calabrese[a,d]

[a] *Centro di Ricerca Interdisciplinare sul Linguaggio, Università del Salento*
[b] *Dipartimento di Scienze della Comunicazione, Università degli Studi di Siena*
[c] Dipartimento Antichità, Medioevo e Rinascimento, Linguistica*, Università degli Studi di Firenze*
[d] *Department of Linguistics, University of Connecticut, USA*

Introduzione

La percezione del linguaggio è possibile grazie a un processo cognitivo che permette di generare rappresentazioni uditive discrete attraverso l'elaborazione del continuum acustico e di collegarle alle rappresentazioni lessicali presenti nella memoria a lungo termine.

A partire da Trubetzkoj (1939) e Jakobson & Halle (1956), le teorie fonologiche hanno identificato tali rappresentazioni con la nozione di *fonema*, descritto da un numero limitato di *tratti distintivi binari* (cioè le caratteristiche acustico-articolatorie proprie di ogni *fono*). A causa però della limitata capacità mnemonica umana, solo i tratti impredicibili sarebbero *specificati* nelle rappresentazioni mentali, laddove i tratti predicibili sarebbero *sottospecificati*, e quindi assenti a livello sottostante (Archangeli & Pulleyblank 1989; Kiparsky 1985; Steriade 1995). Per esempio, la nasale coronale /n/ avrà una rappresentazione mentale specificata come [+ nasale; + coronale], ma il valore per il tratto [sonoro] sarà sottospecificato, poiché tutte le ostruenti nasali sono sonore e il suo valore è dunque desumibile dalla presenza di nasalità.

Pur non esente da critiche (XXXXX 1995; Halle 1995; Stanley 1967), tale assunto di recente è stato ripreso e sviluppato nel modello *Featurally Underspecified Lexicon* (FUL; Lahiri & Reez 2002). Secondo tale teoria, il segnale, una volta analizzato spettrograficamente, sarebbe valutato sulla base delle rappresentazioni lessicali mentali: alle situazioni di *match* (esatta convergenza) e *mismatch* (esatta divergenza) tra le due rappresentazioni, si affiancherebbe una terza di *nomismatch* tra una rappresentazione mentale sottospecificata e una rappresentazione del segnale in entrata completamente specificata.

Lahiri e collaboratori hanno cercato di corroborare il modello FUL facendo uso degli ERPs (Event-Related Potentials) come la *Mismatch Negativity* (MMN), una componente elettroencefalografica preattentiva di polarità negativa generata 100-250 ms dopo la presentazione di uno stimolo deviante inserito saltuariamente durante una serie ripetuta di stimoli frequenti (paradigma *oddball*); essa è un indice robusto delle rappresentazioni mnestiche lessicali (Näätänen et al. 2007), Il modello FUL presuppone che lo stimolo standard, ripetuto più volte, agisca sulla memoria a lungo termine, creando così una rappresentazione fonologica mentale a esso relativa; quando a tale rappresentazione si oppone quella dello stimolo deviante, si instaura una situazione di conflitto di tratti, individuabile nella MMN. Per esempio, in Eulitz & Lahiri (2004) il conflitto in tedesco tra /ø/ (deviante, sottospecificato per [coronale]) e /o/ (standard, specificato per i tratti [dorsale] e [labiale]) elicita una MMN più precoce e più ampia rispetto a quella elicitata da /o/ deviante.

Tuttavia, studi di questo tipo (Lipski et al. 2007; Scharinger et al. 2010) si sono concentrati principalmente su contrasti vocalici, tralasciando il livello consonantico. Infatti, Scharinger et al. (2011), facendo uso di stimoli sillabici VCV

(/awa/ ~ /aja/ e /ava/ ~ /aʒa/ dell'inglese americano), presenta risultati contrari al modello FUL.

Obiettivi dello studio

Sulla scia di Scharinger et al. (2011), con questo lavoro ci proponiamo di verificare ulteriormente gli assunti del modello FUL studiando la MMN elicitata dai contrasti consonantici /aˈta/ e /aˈpa/ dell'italiano. Secondo FUL, /t/ sarebbe sottospecificato per il tratto [coronale], e /p/ invece specificato per [labiale]; quindi, /aˈta/ deviante dovrebbe elicitare una MMN più precoce e più ampia rispetto a quella elicitata da /aˈpa/ deviante.

Metodo sperimentale Dieci esemplari per ogni stimolo, prodotti da un parlante di sesso maschile, sono stati registrati in una camera anecoica tramite software ProTools LE e un microfono Sennheiser E835 (campionamento a 44.1 kHz, risoluzione di ampiezza a 16 bits). Un test comportamentale AX ha testato l'effettiva percezione categoriale della coppia di stimoli.

Per l'esperimento EEG, seguendo il paradigma *oddball*, sono stati preparati due blocchi sperimentali, in ciascuno dei quali i due stimoli alternativamente occorrevano o in posizione standard (p = 0.875) o in posizione deviante (p = 0.125); l'intervallo dell'interstimolo oscillava tra 1200 e 1400 ms.

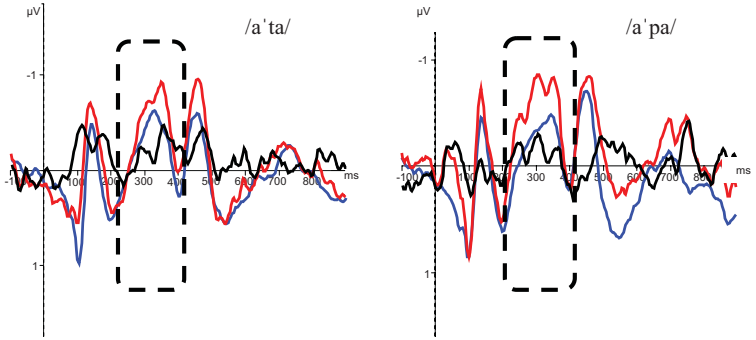
All'esperimento hanno preso parte diciotto soggetti italiani (9 femmine; età media: 23.2). Montaggio della cuffia (actiCAP a 64 canali), acquisizione, filtraggio e analisi del segnale EEG sono stati eseguiti secondo le linee guida di Picton et al. (2000).

Per ridurre la risposta neurofisiologica ai singoli esemplari di ciascuno stimolo, si è calcolata l'*identity MMN* (Pulvermüller & Shtyrov 2006): mentre la MMN tradizionale viene calcolata dalla differenza tra standard e deviante di ciascun blocco, tale metodologia consiste invece nel sottrarre "a incrocio" la risposta dello stimolo deviante alla risposta dello stesso stimolo presentato come standard.

Le MMN dovrebbero emergere a 100-250 ms dopo l'onset della devianza consonantica; poiché la prima vocale /a/ ha una durata media di 190 ms in tutti gli stimoli, la MMN si troverà presumibilmente a 280-450 ms dopo l'onset dello stimolo.

Risultati

I valori in ampiezza e latenza delle MMN ottenute sono stati analizzati con un'ANOVA a due vie per analizzare gli effetti dei fattori contrasto (/aˈta/ vs. /aˈpa/) e elettrodo (Fz, Cz, FCz; e interazione contrasto*elettrodo) su ampiezza e latenza dei picchi delle MMN individuate. Dai nostri calcoli risulta che la componente non è significativamente modulata in ampiezza rispetto ai tre fattori considerati: contrasto: $F(1, 93) = .691$; $p = .408$; elettrodo: $F(2, 93) = 1.879$, $p = .159$; contrasto*elettrodo: $F(2, 93)$; $p = .754$. Lo stesso avviene in latenza: contrasto: $F(2, 93) = .495$; $p = 0.483$; elettrodo: $F(2, 93) = .162$, $p = .851$; contrasto*elettrodo: $F(2, 93) = .274$, $p = .761$).



Onde medie dei soggetti dell'attività neurale in risposta agli stimoli /aˈta/ e /aˈpa/. La curva blu indica la risposta allo standard, la rossa al deviante, e la nera è la risultante onda di differenza. Il quadrato tratteggiato evidenzia la finestra temporale della MMN nelle due condizioni.

Conclusioni

I risultati del nostro studio, in linea con quelli di Scharinger et al. (2011), dimostrano che le predizioni del modello FUL sono disattese per contrasti consonantici.

Nello stesso tempo, i nostri dati sono in linea con la classica letteratura sulla MMN (Näätänen et al. 2007): la presenza di MMN nelle condizioni per entrambi i contrasti indica, infatti, un processo di categorizzazione percettiva, implicando l'estrazione di tratti acustici dagli stimoli presentati e la loro successiva rappresentazione mentale.

Noi interpretiamo la MMN elicitata da contrasti fonologici come la risposta di computazioni cognitive generate dall'individuazione di tratti distintivi identificati mediante *landmarks* e *cues* acustici a cui il nostro sistema percettivo-uditivo è reattivo (Stevens 2002). Infatti, riprendendo il modello *Analysis by Synthesis* (Poeppel et al. 2008), assumiamo che l'elaborazione del segnale acustico avvenga tramite un processo basato sia sull'analisi spettroacustica del segnale in entrata sia sulla verifica *on-line* di predizioni contestualmente coerenti. Tale interpretazione del processo presuppone, a nostro avviso, l'esistenza di rappresentazioni completamente specificate, piuttosto che sottospecificate.

Bibliografia

Anonimo (1995). XXXXXXXXXXXXX.

Archangeli, D., & Pulleyblank, D. (1989). Yoruba Vowel Harmony. Linguistic Inquiry 20: 173–217.

Eulitz, C., & Lahiri, A. (2004). Neurobiological Evidence for Abstract Phonological Representations in the Mental Lexicon during Speech Recognition. Journal of Cognitive Neuroscience, 16(4), 577–583.

Halle, M. (1995). Feature Geometry and Feature Spreading. Linguistic Inquiry, 26(1), 1–47.

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

Kiparsky, P. (1985). Consequences of Lexical Phonology. Phonology Yearbook, 2, 85–138.

Lahiri, A., & Reetz, H. (2002). Underspecified Recognition. In C. Gussenhoven & N. Warner (Eds.), Laboratory Phonology 7 (pp. 637–676). Berlin/New York: Mouton de Gruyter.

Lipski, S. C., Lahiri, A., & Eulitz, C. (2007). Differential Hight Specification in Front Vowels for German Speaker and Turkish-German Bilinguals: An Electroencephalografic Study. In 16th International Conference of Phonetic Science (Vol. XVI). Presented at the 16th International Conference of Phonetic Science, Saarbrücken.

Näätänen, R., Kujala, T., & Winkler, I. (2011). Auditory Processing that Leads to Conscious Perception: a Unique Window to Central Auditory Processing Opened by the Mismatch Negativity And Related Responses. Psychophysiology, 48(1), 4–22.

Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in Basic Research of Central Auditory Processing: A review. Clinical Neurophysiology, 118(12), 2544–2590.

Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. A., Johnson, R. J., Miller, G. A., et al. (2000). Guidelines for Using Human Event-Related Potentials to Study Cognition Recording Standards and Publication Criteria. Psychophysiology, 37, 127–152.

Poeppel, D., Idsardi, W. J., & van Wassenhove, V. (2008). Speech Perception at the Interface of Neurobiology and Linguistics. Philosophical Transactions of the Royal Society B, 363, 1071–1086.

Pulvermüller, F., & Shtyrov, Y. (2006). Language outside the focus of attention: The mismatch negativity as a tool for studying higher cognitive processes. Prog. Nueurobiol., 79, 49–71.

Scharinger, M., Lahiri, A., & Eulitz, C. (2010). Mismatch Negativity Effects of Alternating Vowels in Morphologically Complex Word Forms. Journal of Neurolinguistics, 23(4), 383–399. Elsevier Ltd.

Scharinger, M., Merickel, J., Riley, J., & Idsardi, W. J. (2011). Neuromagnetic Evidence for a Featural Distinction of English Consonants: Sensor- and Source-Space Data. Brain and Language, 116(2), 71–82. Elsevier Inc.

Stanley, R. (1967). Redundancy Rules in Phonology. Language, 43(2), 393–436.

Steriade, D. (1995). Underspecification and Markedness. In J. A. Goldsmith (Ed.), The Handbook of Phonological Theory (Vol. Oxford, pp. 114–174). Blackwell.

Stevens, K. N. (2002). Toward a Model for Lexical Access Based on Acoustic Landmarks and Distinctive Features. The Journal of the Acoustical Society of America, 111(4), 1872–1891.

Trubetzkoy, N. S. (1939), Grundzüge der Phonologie. Göttingen: Vandenhoeck und Ruprecht.

Percezione e produzione di vocali non native da parte di parlanti adulti

Bianca Sisinni[a], Mirko Grimaldi[a], Barbara Gili Fivela[a], Francesco Sigona[a] e Andrea Calabrese,[a,b]

[a]Centro di Ricerca Interdisciplinare sul Linguaggio (CRIL), Dipartimento di Studi Umanistici, Università del Salento

[b] Department of Linguistics, University of Connecticut, USA

Introduzione

Lo studio dei processi di percezione e produzione del linguaggio è stato per molto tempo tenuto separato (Casserly & Pisoni 2008). Uno dei settori in cui la correlazione fra i due processi ha ricevuto maggiore attenzione è l'acquisizione fonologica della seconda lingua (L2) (Flege 2003; Hansen Edwards & Zampini 2008). Secondo Listerrí (1995), la questione centrale è se gli apprendenti possano adeguatamente pronunziare suoni che non sono in grado di percepire bene o se una percezione accurata dei suoni della L2 sia un prerequisito per una buona pronuncia. La risposta a queste questioni non ha solo implicazioni teoriche ma anche ricadute applicative sui metodi per l'insegnamento della seconda lingua.

Allo stato attuale, i risultati in questo settore di studi sono contrastanti. Alcuni dati supportano l'idea che la percezione preceda o sia un prerequisito per la produzione (Trubetzkoy 1939; Neufeld 1988), prospettiva rafforzata dall'ipotesi della *phonological deafness*, secondo cui un suono deve essere adeguatamente percepito per essere coerentemente prodotto (Flege1987; Flege & Eefting1987; Flege1991; Escudero 2006). Tuttavia, studi che hanno indagato più a fondo questi due livelli riportano che la produzione può trascendere la percezione (Strange 1995; Sheldon & Strange 1982; Gass 1984; Bohn & Flege 1996; Kluge et al. 2007). Altri ancora evidenziano una moderata correlazione fra i processi di percezione e quelli di produzione (Flege 1999; Cebrián 2002; Rauber et al. 2005; Koerich, 2006). Infine, studi neurolinguistici hanno dimostrato una parziale dissociazione fra la percezione e la produzione di contrasti fonetici non nativi, a conferma di precedenti lavori sulle patologie del linguaggio (Golestani et al. 2007, Golestani & Pallier 2007).

Obiettivi

Questo lavoro indaga percezione e produzione al di fuori del processo d'apprendimento e si propone di: 1) studiare la correlazione fra percezione e produzione nell'elaborazione di vocali non native; 2) studiare le dinamiche fra i due livelli in funzione del fatto che alcuni fonemi della L2 possano essere condivisi con il sistema della L1, anche se con caratteristiche acustiche leggermente diverse, e altri no. L'ipotesi è che il differente status, sia fonologico che acustico, delle vocali L2 rispetto al sistema L1 si rifletta sul processo di categorizzazione, che, sua volta, si rifletterà su quello di produzione. In particolare, è stata esaminata la percezione e la produzione delle vocali /i/, /e/ ed /ɨ/ del Polacco da parte di parlanti dell'italiano XXXXXX (IS) con un sistema fonologico a cinque vocali: /i, e, a, o, u/. Se l'IS e il Polacco condividono i fonemi /i/ ed /e/, sia pure con differenze acustiche per F1/F2 (minori per /i/ e maggiori per /e/), /ɨ/ è presente solo nel Polacco (cfr. Fig. 1).

Metodi

Tredici parlanti dell'IS (età: 28 ± 2,67) hanno preso parte allo studio. 10 parlanti (6 uomini; età: 29,3 ± 1.54) hanno eseguito un test di categorizzazione e 3 parlanti (età: 24 ± 1) hanno eseguito un test di produzione.

3 parlanti native dell'IS (età: 30 ± 1) e 3 parlanti polacche (età: 24 ± 1) hanno prodotto gli stimoli utilizzati per i test percettivi. Le parlanti dell'IS hanno prodotto le cinque vocali native (6 ripetizioni per vocale) in pseudo-parole bisillabiche ˈbVba inserite nella frase cornice *Dico bVba poi*. Le vocali del polacco /i/, /e/ ed /ɨ/ sono state ottenute tramite la produzione (6 ripetizioni per vocale) di pseudo-parole bisillabiche ˈbVba inserite nella frase cornice "Karol szuka bVba teraz" (Karol cerca ˈbVba adesso): cfr. Fig. 1.

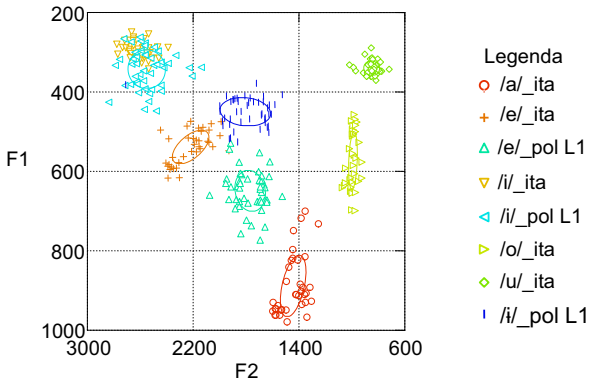


Fig. 1: Aree di esistenza delle vocali dell'IS e delle vocali del polacco.

La prima sillaba /bV/ di tutte le pseudo-parole così ottenute è stata normalizzata (intensità, durata, F0) e le sillabe più rappresentative sono state scelte da due gruppi di parlanti nativi per essere utilizzate nel test percettivo; le pseudo-parole prodotte dalle parlanti del polacco sono state utilizzate anche nel test di produzione L2.
Un test percettivo ha permesso di verificare come i fonemi della L2 sono stati categorizzati rispetto alle cinque vocali dell'IS. Ascoltato lo stimolo contenente la vocale L2 (5 volte) e scelta la vocale L1 a cui associarla, il soggetto doveva giudicare quanto le due vocali fossero simili su una scala da 1 (totalmente differente) a 5 (totalmente uguale). In base alle valutazioni dei soggetti è stato calcolato il *goodness of fit* (Flege & MacKay, 2004) e il *fit index* (Guion et al., 2000).
Per il test di produzione, le 3 parlanti dell'IS hanno realizzato le vocali L1 in una stanza insonorizzata tramite lettura della pseudo-parola ˈbVba, e le vocali L2 attraverso la tecnica della *delayed repetition* (Flege et al., 2003) che evita l'influenza diretta dell'ascolto dello stimolo acustico sulla sua produzione (12 ripetizioni per vocale in L1 e L2).
L'analisi acustica di F1/F2 (nel tratto stabile di 25ms) delle vocali L1 e L2 è stata seguita da un t-test per campioni appaiati ($p = 0.05$), al fine di verificare se i parlanti avessero prodotto i fonemi della L2 in maniera differente rispetto alle categorie native.

Risultati e conclusioni

I processi di categorizzazione (cfr. Tab. 1) evidenziano che /i/ di L2, condiviso con la L1, e simile per proprietà acustiche, è stato categorizzato nella quasi totalità dei casi come /i/ dell'IS con un giudizio di similarità medio-alto. Il fonema /e/ di L2, presente nell'inventario dell'IS, ma acusticamente dissimile, è stato categorizzato come L1 /e/ con percentuali elevate e con un giudizio di similarità intermedio. Infine, il fonema /ɨ/ di L2, non presente nel sistema nativo, è stato categorizzato con i fonemi L1 /e/, /u/ con percentuali intermedie e giudizi di similarità bassi, quindi percepito come dissimile dalle categorie native.
I processi di produzione delle tre parlanti analizzate sembrano riflettere in generale i processi di categorizzazione (cfr. Fig. 2): /i, e/ di L2 sono stati tendenzialmente prodotti con valori F1 o F2 simili a quelli di /i, e/ di L1 (L1 /i/ vs L2 /ɨ/: P1 = per F2, P3 = per F1; L1 /e/ vs L2 /e/: P1 e P3 = per F1; P2 = per F1 e F2). Quindi, benché fra /i, e/ L1 e /i, e/ L2 non ci sia una sovrapposizione netta, non sono state create categorie fonetiche completamente nuove, ma solo leggermente diverse. Al contrario, /ɨ/ di L2 è stato prodotto con valori F1/F2 diversi dalle categorie L1 a cui è stato associato percettivamente (eccetto per P3 che mostra una F1 di /ɨ/ uguale a quella di /e/ di L1).

Tuttavia, all'interno di questi processi generali si possono individuare modulazioni personali nei processi di produzione. Infatti, P1, in modo più marcato di P2 e P3, in corrispondenza dei fonemi della L2 tende a creare categorie fonetiche diverse fra loro e diverse anche dalle categorie di L1. Ciò può essere dovuto a sensibilità personali alla percezione e produzione di suoni non nativi (cfr. Mechelli, 2004).
In sintesi, si può concludere che le modalità di categorizzazione delle vocali L2 rispetto a quelle L1 si riflettono in produzione, anche se in modo diverso a seconda dei soggetti. In particolare, per i fonemi L2 /i, e/ condivisi con L1, sia pure con caratteristiche fonetiche leggermente diverse, sono state create categorie simili, mentre per il fonema non condiviso /ɨ/ due soggetti su tre hanno creato una categoria completamente nuova.

L1 / L2	/a/			/e/			/i/			/o/			/u/		
	%	GoF	fit	%	GoF	fit	%	GoF	fit	%	GoF	Fit	%	GoF	fit
/e/	2%	3	0	84%	3,1	2,6	1%	2	0	1%	3	0	13%	2,2	0,3
/i/	2%	3	0	3%	3,2	0,1	97%	3,7	3,6						
/ɨ/				47%	2,2	1	4%	2,2	0				45%	2,5	1,1

Tab. 1: Percentuale di categorizzazione (%), goodness of fit (GoF) e fit index (fit) dei fonemi del polacco L2 categorizzati in termini dei fonemi L1.

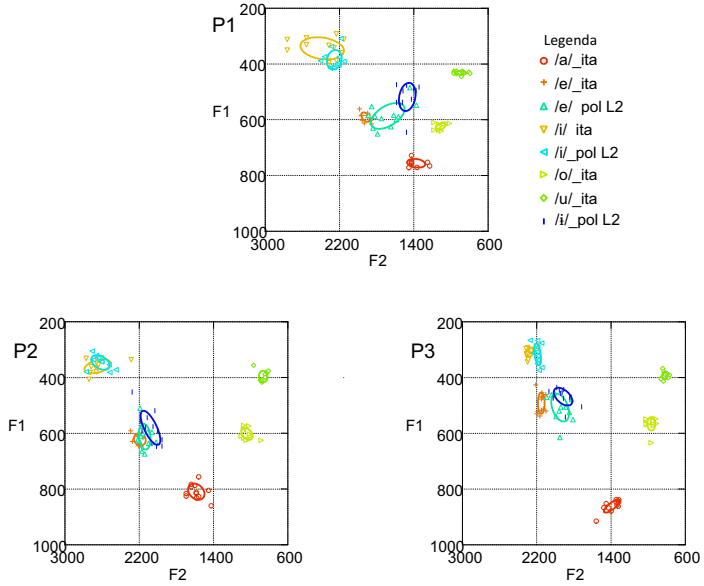


Fig. 2: aree di esistenza delle vocali dell'IS (*_ita) e delle vocali del polacco L2 (*_pol L2) prodotte dalle 3 parlanti dell'IS (P1, P2, P3).

Bibliografia

Bohn, O.-S., &Flege, J. E. (1996). Perception and production of a new vowel category by adult second language learners. In A. James &J. H. Leather (Eds.), Second-language speech. Structure and Process (pp.53-73). New York: Mouton de Gruyter.
Casserly, E. D. & Pisoni, D. B. (2008), Speech Perception and Production, Research on Spoken Language Processing – Progress Report 29, Indiana University: 232-254.

Catford, J. C., & Pisoni, D. (1970). Auditory versus articulatory training in exotic sounds. The Modern Language Journal, 54, 477-481.
Cebrián, J. (2002). Phonetic Similarity, syllabification and phonotactic constraints in the acquisition of a second language contrasts. PhD Dissertation. Toronto Working papers in Linguistics Dissertation Series. University of Toronto, Toronto, Canada.
Escudero, P. (2006). Second Language Phonology: The Role of Perception. In: Pennington, M. C. (ed.). *Phonology in Context*. New York: Palgrave Macmillan, 109-134.
Flege, J. (1999). The relation between L2 production and perception. In J. Ohala, Y. Hasegawa, M. Granveille & A. Bailey (Eds.) Proceedings of the XIVth International Congress of Phonetics Sciences (pp. 1273-1276). Berkely, United States.
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.
Flege, J.E. (2003). Assessing constraints on second-language segmental production and perception. In: Meyer, A. and Schiller, N. (eds). *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*. Berlin: Mouton de Gruyter, 319-355.
Flege, J. E., & Eefting, W. (1987). Production and perception of English stops by native Spanish speakers. Journal of Phonetics, 15, 67-83.
Flege, J.E., MacKay, I.R.A. (2004), Perceiving vowel in a second language, *Studies in Second Language Acquisition*, 26, 1-34.
Gass, S. (1984). Development of speech perception and speech production in adult second language learners. Applied Psycholinguistics, 5, 51-74.
Golestani, N. &Pallier, C. (2007). Anatomical Correlates of Foreign Speech Sound Production, Cerebral Cortex, 17(4), 929-934.
Golestani, N., Molko, N., Dehaene, S., Le Bihan, D. & Pallier, C. (2007). Brain Structure Predicts the Learning of Foreign Speech Sounds, Cerebral Cortex, 17(3), 575-582.
Guion, S.G., Flege, J.E., Ahahane-Yamada, R., Pruitt J.C. (2000), An investigation of current models of second language speech perception: the case of Japanese adults' perception of English consonants, *Journal of the Acoustical Society of America*, 107, 2711-2725.
Hansen Edwards, J.G.& Zampini, M.L. (2008). *Phonology and Second Language Acquisition*. Amsterdam/Philadelphia: John Benjamins.
Kluge, D. C., Rauber, A. S., Reis, M. S. & Bion, R. A. H (2007). The relationship between the perception and production of English nasal codas by Brazilian learners of English. Proceedings of Interspeech 2007 (pp. 2297-2300). Antwerp, Belgium.
Koerich, R. D. (2006). Perception and Production of vowel paragorge by Brazilian EFL students. In B. Baptista& M. Watkins (Eds.), English with a Latin Beat. Studies in Portuguese/Spanish – English Interphonology. Studies in Bilingualism 31. Amsterdam: John Benjamins.
Llisterrí, J. (1995). Relationships between Speech Production and Speech Perception in a Second Language, *Proceedings of the 13th International Congress of Phonetic Sciences*, Vol. 4, 92-99.
Mechelli, A. 2004. Structural plasticity in the bilingual brain: proficiency in a second language and age of acquisition affect grey-matter density. *Nature* 431: 757.
Neufeld, G. (1988). Phonological asymmetry in second-language learning and performance. Language Learning, 38 (4), 531-559.
Rauber, A., Escudero, P., Bion, R. A. H., & Baptista, B., O (2005). The interrelation of Perception and Production of English Vowels by Native Speakers of Brazilian Portuguese. Proceedings of Interspeech 2005 (pp. 2913-2916). Lisbon, Portugal.
Sheldon, A. & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. Applied Psycholinguistics, 3 (3), 243-261.
Strange, W. (1995). Phonetics of second language acquisition: past, present and future. Proceedings of the XIIIth International Congress of Phonetic Sciences, 4, (pp. 84-91). Stockholm, Sweden.
Trubetzkoy, N. S. (1939). GrundzügederPhonologie. Travaux duCercle Linguistique de Prague VII, English Translation by Chr. Baltaxe, University of California Press.

Coordinazione interarticolatoria nella produzione dell'intonazione L2

Antonio Stella, Barbara Gili Fivela, Francesco Sigona
CRIL & Università del Salento

Produrre correttamente l'intonazione degli enunciati in lingua straniera può essere un compito molto difficile per un discente. Il fattore che sembra esercitare un'influenza maggiore è l'interazione con il sistema prosodico nativo. Il contrasto tra il sistema prosodico nativo e quello della lingua target può implicare influenze di tipo fonologico e/o fonetico, con effetti diversi sulle produzioni in L2. Nel caso degli accenti tonali, le influenze fonologiche si presentano quando viene utilizzata una categoria al posto di un'altra (ad esempio, un accento tonale ascendente in luogo di uno discendente); le seconde agiscono sulla realizzazione fonetica di una categoria che appartiene ad entrambi i sistemi: ad esempio, l'uso di uno stesso accento ascendente implementato con differente allineamento nelle due lingue [1]. Alcuni studi [2, 3, 4] mostrano che questi due tipi di influenze agiscono in maniera diversa sulle produzioni in L2, a seconda del livello di competenza del parlante: in un primo stadio di apprendimento vengono individuate e superate le influenze di tipo fonologico; solo in un secondo momento si superano le influenze fonetiche dovute, ad esempio, all'uso di categorie simili per la stessa funzione pragmatica. Inoltre, Mennen [5] mostra che le caratteristiche fonetiche sono molto difficili da modificare anche per parlanti con competenza molto alta della lingua straniera, in particolare quando le categorie fonologiche sono condivise dalla L1 e dalla L2 e differiscono solo per caratteristiche fonetiche.

Questo contributo ha lo scopo di evidenziare le influenze della L1 sull'allineamento di accenti tonali utilizzati in contesti di focus largo (FL), analizzando le produzioni in tedesco L2 da parte di parlanti nativi dell'italiano parlato a XX (XX per garantire l'anonimato) con differente livello di competenza in L2 (alto e basso). Gli accenti tonali realizzati in posizione iniziale di enunciato con FL sono ascendenti nei sistemi nativi delle due lingue e sono etichettati come L+H* nell'italiano di XX, e come L+H* o L*+H nel tedesco [6] (alcune analisi del tedesco [7] riportano un uso indistinto delle due categorie sia in caso di FL che in caso di focus contrastivo (FC)).

Dal punto di vista fonetico, in entrambe le lingue gli accenti prodotti in enunciati con FL sono realizzati con un'ascesa tonale in corrispondenza della sillaba tonica: il target basso iniziale dell'accento è allineato all'inizio della sillaba in italiano e all'interno della vocale in tedesco, mentre il tono alto è posizionato alla fine della sillaba tonica in italiano e alla fine della post-tonica (oppure in fine parola) in tedesco (Fig. 1). Uditivamente, l'accento è realizzato come ascendente in entrambe le lingue con una più forte percezione del tono basso nel tedesco.

Per raggiungere l'obiettivo di questo lavoro le differenze nell'allineamento degli accenti tonali sono state valutate dal punto di vista articolatorio. L'analisi articolatoria dell'allineamento tonale mette in relazione il contorno intonativo con i gesti di apertura e chiusura degli articolatori coinvolti nella produzione della sillaba tonica. Il quadro teorico di riferimento è quello della *Articulatory Phonology* [8]: gli accenti sono assimilati a gesti tonali e viene valutata la sincronia tra questi ultimi e i gesti sopralaringali utilizzando il *Coupled Oscillator Model* [9]. Tale analisi si è rivelata molto utile nello studio di differenze cross-linguistiche nell'allineamento [10, 11, 12, 13] e sembra evidenziare risultati molto più stabili rispetto alle stesse indagini condotte su base acustica.

L'analisi articolatoria dell'allineamento nel caso di FC, condotta in uno studio precedente (nel quale sono stati analizzati dati acustici e articolatori simili a quelli qui considerati e prodotti dagli stessi parlanti nella medesima sessione sperimentale), mostra che un corretto allineamento dei toni è implementato solo dal parlante con alta competenza del tedesco, mentre l'influenza della lingua nativa è ancora ben visibile nelle produzioni in tedesco del parlante con bassa competenza. Nel caso di FC però gli accenti tonali prodotti in posizione iniziale mostravano differenze fonologiche e fonetiche ancora più marcate tra le due lingue rispetto a quelle realizzate in caso di FL: mentre nell'italiano di XX l'accento ha un contorno ascendente-discendente H*+L, in tedesco l'accento ha un contorno ascendente L+H* o L*+H (Fig. 2).

Nello studio qui presentato, oltre all'analisi dei dati relativi al FL, si compareranno i risultati ottenuti per i due tipi di focus. L'ipotesi alla base del lavoro è che per i parlanti sia più facile implementare correttamente

gli accenti in caso di FC rispetto agli accenti in caso di FL, per via della differenza più marcata tra gli accenti in FC nelle due lingue.

I materiali sperimentali analizzati sono stati ricavati utilizzando 2 corpora (uno per lingua) formati da 4 minidialoghi composti da due coppie domanda/risposta: nella prima è elicitata una risposta con FL, mentre nella seconda una con FC (usata per lo studio menzionato in precedenza). L'accento tonale considerato per l'analisi occorre sulla sillaba [ma.l] o [mal] in parole piane o sdrucciole. Alle registrazioni acustiche e articolatorie (effettuate con l'AG500) hanno preso parte due parlanti nativi di XX - rispettivamente con alto e basso livello di competenza del tedesco L2 - i quali hanno prodotto entrambi i corpora per 7 volte, ed un parlante nativo tedesco, che ha prodotto il solo corpus in tedesco per 7 volte. L'analisi è stata effettuata ispezionando i tracciati di velocità dei singoli articolatori e misurando la latenza tra i target tonali e i gesti di apertura e chiusura sull'asse verticale per 3 sensori: il sensore posto sul labbro inferiore (LL) per isolare i gesti labiali prodotti per [m]; il sensore sulla punta della lingua (TT) per i gesti alveolari prodotti per [l]; il sensore sul dorso della lingua (TD) per i gesti vocalici. La sincronia tra gesti tonali e gesti sopralaringali è stata valutata utilizzando un diagramma con 5 livelli separati e sincronizzati tra loro, contenenti: 1) le medie dei segmenti, 2) le medie della durata di ascesa e discesa degli accenti tonali, e le medie degli intervalli di attivazione dei gesti di 3) TT, 4) TD e 5) LL, con i relativi picchi di velocità. Esempi di tale diagramma sono riportati in Fig. 3.

Dai risultati preliminari sugli enunciati con FL (Fig. 3) emerge che nelle produzioni in tedesco L2 entrambi i parlanti allineano i target tonali dell'ascesa in maniera differente rispetto al parlante nativo tedesco: infatti mentre per quest'ultimo l'ascesa inizia dopo l'onset del gesto di chiusura (apparentemente con il picco di velocità del gesto) della punta della lingua per la consonante post-tonica, nei due parlanti italiani l'ascesa è allineata con l'inizio dello stesso gesto e mostra un allineamento simile a quello delle produzioni in italiano. Sebbene vi siano delle differenze fonetiche tra i due parlanti italiani legate a diverse strategie di produzione, in generale sembra che nella produzione di FL l'accento non sia implementato correttamente neanche dal parlante con alta competenza (diversamente da quanto osservato per FC).

Lo studio sarà completato con una dettagliata analisi articolatoria delle produzioni per verificare che tale differenza sia attestata stabilmente in tutte le condizioni sperimentali e per inserire tali risultati in un quadro teorico sull'apprendimento delle caratteristiche intonative, anche alla luce del confronto con i risultati ottenuti per il FC.

Figure

Figura 1: produzioni dell'accento tonale iniziale in enunciati con FL di un parlante nativo dell'italiano di XX (a sinistra) e di un parlante nativo tedesco (destra).

ITALIANO TEDESCO

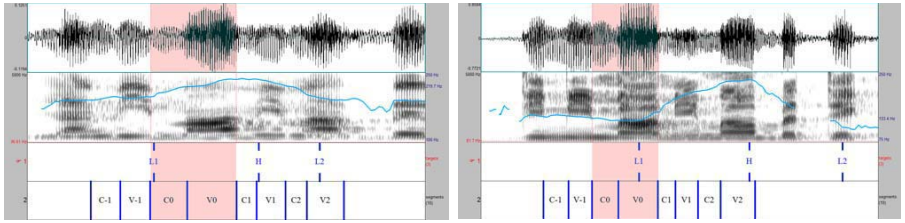


Figura 2: produzioni dell'accento tonale iniziale in enunciati con FC di un parlante nativo dell'italiano di XX (a sinistra) e di un parlante nativo tedesco (a destra).

ITALIANO TEDESCO

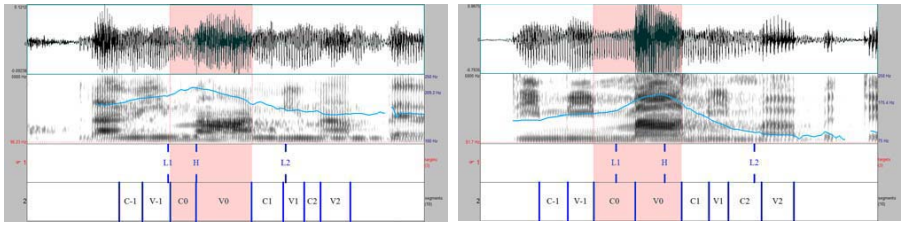
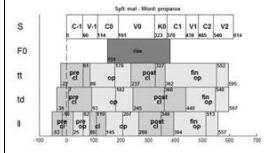
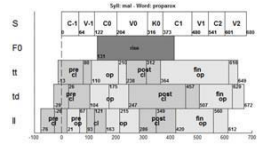
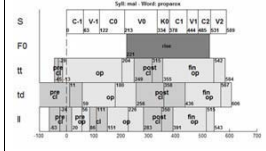
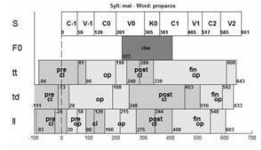
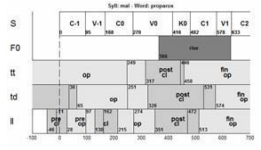


Figura 3: Diagrammi di allineamento.

	Parlante italiano (bassa competenza)	*Parlante italiano (alta competenza)*	*Parlante nativo tedesco*
ITA			
DEU			

Bibliografia

[1] Mennen, I. (2007), Phonological and phonetic influences in non-native intonation, in J. Trouvain & U. Gut (eds.), *Non-native Prosody: Phonetic Descriptions and Teaching Practice,* The Hague: Mouton De Gruyter, 53-76.

[2] Ueyama, M. & Jun, S.-A. (1996), Focus realization in Japanese English and Korean English intonation, in *UCLA Working Papers in Phonetics*, 94.

[3] Ueyama, M. (1997), The phonology and phonetics of second language intonation: the case of "Japanese English", in *Proceedings of the 5th European Speech Conference*, 2411-2414.

[4] Jun, S.-A. & Oh, M. (2000), Acquisition of second language intonation, in *Proceedings of International Conference on Spoken Language Processing*, 4, 76–79.

[5] Mennen, I. (2004), Bi-directional interference in the intonation of Dutch speakers of Greek, in *Journal of Phonetics*, 32, 543-563.

[6] Grice, M., Baumann, S., & Benzmueller, R. (2005), German intonation in autosegmental-metrical phonology. In S.-A. Jun (ed.), *Prosodic typology*, (pp.55 – 83), Oxford: Oxford University Press.

[7] Braun, B. (2006), Phonetic and phonology of thematic contrast in German, *Language and Speech, 49*, 451-493.

[8] Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: an overview. *Phonetica*, *49*(3-4), 155–180.

[9] Nam, H. & Saltzman, E. (2003), A competitive, coupled oscillator model of syllable structure, in *Proceedings of 15th ICPhS*, Barcelona, 2253-2256.

[10] D'Imperio, M., Nguyen, N. & Munhall, K. G. (2003), An articulatory hypothesis for the alignment of tonal targets in Italian, in *Proceedings of 15th ICPhS*, Barcelona, 253-256.

[11] D'Imperio, M., Espesser, R., Loevenbruck, H., Menezes, C., Nguyen, N. & Welby, P. (2007), Are tones aligned with articulatory events? Evidence from Italian and French, in J. Cole & J. I. Hualde (eds), *Papers in Laboratory Phonology IX*, Mouton de Gruyter, 577-608.

[12] Prieto, P., Mücke, D., Becker, J. & Grice, M. (2007), Coordination patterns between pitch movements and oral gestures in Catalan, in J. Tourvain & W. J. Barry (eds.), *Proceedings of the 16th ICPhS*, Dudweiler: Pirrot GmbH, 989-992.

[13] Mücke, D., Grice, M., Becker, J. & Hermes, A. (2009), Sources of variation in tonal alignment: evidence from acoustic and kinematic data, in *Journal of Phonetics, 37* (3), 321–338.

IL CONTROLLO DELLA QUALITA' NEL PROCESSO DI CONSERVAZIONE DELLE REGISTRAZIONI SONORE COME FONTE DOCUMENTARIA NELLE INDAGINI LINGUISTICHE.

Federica Bressan
Dip. di Informatica
Università di Verona
Strada Le Grazie 15, 34134 Verona
federica.bressan@dei.unipd.it

Sergio Canazza
Dip. di Ingegneria dell'Informazione
Università di Padova
Via G. Gradenigo 6/B, 35131 Padova
canazza@dei.unipd.it

Gli strumenti informatici dedicati all'analisi del linguaggio parlato hanno raggiunto livelli di sofisticazione impensabili fino a pochi decenni fa, e si stanno aprendo nuovi orizzonti per la ricerca nel campo dei dispositivi mobili e delle traduzioni automatiche del dialogo in tempo reale. I documenti orali – ovvero il linguaggio parlato fissato su supporti audio di diversa natura – hanno acquistato un valore crescente in molti settori della linguistica, ora riconosciuti pienamente nella loro funzione di fonti documentarie in grado di esprimere sia caratteristiche esplicite del linguaggio, sia caratteristiche implicite che gli attuali strumenti di analisi non sono ancora in grado di apprezzare e di codificare. La riconosciuta importanza delle registrazioni sonore come fonte documentaria ha fatto sì che aumentasse l'attenzione verso gli archivi sonori, pochissimi dei quali erano stati tutelati e valorizzati adeguatamente nel tempo. La situazione più generale degli archivi, inclusi quelli che custodiscono materiale di interesse linguistico, si è rivelata tutt'altro che confortante: migliaia di ore di indagini linguistiche e di testimonianze orali si trovano ancora in situazioni di serio degrado, per lo più fissate su supporti chimicamente instabili e negletti in scatoloni nei magazzini dei dipartimenti universitari o nelle abitazioni private dei raccoglitori degli archivi o dei loro eredi (Edmonson, 2002). Per prendere ad esempio un'area che è già stata oggetto di azioni di recupero e conservazione, la regione Toscana, la situazione degli archivi a rischio è impressionante per la quantità di materiale inedito e non ancora segnalato nei censimenti ufficiali o ufficiosi. Il progetto di ricerca "Grammo-foni. Le soffitte della voce" (2011-2013), condotto dalla Scuola Normale Superiore di Pisa e dall'Università degli Studi di Siena (PAR FAS 2007-2013 Regione Toscana Linea di Azione 1.1.a.3.), ha dimostrato che nonostante l'esistenza di un censimento dettagliato per quanto concerne i beni vocali di Toscana (Andreini et alii, 2007), il numero di archivi che richiedono un intervento di recupero e di restauro, pena una perdita irreversibile delle registrazioni, appare in continua crescita (citazione bibliografica omessa per evitare autoriferimenti). E' evidente che il *corpus* vocale costituito dall'insieme delle registrazioni sonore non diminuisce con il tempo il proprio ruolo all'interno della ricerca né di quello che riveste per la comunità non specialistica, anzi esso aumenta alla luce delle nuove frontiere scientifico-tecnologiche in grado di: (a) sfruttare tali registrazioni in maniera originale nell'ambito di nuove applicazioni, e (b) di estrarre dalle registrazioni informazioni sinora "imprigionate" nel segnale audio. La letteratura di ambito archivistico e di ambito ingegneristico degli ultimi trent'anni riporta innumerevoli studi legati al trattamento più adeguato dei documenti sonori, in particolare mettendo in guardia sulle manipolazioni, volontarie e involontarie, consapevoli o meno consapevoli, che possono alterare la natura del segnale audio fino a tradire completamente il documento di origine, con le gravissime conseguenze che ne derivano per gli studiosi che a tali documenti si affidano per formulare le proprie architetture di pensiero. Alle soglie di un nuovo capitolo per la ricerca linguistica in cui il paletto è stato spostato molto più in avanti nel campo computazionale, è sempre più importante diffondere la conoscenza sul corretto trattamento dei documenti sonori, per non rischiare di trasferire gran parte del *corpus* vocale nel cosiddetto "mondo digitale" e scoprire, o peggio non scoprire mai, che quelle voci e quei suoni sono stati falsificati, che non sono *autentici* e quindi invalidanti per ogni applicazioni di cui stanno alla base.

Nell'era digitale, occorre operare una riflessione per ridefinire i concetti di *affidabilità*, di *accuratezza* e di *autenticità* dei documenti, non solo sonori (Factor, 2009). Anche in questo campo, fortunatamente, l'informatica è in grado di offrire strumenti, esistenti e in fase di sviluppo, per il

controllo e per l'automazione in grado di *garantire* che i parametri di affidabilità, di accuratezza e di autenticità vengano rispettati.

Nell'articolo verranno esposte le principali posizioni nel dibattito storico sull'etica del restauro dei segnali audio e sulle metodologie di conservazioni più accreditate dalla comunità archivistica internazionale. Sarà presentato un protocollo pensato appositamente per la ri-mediazione dei documenti orali e saranno evidenziate le azioni di manipolazione che vengono operate comunemente sui documenti sonori e che sono la principale causa dell'alterazione del segnale, di cui fanno le spese lo studioso e l'utente del *file* sonoro messo a disposizione dall'archivio in loco oppure via web. Verranno altresì illustrati alcuni casi particolarmente esemplificativi di restauro operati in maniera filologicamente corretta.

Riferimenti bibliografici

Andreini, A. & Clemente, P. (a cura di) (2007), "I custodi delle voci. Archivi orali in Toscana: primo censimento", Firenze, Regione Toscana.

Edmonson R. (2002), "Memory of the World: General Guidelines to Safeguard Documentary Heritage", UNESCO.

Factor, M. & Henis, E. & Naor, D. & Rabinovici-Cohen, S. & Reshef, P. & Ronen, S. & Guercio, M. (2009), "Authenticity and provenance in long term digital preservation: Modeling and implementation in preservation aware storage", Proceedings of the First workshop on the theory and practice of provenance, San Francisco, California, USA.

Atlante Multimediale dei Dialetti Veneti

GRAZIANO TISATO, PAOLA BARBIERATO, GIACOMO FERRIERI,
CARLA GENTILI, MARIA TERESA VIGOLO

ISTC- Istituto di Scienze e Tecnologie della Cognizione
CNR - Centro Nazionale delle Ricerche, Padova
tisato@pd.istc.cnr.it

SOMMARIO

Il lavoro presenta la metodologia di realizzazione e le caratteristiche dell'Atlante Multimediale dei Dialetti Veneti (AMDV).

L'AMDV è un progetto interdisciplinare che ha riunito un gruppo di esperti di dialettologia, di etimologia, di fonetica e di etnografia per creare un atlante sonoro dei dialetti veneti che sfruttasse le moderne metodologie della linguistica geografica.

Il progetto triennale è stato finanziato dalla Fondazione della Cassa di Risparmio di Padova e Rovigo (Cariparo), come progetto di eccellenza 2007-2008.

L'AMDV si è ispirato ad analoghi atlanti parlanti (ALD [1], ALEPO [2], VIVALDI [3], ecc.), il cui obbiettivo è di restituire la dimensione fonetico-acustica originale che sta alla base di trascrizioni più o meno discutibili [1].

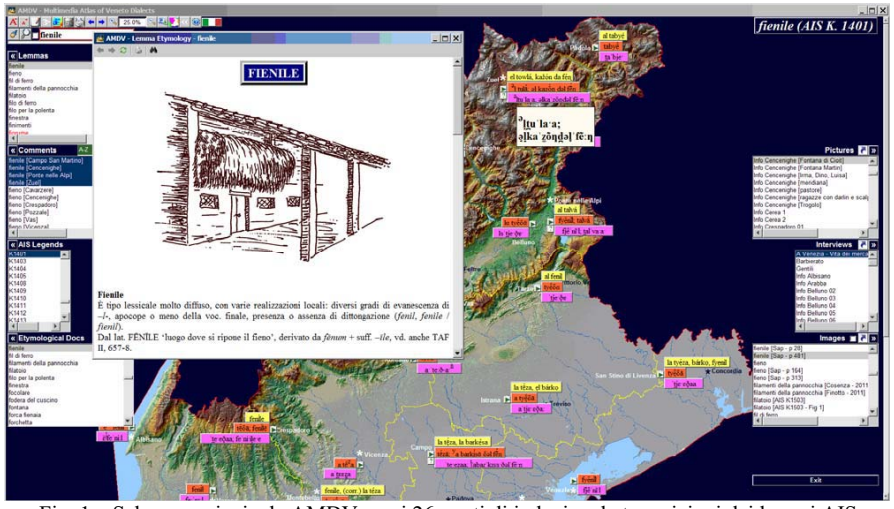


Fig. 1 – Schermo principale AMDV con i 26 punti di indagine, le trascrizioni dei lemmi AIS originali (etichetta gialla) ed attuali (etichetta arancio con trascrizione AIS-like e magenta con trascrizione IPA), la casella per la ricerca in tutti i documenti, la lista dei commenti sonori, delle legende delle tavole AIS, delle schede etimologiche, dei disegni, dei video, delle fotografie. In alto i pulsanti per l'ingrandimento dei caratteri e della mappa, per la ricerca fonetica, il dizionario, il sonogramma, ecc. Sovrapposta compare una scheda lessicale-etimologica.

La principale caratteristica dell'AMDV riguarda il confronto diacronico fra il repertorio lessicale raccolto nella regione Veneto da Paul Scheuermeier del 1921-1928, e pubblicato negli 8 volumi dell'AIS (Atlante Italo-Svizzero - *Sprach- und Sachatlas Italiens und der Südschweiz*) di Karl Jaberg e Jakob Jud, e il lessico raccolto dall'AMDV nelle inchieste del 2009-2010, esattamente

negli stessi luoghi, usando un questionario modellato, sia pure in forma ridotta, su quello predisposto per l'AIS (Fig. 1). Un'idea simile era stata realizzata su scala molto più ridotta per un progetto sui dialetti trentini del 2003 [4].

Il lavoro descriverà la metodologia innovativa che è stata sviluppata per quanto riguarda l'acquisizione dei materiali dialettali e la successiva elaborazione e trascrizione, e come si è cercato di recuperare e valorizzare una miniera inesauribile di informazioni dialettali ed etnografiche, che giace inutilizzato nei volumi dell'AIS per le difficoltà di accesso.

In effetti, il passo preliminare al progetto AMDV è stato la realizzazione di una versione digitale navigabile dell'atlante AIS che doveva servire sia per il controllo in tempo reale delle risposte degli informatori, sia per la trascrizione in un *database* del lessico AIS per la regione Veneto, sia per ricavare i disegni da utilizzare nel multimediale.

Il programma (chiamato NavigAIS) è stato scritto in Matlab ed è disponibile a questo indirizzo http://www3.pd.istc.cnr.it/navigais [5].

Fig. 2 – NavigAIS: La finestra di navigazione (in alto a sinistra), la mappa AIS 225 (La Pialla), e la finestra con la ricerca dei lemmi e dei punti (in alto a destra).

Un altro criterio metodologico adottato è stato quello della ricerca del massimo livello di automazione in tutte le fasi dell'elaborazione, in cui questo fosse ovviamente possibile: registrazione ed elaborazione audio-video, creazione dei *database*, controllo dell'informazione, ecc.

Per acquisire il materiale sonoro, è stato sviluppato un programma originale (SynRec) che consentisse di presentare al soggetto un disegno, una foto, o un testo, secondo una lista predeterminata di eventi (Fig. 3), in modo da minimizzare l'interferenza con l'informatore, e contemporaneamente far partire una registrazione in sincrono con le domande poste (Fig. 4). Una volta completata la risposta, il programma salva lo spezzone sonoro con il nome stesso dell'oggetto, facilitando tutte le fasi successive di elaborazione dell'audio e costruzione dei *database* dei lemmi e dei commenti.

	A	B	C	D	E	F
1	Step N.	Record	Random	Audio	Image	Text
2	Step 0	N	N	proverbi in dialetto di Cencenighe (storia)		Introduction
3	Step 1	N	N	silence(1)		
4	Step 2	N	N			Audio+Images
5		N	Y	scandola	scandola	
6		N	Y		albero	
7		N	Y		ceppo	
8		N	Y		tronco	
9		N	Y		roncola	
10		N	Y		accetta	
11		N	Y		scure	
12		N	Y		scure da squadratura	
13		N	Y		cuneo	
14		N	Y		mazza di legno	
15	Step 3	N	N	silence(1)		
16	Step 4	N	N			Record
17		Y	Y		sega	
18		Y	Y		sega a telaio	
19		Y	Y		sega lunga	
20		Y	Y		pialla	

Fig. 3 – SyncRec: Partitura degli eventi della sessione: le linee 5-14 comandano la sequenza casuale di immagini e/o testo di addestramento. La sessione di registrazione comincia dalla riga 16 in poi.

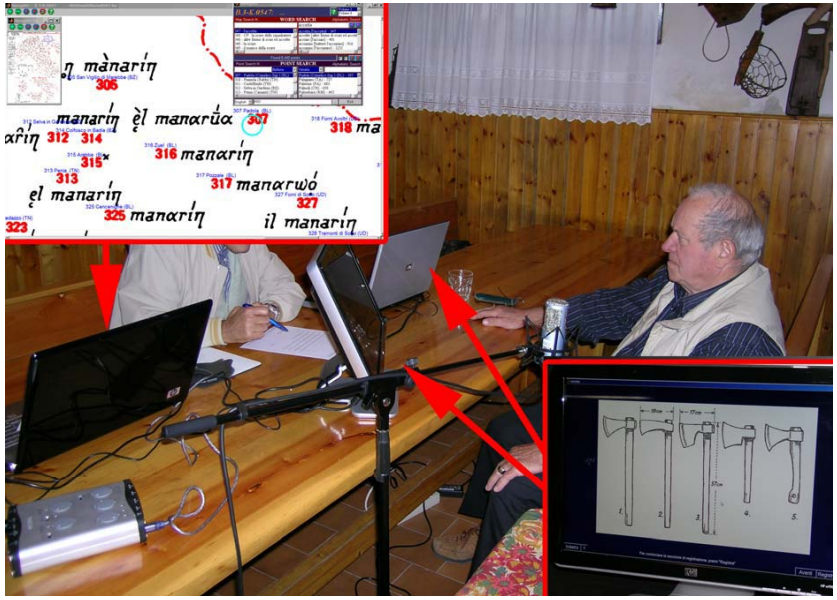


Fig. 4 – Metodologia di acquisizione nelle inchieste AMDV. In alto compare lo schermo di NavigAIS, visto dal dialettologo. In basso, il disegno dell'oggetto visto dall'informatore, dal linguista e dal responsabile della registrazione audio.

Mentre il soggetto parlava, il dialettologo poteva verificare la risposta corrispondente sulla pagina di NavigAIS e rilevare immediatamente le concordanze o le discordanze, ed eventualmente intervenire e chiedere spiegazioni. Questa possibilità ha dato ai ricercatori dell'AMDV un vantaggio notevole in termini di precisione e ha favorito quantitativamente il numero di lemmi

raccolti (14.500), rispetto a quelli acquisiti da Scheuermeier (11.600), che a quel tempo non poteva disporre di simili fonti di informazione e di paragone.

Fig. 5 – NavigAIS mostra le parole inserite nel *database* vicino all'originale sulla mappa AIS (in verdazzurro), per il controllo della trascrizione dell'AIS.

L'automazione delle procedure e la realizzazione di NavigAIS ha facilitato anche il controllo dei dati inseriti nei *database*, permettendo ad esempio l'immediata verifica della correttezza della trascrizione dei lemmi AIS visualizzandola automaticamente vicino all'originale nelle mappe AIS (Fig. 5).

La trascrizione dei materiali sonori è stata fatta oltre che con lo standard internazionale IPA anche con lo stesso simbolismo usato da Scheuermeier nell'AIS, per consentire il confronto fra le inchieste del 1921 con quelle odierne. Per realizzare la trascrizione, si è evitato l'uso di un *font* proprietario per la conseguente necessità di installare il *font* e di una programmazione *software* molto più complessa per l'ordinamento alfabetico e per la ricerca dei lemmi o di sequenze fonetiche. In Fig. 6 si vede l'inventario fonetico AMDV e un esempio di ricerca di una sequenza.

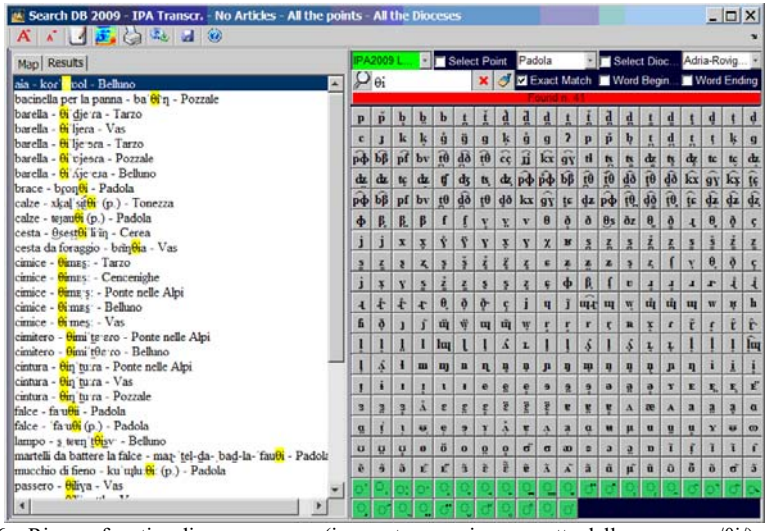


Fig. 6 – Ricerca fonetica di una sequenza (in questo caso ricerca esatta della sequenza /θi/), di cui compaiono i risultati nella colonna di sinistra.

Il lavoro illustrerà anche gli strumenti di analisi acustica e fonetica, ed un originale sistema di mappatura dei vocoidi nello spazio della vocali italiane ricavato dai dati elaborati da F. Ferrero all'ISTC (Fig. 6).

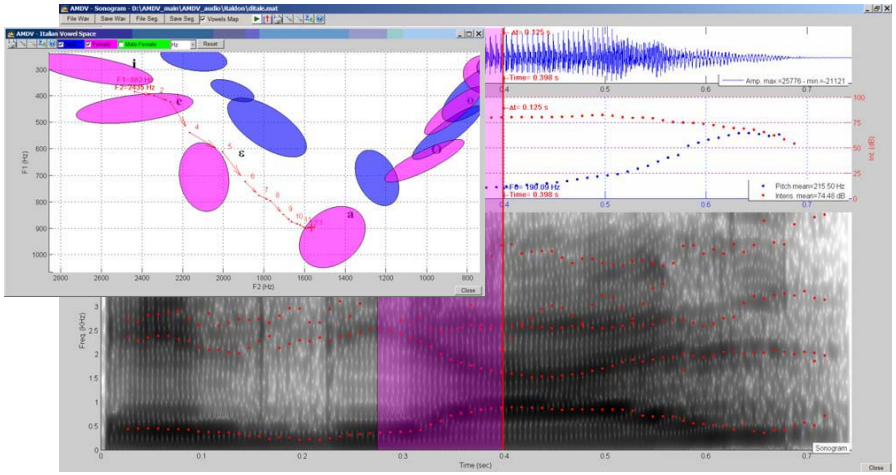
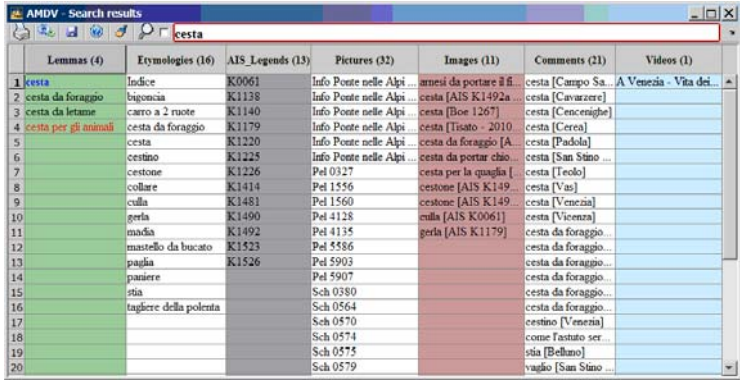


Fig. 6 – Sonogramma, e andamenti del pitch e dell'intensità per la parola *il ditale* [eḷ ˈd̪jaˑaḷ] pronunciata da un soggetto femminile di Raldon (VR). Nella finestra a sinistra si vede una transizione /ja/ di 125 ms in passi di 10ms nello spazio vocalico italiano: la lunghezza delle frecce misura la velocità articolatoria istantanea.

Si spiegherà anche il lavoro che è stato fatto per indicizzare ed annotare le diverse fonti audio-visuali in modo da poter recuperare in maniera omogenea l'informazione in tutti i media AMDV (Fig. 7).



AMDV - Search results — cesta

	Lemmas (4)	Etymologies (16)	AIS_Legends (13)	Pictures (32)	Images (11)	Comments (21)	Videos (1)
1	cesta	Indice	K0061	Info Ponte nelle Alpi ...	arnesi da portare il fi...	cesta [Campo Sa...	A Venezia - Vita dei...
2	cesta da foraggio	bigoncia	K1138	Info Ponte nelle Alpi ...	cesta [AIS K1492a ...	cesta [Cavarzere]	
3	cesta da letame	carro a 2 ruote	K1140	Info Ponte nelle Alpi ...	cesta [Boe 1267]	cesta [Cencenighe]	
4	cesta per gli animali	cesta da foraggio	K1179	Info Ponte nelle Alpi ...	cesta [Tisato - 2010...	cesta [Cerea]	
5		cesta	K1220	Info Ponte nelle Alpi ...	cesta da foraggio [A...	cesta [Padola]	
6		cestino	K1225	Info Ponte nelle Alpi ...	cesta da portar chio...	cesta [San Stino ...	
7		cestone	K1226	Pel 0327	cesta per la quaglia [...	cesta [Teolo]	
8		collare	K1414	Pel 1556	cestone [AIS K149...	cesta [Vas]	
9		culla	K1481	Pel 1560	cestone [AIS K149...	cesta [Venezia]	
10		gerla	K1490	Pel 4128	culla [AIS K0061]	cesta [Vicenza]	
11		madia	K1492	Pel 4135	gerla [AIS K1179]	cesta da foraggio...	
12		mastello da bucato	K1523	Pel 5586		cesta da foraggio...	
13		paglia	K1526	Pel 5903		cesta da foraggio...	
14		paniere		Pel 5907		cesta da foraggio...	
15		stia		Sch 0380		cesta da foraggio...	
16		tagliere della polenta		Sch 0564		cesta da foraggio...	
17				Sch 0570		cestino [Venezia]	
18				Sch 0574		come l'astuto ser...	
19				Sch 0575		stia [Belluno]	
20				Sch 0579		vaglio [San Stino ...	

Fig. 7 - Risultati della ricerca della parola "cesta" in tutti i documenti AMDV.

Si mostrerano infine le schede di commento lessicale-etimologico che rappresentano uno dei contributi più fondamentali di questa ricerca, in quanto mostrano il substrato che spiega analogie e differenze del lessico

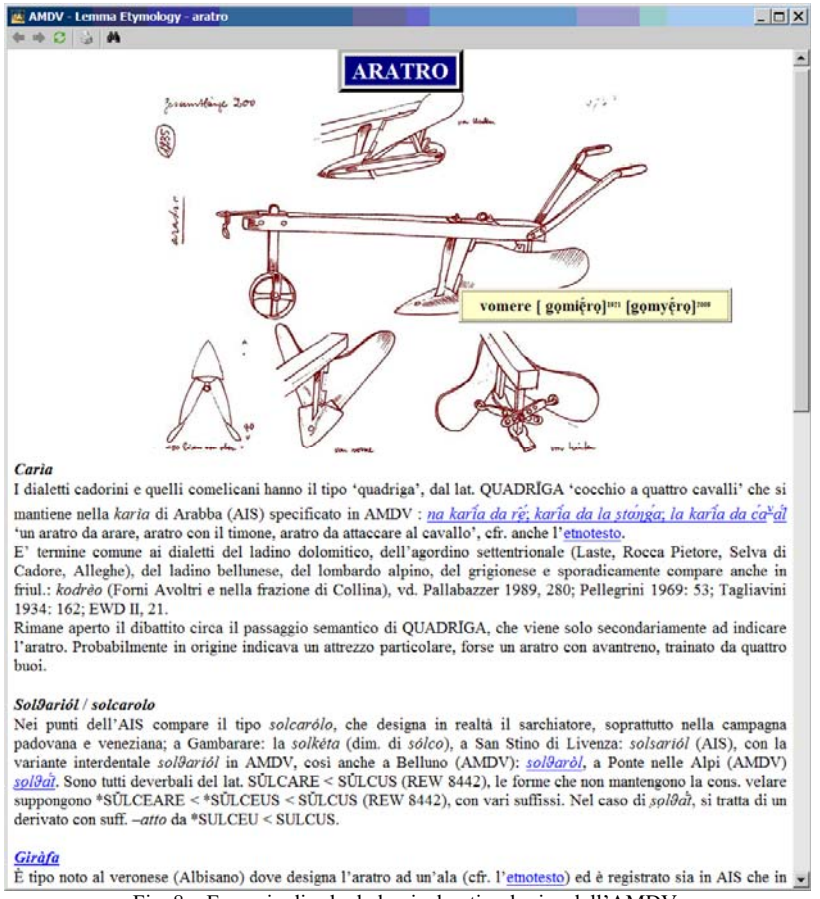


Fig. 8 – Esempio di scheda lessicale-etimologica dell'AMDV.

1. Goebl, H., (1994). *L'Atlas linguistique du ladin central et des dialectes limitrophes (première partie, ALD-I)*. In: Pilar Garcia, Mouton, (ed.): *Geolingüística. Trabajos europeos*, Madrid, 155-168, http://ald2.sbg.ac.at/a/index.php/en/the-project/
2. Telmon, T., - Canobbio, S., (ed.) (1985). *Atlante Linguistico ed Etnografico del Piemonte Occidentale*. Regione Piemonte, CELID, Torino, http://www.alepo.unito.it/default.htm
3. Kattenbusch, D., (1995). *Atlas parlant de l'Italie par régions: VIVALDI,* in: *Estudis de lingüística i filologia oferts a Antoni M. Badia i Margarit*, Barcelona 1995, 443-455, http://www2.hu-berlin.de/vivaldi/
4. Mott A., Kezich G., Tisato G., (2003). *Il Trentino dei contadini. Piccolo atlante sonoro della cultura materiale*. Museo degli Usi e Costumi della Gente Trentina, San Michele all'Adige (TN), http://www.museosanmichele.it/editoria/editNov/CDTrentino.html.
5. Tisato, G., (2010). *NavigAIS – AIS Digital Atlas and Navigation Software*, VI Convegno AISV (Associazione Italiana delle Scienze della Voce) 2010, Napoli, 451-461, http://www3.pd.istc.cnr.it/navigais

A New Language and a New Voice for MaryTTS

Fabio Tesser, Giulio Paci, Giacomo Sommavilla, Piero Cosi

ISTC CNR - UOS Padova
Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche
Unità Organizzativa di Supporto di Padova
Italy
[*fabio.tesser, giulio.paci, giacomo.sommavilla, piero.cosi*]*@pd.istc.cnr.it*

ABSTRACT

This paper describes the development of the Italian modules and the building of new Italian female voice for the MARY (Modular Architecture for Research on speech sYnthesis) Text-To-Speech synthesis system which was originally developed for the German language (Schröder, and Trouvain, 2001).

MARY TTS is a flexible and modular tool for research, development and teaching in the domain of text-to-speech (TTS) synthesis (Schröder and Trouvain, 2003).

Our activities were focused on the creation of a new Italian female voice by using a multilingual "Voice Creation Toolkit" (Pammi et al., 2010) for the MARY TTS Platform, whose workflow, illustrating the steps required to add support for a new language from scratch, is illustrated in Figure 1. As underlined by their authors (Pammi et al., 2010), "…Two main tasks can be distinguished: (i) building at least a basic set of natural language processing (NLP) components for the new language, carrying out tasks such as tokenization and phonemic transcription (left branch in Figure 1); (ii) the creation of a voice in the new language (right branch in Figure 1). …".

In particular, this toolkit, includes graphical user interfaces (GUIs) for most of the common tasks required in the creation of a synthetic voice, this facilitating the understanding of the whole process. This toolkit aims to simplify the task of building new synthesis voices so that users who do not have detailed technical knowledge of speech synthesis can build their own voices. The Voice Import Tools cover the following steps in voice building:

- Feature Extraction from Acoustic Data
- Feature Vector Extraction from Text Data
- Automatic Labeling (ehmm from Festvox 2.4-current1)
- Unit Selection voice building
- HMM-based voice building (SPTK 3.2 [13], HTS 2.1 [6])
- Voice Installation to MARY

As for our Italian voice, we started with the porting of some of the existing Italian FESTIVAL (Black etal., 1999) TTS modules (Cosi et al., 2001). An Italian lexicon has been adapted for MARY converting the Italian FESTIVAL lexicon and an Italian Letter To Sound rules have been obtained together with a simple Part Of Speech Tagger.

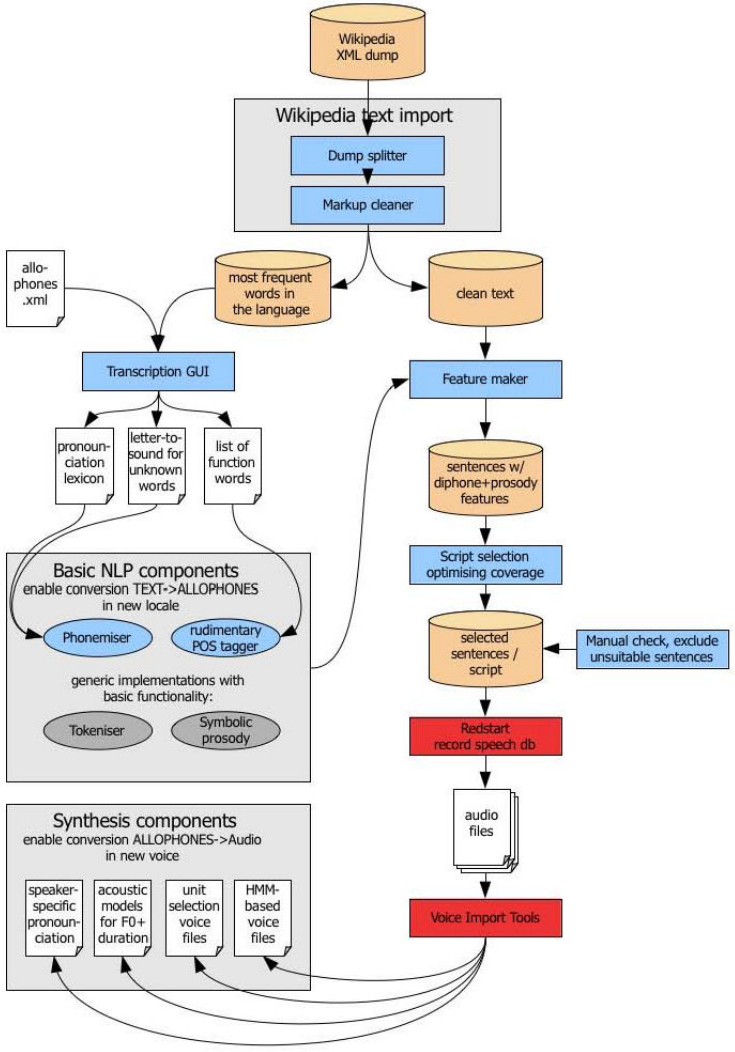


Figure 1: Workflow for multilingual voice creation in MARY TTS (http://mary.opendfki.de/wiki/VoiceImportToolsTutorial)

The first step to face with in order to add a new language in a TTS system is that of fixing the allophone set for the new language. The SAMPA (SAMPA, 1989). alphabet has been chosen because it is simple, it is a well distributed standard and because it was already used in the Italian FESTIVAL voices. The words' pronunciation can be

obtained directly from the pronunciation lexicon or, for the words not present in the lexicon, from the Letter To Sound rules. An Italian lexicon for MARY has been created converting the Italian Festival lexicon, containing around 450K words and their transcriptions, into the format specified by MARY and successively compiled into its efficient finite state transducer form. Afterwards the Letter To Sound rules have been obtained using a MARY procedure to automatically get LTS rules from lexicon examples.

A context dependent Part-of-speech (POS) tagger has been developed to predict whether words are nouns, verbs, or other grammatical classes depending on their surrounding context. Some manually annotated POS data for Italian has been kindly provided (TANL, 2008), (Zanchetta and Baroni, 2005). This corpus contains 4000 sentences for a total of 113K words, annotated with 36 POS classes (see Table 1) using the TANL tagset (TANL, 2008). This data has been used in order to train the an Italian OpenNLP POS tagger (OPENNLP, 2012) using the Maximum Entropy model (Ratnaparkhi, 1997).

Table 1. TANL tagset. Description of the Part-of-Speech tags.

Value	Description
A	adjective
AP	possessive adjective
B	adverb
BN	negation adverb
CC	coordinate conjunction
CS	subordinate conjunction
DD	demonstrative determiner
DE	exclamative determiner
DI	indefinite determiner
DQ	interrogative determiner
DR	relative determiner
E	preposition
EA	articulated preposition
FB	balanced punctuation
FC	clause boundary punctuation
FF	comma
FS	sentence boundary punctuation
I	interjection
N	cardinal number
NO	ordinal number
PC	clitic pronoun
PD	demonstrative pronoun
PE	personal pronoun
PI	indefinite pronoun
PP	possessive pronoun
PQ	interrogative pronoun

Value	Description
PR	relative pronoun
RD	determinative article
RI	indeterminative article
S	common noun
SA	abbreviation
SP	proper noun
T	predeterminer
V	main verb
VA	auxiliary verb
VM	modal verb
X	residual class
SW	foreign word

These components, together with a generic tokeniser and a generic rule-based prediction of symbolic prosody (TOBI) (Silverman et al., 1992), are able to predict a symbolic representation of speech, efficiently represented in the MARY XML language.

An automatic procedure based on the analysis of the freely available Wikipedia dumps (see Table 2) for optimal text selection able to insure good phonetic and prosodic coverage, has been applied for Italian, and, finally, a new Italian female voice (Lucia) has been created using the Voice Import Tools by recording in a quasi soundproof chamber around 1400 sentences automatically extracted by the automatic selection procedure (see Table 3) and uttered by a young Italian native female speaker. The original sentence selection procedure has been modified in order to select only those sentences for which it was possible to obtain a phonetic transcription using only the lexicon. The final text selection has been obtained by 4 iterations 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 2: Description of the Italian Text corpus for Mary TTS.

Feature	Description
Wikipedia dump date	2011/08/15 (201108151941131343006 2)
DB Size (sent.)	1400
Coverage method	SimpleDiphones+SimpleProsody

Table 3: Description of the Lucia TTS recording corpus.

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

Both a Unit Selection and an HMM voice have been created using the Voice Import Tools and the resulting voices were 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 little bit lower audio quality, but it has an higher intelligibility.

A Vocal-Tract-Scaler has also been applied to simulate a child-like voice which has been chosen as suitable voice for the NAO robot during the ALIZ-E Project experiments. Moreover, various experiments on spoken output prosody modification targeting “emotional” or “focus/prominence” modeling have been exploited using symbolic mark-up of speech rate, pitch and contour.

Acknowledgements

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: Text-to-Speech, Speech Synthesis, Markup languages, Teaching in Speech Technology, Emotions

References

Black, A., Taylor, P., and Caley, R. (1999). *Festival speech synthesis system, edition 1.4*. Technical report, Centre for Speech Technology Research, University of Edinburgh, UK. http://www.cstr.ed.ac.uk/projects/festival.

Cosi, P., Tesser, F., Gretter, R., and Avesani, C. (2001), "Festival Speaks Italian!", *Proceedings of Eurospeech 2001*, Aalborg, Denmark, September 3-7, 2001, 509-512.

OpenNLP. (2010). http://opennlp.apache.org/

Pammi, S., Charfuelan, M., and Schröder, M. (2010). "Multilingual Voice Creation Toolkit for the MARY TTS Platform", in *Proceedings of Language Resources and Evaluation Conference, LREC 2010*, 17-23 May 2010, Malta.

Ratnaparkhi, A. (1997). A Simple Introduction to Maximum Entropy Models for Natural Language Processing. *IRCS Technical Reports Series*. University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-97-08.

SAMPA for Italian. (1989). http://www.phon.ucl.ac.uk/home/sampa/italian.htm, 1989 (accessed February 22, 2011).

Schröder, M., and Trouvain, J. (2001). The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. In Proceedings of the *4th ISCA Workshop on Speech Synthesis, Blair Atholl, Scotland.*

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, pp 365-377.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). ToBI: A standard for labeling English prosody. In *Proceedings of the 2nd International Conference of Spoken Language Processing*, Banff, Canada, 1992, pp. 867–870.

Sproat, R., Hunt, A., Ostendorf, M., Taylor, P., Black, A., Lenzo, K., and Edgington, M. (1998). SABLE: A standard for TTS markup. In *Proceedings of the 5th International Conference of Spoken Language Processing*, Sydney, Australia, pp. 1719–1724 .

TANL (2008). Text Analytics and Natural Language processing, by Medialab, Attardi G., et al. Project Analisi di Testi per il Semantic Web e il Question Answering, 2008. http://medialab.di.unipi.it/wiki/SemaWiki.

Zanchetta, E., and Baroni, M. (2005). Morph-it! A free corpus-based morphological resource for the Italian language. In *Proceedings of Corpus Linguistics 2005*, January 2005.*

Prime indagini su un corpus di dialogo uomo-macchina raccolto nell'ambito del progetto Speaky Acu-tattile

Poroli Fabio, Delogu Cristina, Falcone Mauro, Paoloni Andrea, Todisco Massimiliano

La diffusione di sistemi automatici di dialogo, che procede di pari passo con la realizzazione di sistemi sempre più *user-oriented*, ha portato alla necessità di approfondire gli aspetti legati alla situazione comunicativa uomo-macchina, differente dalle conversazioni ordinarie per *setting* (la presenza di un interlocutore non umano, cfr. Bazzanella 2005, le cui caratteristiche determinano, inoltre, la rilevanza o meno di altri tratti che caratterizzano la situazione, come la condivisione spaziale e temporale) e per tipo di interazione intrattenuta (*practical dialogues*, Allen 2000: 2).

La nostra ricerca si inquadra nell'ambito del Progetto Speaky Acu-tattile, una nuova piattaforma inclusiva di assistente intelligente vocale multicanale, a cui la Fondazione Bordoni partecipa contribuendo alla progettazione di un sistema di dialogo automatico (VUI, *Voice User Interface*) per aiutare alcune tipologie di persone (anziani, ciechi, disabili motori) nell'uso del PC, nella navigazione sul web, nell'uso della televisione e nella gestione della casa (quando integrato in un sistema domotico controllabile sia dentro casa, sia via telefono).

Vista la priorità data all'accessibilità, la progettazione del sistema di dialogo è fortemente orientata verso l'utente e verso il suo comportamento nella situazione uomo-macchina. Per questo si è scelto, per la raccolta del *corpus*, di usare la tecnica del Mago di Oz: una simulazione che consiste nel far interagire un uomo con una macchina "finta", impersonata dallo sperimentatore (chiamato *wizard*), senza che il primo ne sia a conoscenza, fornendo così dati sulle interazioni (necessari per progettare il sistema) prima ancora di avere a disposizione il sistema (Fraser – Gilbert 1991).

L'esperimento, che consente di prescindere dalle possibilità tecnologiche degli attuali sistemi di dialogo, richiede comunque la definizione di alcune variabili, legate al sistema che si intende progettare e agli scenari in cui sarà coinvolto. Nel nostro esperimento sono stati definiti 48 compiti (suddivisi in quattro domini: assistenza sanitaria, domotica, intrattenimento, servizi esterni) da far svolgere a 24 soggetti coinvolti (8 per ogni categoria di utenza), per un totale di 384 dialoghi registrati. Sul lato della comprensione non sono state imposte particolari restrizioni al wizard (a parte le richieste fuori dominio), mentre sul lato della produzione è stato definito *a priori* un protocollo che lega il wizard a un comportamento omogeneo e "naturale" con ogni soggetto coinvolto nell'esperimento, consentendogli inoltre una rapida reazione agli input.

A tale scopo sono stati pre-stilati gli output usati dal wizard, e successivamente organizzati in alberi di dialogo che ricalcano il compito (formalizzato secondo un'architettura *frame-based*, ovvero per il suo completamento è necessario che l'input contenga alcuni dati necessari predefiniti) e la macrostruttura del "dialogo pratico", il tipo di interazione che generalmente caratterizza la situazione comunicativa uomo-macchina. I dialoghi pratici possono essere distinti, infatti, da quelli definiti, per comodità, "ordinari" (Leech 2005), per il forte orientamento verso la risoluzione di un compito (la richiesta di informazioni, l'acquisto di un prodotto, ecc.) e per la delimitazione netta dei domini su cui vertono. Il livellamento sul compito e la limitazione del dominio porta, quindi, da una parte alla forte riduzione della variabilità linguistica e dall'altra a una macrostruttura *grosso modo* omogenea che prescinde dai domini e che si può

schematizzare in cinque fasi (Patzold et al. 1995, Alexandersson et al. 1997): (1) saluti, (2) apertura del compito, (3) negoziazione, (4) chiusura del compito, (5) saluti.

La simulazione rende possibile, inoltre, l'iniziativa mista: come per le conversazioni ordinarie, in cui è normale che l'iniziativa (o il controllo della conversazione) slitti di parlante in parlante durante l'interazione (Walker & Whittaker 1990), determinando di volta in volta chi, con il proprio turno, gestisce, in parte, il turno successivo (Burke 1994: 99) e, localmente, la risoluzione del compito (Novick – Sutton 1997), così il soggetto può rispondere a una domanda precisa fornendo più informazioni di quelle "obbligate" dalla singola richiesta del sistema o correggere il sistema direttamente di fronte a richieste di conferma su informazioni erroneamente acquisite (ad esempio: Utente: "Voglio andare da Roma a Milano" – Sistema: "Vuoi andare da Roma a Merano?" – U: "No, da Roma a Milano").

Una volta terminata la raccolta si procederà con una prima indagine volta a modellizzare il comportamento dei soggetti, determinato dall'idea che il parlante umano della macchina, inquadrabile, come già visto in letteratura, tra due estremi: la macchina come interlocutore umano e la macchina come interfaccia applicativa (Edlund et al. 2008). A tale scopo si porrà particolare attenzione ad alcuni parametri, come la gestione dell'iniziativa da parte dell'utente (Fischer - Bateman 2006) nei turni in cui è possibile prenderne il controllo, la presenza della funzione della cortesia, le riformulazioni in caso di errore, e la più generale tendenza alla semplificazione linguistica negli input (a partire dalla morfosintassi, cfr. Danieli 2004, fino alla pragmatica, come la presenza o meno di segnali discorsivi).

Riferimenti bibliografici

ALEXANDERSSON ET AL. 1997 = J. Alexandersson, B. Buschbeck-Wolf, T. Fujinami, M. Kipp, S. Koch, E. Maier, N. Reithinger, B. Schmitz, M. Siegel, *Dialogue Acts in VERBMOBIL-2*, Verbmobil-Report 226, DFKI Saarbrucken, Universitat Stuttgart, Technische Universitat Berlin, Universitat des Saarlandes

ALLEN 2000 = J. F. Allen, D. Byron , M. Dzikovska, G. Ferguson, L. Galescu, A. Stent, *An architecture for a generic dialogue shell*, Natural Language Engineering, vol. 6 (3), pp. 1-16.

BAZZANELLA 2005 = C. Bazzanella, *Parlato dialogico e contesti di interazione,* in Hölker K., Maaß C. (a cura di), Aspetti dell'italiano parlato, Münster, Hamburg, London, LIT, Verlag, pp. 1-22

BURKE 1994 = P. Burke, *Segmentation and control of a dissertation defense*, in A. Grimshaw (a cura di), *What's going on here? Complementary studies of talk*, Norwood, Ablex, pp. 95-124.

DANIELI 2004 = M. Danieli, *Il parlato telegrafico tra persone e sistemi artificiali*, in Atti del convegno nazionale Associazione Italiana di Scienze della Voce 13-15 febbraio 2003, Napoli, D'Auria Editore

EDLUND ET AL. 2008 = J. Edlund, J. Gustafson, M. Heldner, A. Hjalmarsson, *Towards human-like spoken dialogue systems*, Speech Communication, 50, pp. 630-645.

FISCHER – BATEMAN 2006 = K. Fischer, J. A. Bateman, *Keeping the initiative: An empirically motivated approach to predicting user-initiated dialogue contributions in HCI*, Proceedings of the EACL'06, Trento.

FRASER – GILBERT 1991 = N. Fraser, N. Gilbert, *Simulating speech systems*, Computer Speech and Language, 5, pp. 81-99

NOVICK – SUTTON 1997 = D. Novick, S. Sutton, *What is mixed-initiative interaction?*, Papers from the 1997 AAAI Spring Symposium on Computational Models for Mixed Initiative Interaction, Stanford University, Technical Report

PÄTZOLD ET AL. 1995 = M. Pätzold, M. Scheffers, A.P. Simpson, W. Thon, *Controlled elicitation and processing of spontaneous speech in Verbmobil*, Proc. XIIIth ICPhS, vol. 3, Stockholm, pp. 314-317.

WALKER & WHITTAKER 1990 = Walker, M., and Whittaker, S. (1990). *Mixed initiative in dialogue: An investigation into discourse segmentation*, *Proceedings of the 28th Meeting of the ACL*, pp. 70-78.

Tecnologie del parlato in sala operatoria: il progetto DOMHOS

Mirco Ravanelli[1], Matteo Benetti[2], Giovanni Pedrotti[3],
Marco Matassoni[1], Luca Cristoforetti[1], Maurizio Omologo[1]
[1] Fondazione Bruno Kessler, Trento
[2] Unihospital, Trento
[3] Ospedale S. Chiara, Trento
mravanelli@fbk.eu

Le tecnologie del parlato possono oggi offrire un utile supporto in particolari applicazioni: dettatura, trascrizione, risponditori automatici. Negli ultimi anni, grazie al grande sviluppo del potenziale computazionale e al continuo miglioramento delle tecniche di riconoscimento del parlato, tali sistemi stanno guadagnando sempre maggiori spazi applicativi. La maggior parte di questi sistemi, tuttavia, è ancora basato sull'utilizzo di microfoni vicini al parlatore, una modalità che se da una parte permette di ottenere prestazioni molto interessanti, dall'altra pone all'utente il vincolo di parlare molto vicino al microfono.

Stanno recentemente comparendo nuovi campi d'impiego in cui l'utente non vuole o non può essere vincolato dal microfono per diversi motivi; in tali situazioni risulta vantaggioso studiare soluzioni in cui l'interazione vocale opera in modalità cosiddetta *hands-free*. Alcuni protocolli previsti nell'ambito chirurgico, per esempio, sembrano naturalmente predisposti per essere accoppiati a sistemi di riconoscimento automatico del parlato operanti attraverso quest'ultima modalità.

Il progetto DOMHOS, iniziato nel gennaio 2012, è guidato da un'azienda attiva nel settore di servizi a custodia della salute con il supporto tecnico-scientifico di un istituto di ricerca e coinvolge direttamente il principale ospedale della città, in particolare alcuni chirurghi del reparto di neurochirurgia. Il progetto si propone di introdurre l'interazione vocale all'interno della sala operatoria: uno scenario allo stesso tempo poco esplorato in passato e particolarmente sfidante per la tecnologia attualmente disponibile.

Nel contesto previsto l'obiettivo è duplice. Il primo obiettivo prevede di consentire all'equipe medica di operare normalmente mentre contestualmente un sistema automatico registra e trascrive gli appunti vocali dettati dal personale in sala. Tale pratica facilita la stesura del verbale operatorio, un documento obbligatorio per legge che in genere viene redatto dai chirurghi dopo l'operazione; e questo ritardo nella ricostruzione dell'operazione può pertanto determinare delle imprecisioni nel referto. Come confermato dai medici, l'utilizzo delle tecnologie vocali nell'ambito di questo scenario applicativo può essere particolarmente gradito al chirurgo, il quale al termine dell'operazione disporrà già di una prima trascrizione che dovrà solamente controllare ed eventualmente correggere. L'altra applicazione prevista è l'implementazione per mezzo delle tecnologie vocali del protocollo di *checklist*, recentemente introdotto anche in Italia dal Ministero della Salute. Tale procedura è uno strumento guida per l'esecuzione di controlli a supporto dell' equipe operatoria con la finalità di favorire in modo

sistematico l'adozione di uno standard di sicurezza in grado di prevenire errori, mortalità e complicanze post-operatorie. Attualmente la procedura della *checklist* viene eseguita grazie ad un infermiere dedicato, che nelle varie fasi dell'operazione pone opportune domande al personale medico compilando un documento cartaceo sulla base delle risposte ricevute. Il secondo scopo del progetto DOMHOS è quindi quello di studiare e sperimentare se in questo scenario applicativo l'utilizzo delle tecnologie vocali possa velocizzare e rendere più sicura l'intera procedura di *checklist*.

La modalità hands-free che deve essere adottata per entrambi gli scenari applicativi pone tuttavia numerose complicazioni rispetto alla modalità con microfoni vicini al parlatore. Una problematica di notevole rilievo è dunque costituita dal contesto ambientale di utilizzo del sistema, molto sfidante per lo stato dell'arte della tecnologia. In particolare l'elevata riverberazione e l'elevato livello di rumorosità dovuto alle numerose apparecchiature utilizzate durante l'operazione (monitor, aspiratore, ecc.) rendono necessaria l'adozione di opportuni sistemi multi-microfonici in grado di limitare il più possibile gli effetti di questi disturbi.

In tal senso, si è già equipaggiata una sala operatoria di neurochirurgia dell'ospedale con 8 microfoni in 3 diverse configurazioni sperimentali. Si stanno dunque effettuando diversi esperimenti preliminari per definire la migliore configurazione multi-microfonica, le migliori caratteristiche del front-end da sviluppare e la migliore modalità di addestramento dei modelli acustici in questo particolare contesto operativo. Oltre alle questioni citate, altre problematiche sono state considerate già nei primi mesi di attività. In particolare, grazie ai verbali operatori prodotti in passato dai vari medici dell'ospedale, è stato possibile addestrare un modello del linguaggio preliminare sufficientemente preciso per modellare i registri operatori che verranno presumibilmente dettati. Attraverso la collaborazione con i medici si stanno parallelamente definendo le migliori modalità di interazione con il sistema di riconoscimento ed i feedback che esso dovrà fornire. Dal momento che i chirurghi desiderano avere il controllo sulla registrazione stessa, la modalità che appare più idonea è quella che prevede di attivare la dettatura solo dopo la pronuncia di una certa parola chiave da parte dell'operatore. Si è pertanto previsto l'utilizzo di tecniche di *keyword-spotting* operanti in real-time ed in continuo ascolto. Le prestazioni di quest'ultimo sistema assumono un aspetto cruciale: se per la dettatura della nota vocale il medico può tollerare qualche errore, facilmente correggibile successivamente, nell'ambito del *keyword-spotting* malfunzionamenti come falsi o mancati allarmi sono decisamente più critici in quanto possono introdurre dei fastidiosi ritardi e delle inefficienze. Di notevole rilevanza ha dunque sia la scelta della parola chiave che la progettazione di una grammatica di rigetto robusta nel contesto operativo di funzionamento del sistema.

Nell'articolo completo verranno presentati dunque i risultati preliminari di questa prima fase attraverso una descrizione dell'architettura hardware e software, le caratteristiche dei dati acustici raccolti in sala operatoria e le prestazioni del sistema di riconoscimento sui segnali reali.

JULIUS ASR for Italian Children Speech

Giulio Paci, Giacomo Sommavilla, Fabio Tesser, Piero Cosi

ISTC CNR - UOS Padova
Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche
Unità Organizzativa di Supporto di Padova
Italy

[giulio.paci, giacomo.sommavilla, fabio.tesser, piero.cosi]@pd.istc.cnr.it

ABSTRACT

In this paper we describe the JULIUS ASR engine we adapted to Italian and we report on the results obtained for some children speech ASR experiments developed for the EU-FP7 Project ALIZ-E (ALIZ-E, 2012) in which we are involved.

After a comparison with Sphinx-3, we give an overview on the JULIUS' Acoustic Model (AM) training procedure and on the different types of Language Model (LM) supported. Finally we end describing the acoustic model training procedures and the language models design we adopted in a preliminary ASR setup for ALIZ-E experiments.

Open-Source Large Vocabulary Continuous Speech Recognition Engine JULIUS (Lee et.al. 2001), (JULIUS, 2012) 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-the-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 include small memory footprint, core engine as a separate C library, modular configuration file structure, parallel configurations decoding. Moreover, it is Open-source software.

We tried also CMU Sphinx-3 (Lee et al., 1990) for speech recognition. However it has been difficult to implement live decoding and run-time features with it and Sphinx-3 upstream code is no longer maintained. So Sphinx-3 has been replaced with Open-Source Large Vocabulary CSR Engine JULIUS as the ASR engine in our recognition experiments with children speech. With JULIUS, it has proven to be very easy to implement the desired features and to integrate them into the system. Also, in comparison with Sphinx-3, JULIUS decoder API is very well designed, its language model can be swapped at run-time and its configuration is modular.

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. The HTK tools (Young et al., 2006) have been used for this task, following the Voxforge HTK training for JULIUS tutorial (VoxForge, 2012).

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.

The JULIUS engine supports N-gram LMs in ARPA format and SRI-LM toolkit (Stolcke, 2002) can be used to train simple LMs.

The JULIUS engine distribution includes some tools that allow to express a Grammar in a simple format and then to convert it 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 (JULIUS, 2012). 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 al., 2006).

Two Italian Corpora have been tested so far with HTK and JULIUS:

- the training data provided for the EVALITA 2011 Forced Alignment task (Cutugno et al., 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 al., 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.

The Quiz questions and answers database of the Quiz Game ALIZ-E scenario has been used as training material for this "question recognition" 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. A simple model for Quiz answers recognition was 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.

Table 3 shows the results of ASR applied to 64 utterances (561 words), where a child poses a quiz question to the NAO robot. On average, we get 74% correct words, 11.5% inserted words and 38% WER. Taking the ASR hypotheses as input to a specific Natural Language Understanding (NLU) module, specifically designed and implemented for ALIZ-E, questions were correctly identified by fuzzy matching against the quiz database contents.

Experiments ID	#Snt	#Wrd	WCR%	Ins%	WER%
1	4	22	77.3	31.8	54.5
2	6	82	75.5	31.7	56.1
3	5	40	80.0	17.5	37.5
4	7	63	74.6	3.2	28.6
5	15	114	90.4	4.4	14.0
6	4	49	59.2	8.2	49.0
7	12	107	62.6	5.6	43.0
8	11	84	67.9	8.3	40.5
Total	**64**	**561**	**73.8**	**11.4**	**37.6**

Table 3: Preliminary ASR results on quiz question recognition

This is an encouraging first result and further experiments will show whether this level of ASR+NLU performance suffices to sustain the interaction.

Acknowledgements

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: ASR, Children Speech, JULIUS

References

ALIZ-E (2012), http://ALIZ-E.org/.

Bisani, M., and Ney, H. (2008). "Joint-sequence models for grapheme-to-phoneme conversion". In Speech Communication 50.5 (2008), pp. 434-451. issn: 0167-6393. doi: 10.1016/j.specom.2008.01.002. http://www.sciencedirect.com/science/article/pii/S0167639308000046.

Cutugno, F., Origlia, A., and Seppi, D. (2012) "EVALITA 2011: Forced alignment task". Tech. rep. 2012. http://www.evalita.it/sites/evalita.fbk.eu/files/working_notes2011/Forced_Alignment/FORCED_ORGANIZERS.pdf.

Gerosa, M., Giuliani, D., and Brugnara, F. (2007). "Acoustic variability and automatic recognition of children's speech". In Speech Communication 49 (2007), 847-860.

JULIUS development team. (2012) "*Open-Source Large Vocabulary CSR Engine JULIUS*". Mar. 2012. url: http://julius.sourceforge.jp/.

Lee, A., Kawahara, T., and Shikano, K. (2001). “JULIUS - an open source real-time large vocabulary recognition engine". In *Proceedings of INTERSPEECH 2001*, 1691-1694.

Lee, K.,F., Hon, H.,W., and Reddy R. (1990), “*An overview of the SPHINX speech recognition system*". In *IEEE Transactions on Acoustics, Speech and Signal Processing* 38.1 (1990), 35-45.

Stolcke A. (2002). “SRILM - An Extensible Language Modeling Toolkit". In Proceedings of ICSLP-2002, International Conference on Spoken Language Processing, 2002. 901--904.

VoxForge (2012a). *Tutorial: Create Acoustic Model - Manually*. Mar. 2012. http://www. voxforge.org/home/dev/acousticmodels/linux/create/htkjulius/tutorial.

VoxForge. (2012b). Free Speech... Recognition (Linux, Windows and Mac). March 2012. http://www.voxforge.org/.

Young, S.J., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. (Andrew), Moore, G., Odell, J., Ollason, D., Povey, D., and Valtchev, V. (2006), *The HTK Book, version 3.4*.1. Cambridge, UK: Cambridge University Engineering Department, 2006.

Prime Note sulla Valutazione Soggettiva dell'Intelligibilità

Giovanni Costantini[1,2], Andrea Paoloni[3], Massimiliano Todisco[1,3]

[1] Dipartimento di Ingegneria Elettronica, University of Rome “Tor Vergata”, Roma, Italia
[2] Istituto di Acustica “O. M. Corbino”, Roma, Italia
[3] Fondazione “Ugo Bordoni”, Roma, Italia

Introduzione

Come è noto la “qualità” di un segnale audio viene valutata in base a tre caratteristiche: l’intelligibilità, ovvero la possibilità di comprendere con precisione quello che viene detto, la naturalezza, ovvero quanto il segnale corrisponda a quello ottenibile nell'ascolto diretto e qualità in senso proprio, ossia quanto il segnale sia gradevole.
Queste definizioni sono state formulate pensando all'analisi delle prestazioni di un sistema di trasmissione; in altri termini si voleva capire quale fosse la qualità audio che un sistema di trasmissione con determinate caratteristiche (banda passante, rapporto segnale rumore, tipo di codifica) era in grado di garantire. La misura dell'intelligibilità di conseguenza non è altro che la misura della differenza tra l’intelligibilità del segnale in uscita rispetto a quella del segnale di ingresso.
Esistono tuttavia alcune applicazioni, in particolare quella forense, per le quali quello che si vorrebbe misurare è l'intelligibilità di un segnale senza avere a disposizione il messaggio di partenza, né in forma di testo né tantomeno come segnale audio.
Il problema di valutare l'intelligibilità di un segnale "single side” ovvero avendo a disposizione solo il file audio corrotto di cui si intende valutare l'intelligibilità, è molto complesso perché l'intelligibilità residua dipende da molti parametri: la larghezza di banda, il rapporto segnale rumore, il tipo di rumore, il tipo di segnale, la distorsione, la codifica. Inoltre i parametri che abbiamo elencato non sono di facile stima a se si deve partire dal segnale stesso di cui si vuole conoscere intelligibilità.
Nel presente lavoro ci concentreremo sul tipo di segnale, ovvero su come, a parità di rapporto segnale rumore, la differente tipologia di segnali porti a differenti valutazioni di intelligibilità. Secondo la letteratura [1] a parità di rapporto segnale rumore l’intelligibilità è minore per i logatomi (ossia per le sillabe di cui si vuole conoscere la consonante iniziale), maggiore per le parole isolate e ancora maggiore per le frasi. Se consideriamo tuttavia che una frase contiene almeno una trentina di fonemi e se la consideriamo errata se uno solo di questi non viene correttamente riconosciuto, sembrerebbe logico un risultato opposto, ossia che sono proprio le frasi a presentare la minore intelligibilità.

Obiettivo del lavoro

Obiettivo primario del presente lavoro è valutare l’intelligibilità utilizzando le diverse tipologie di segnale sopra elencate (logatomi, frasi, parole) in funzione di differenti rapporti S/N. Un altro importante obiettivo è costruire un corpus di segnali di cui sia nota l’intelligibilità, misurata soggettivamente, al fine di poter valutare le prestazioni di sistemi oggettivi di misura.
Il disturbo preso in considerazione è di tipo “mormorio” o Babble in quanto questa tipologia di rumore, che imita il rumore provocato de diverse persone che parlano tra loro è quello che meglio rappresenta le reali situazioni di disturbo.

I corpora

I corpora utilizzati per i test soggettivi sono stati costruiti utilizzando il corpus del progetto europeo SAM EUROM 1 e CLIPS. In particolare, sono state utilizzate 10 frasi, 15 parole e 19 fonemi della lingua italiana. Il segnale, che era stato a suo tempo equalizzato per quanto attiene al livello audio, è stato poi degradato con rumore di tipo additivo in modo da ottenere 5 diversi gradi di rapporto segnale / rumore (S / N = 6, 3, 0, -3, -6 dB). La Tabella I mostra i corpora utilizzati.

Misure soggettive di intellegibilità

I corpora del parlato sono stati quindi sottoposti a un gruppo di 10 ascoltatori, al fine di ottenere il risultato di intelligibilità soggettiva, usando il software sviluppato appositamente per questo scopo in ambiente Max/MSP. 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 una delle 3 sessioni composta da Phonemes, Words e Sentences, quindi si attiva il tasto “play” per la riproduzione dello stimolo

sonoro, infine si scrive quello che si ritiene di aver ascoltato. Al termine di ogni sessione viene registrato il testo contenente i risultati forniti dal soggetto. Esiste inoltre la possibilità di un addestramento che consente di comprendere meglio la prova e di regolare livello del segnale audio. I risultati dei test soggettivi sono mostrati in Fig. 1.

Tabella I

	SENTENCES		WORDS		PHONEMES
1	IL FULMINE HA COLPITO L'ALBERO	1	AEROPLANO	1	PALE
2	QUEL CANTANTE HA UNA BELLA VOCE	2	BIGLIETTO	2	TALE
3	ABBIAMO PREPARATO UNA TORTA MOLTO DOLCE	3	COLAZIONE	3	CALE
4	IL TRENO PARTIRÀ IN RITARDO	4	ELEGANTE	4	BALE
5	UN MESE DI VACANZA PASSA IN FRETTA	5	FATICA	5	DALE
6	QUEI SIGNORI NON SANNO MAI COSA FARE NÉ DOVE ANDARE	6	SETTIMANA	6	GALE
7	NEL GRANDE PARCO UN BAMBINO GIOCAVA CON SUO PADRE	7	GINOCCHIO	7	FALE
8	LA RAGAZZA CHE È APPENA ENTRATA, NON LA CONOSCO	8	GOVERNO	8	SALE
9	CHIAMAI IL MEDICO PERCHÉ AVEVO MALE AGLI OCCHI	9	INDUSTRIA	9	SCIALE
10	QUEL RAGAZZO NON DICE MAI LA VERITÀ	10	MACCHINA	10	VALE
		11	MODELLO	11	MALE
		12	OROLOGIO	12	NALE
		13	PADRONE	13	GNALE
		14	PRINCIPE	14	GLIALE
		15	RAGAZZO	15	LALE
				16	RALE
				17	ZALE
				18	CIALE
				19	GIALE

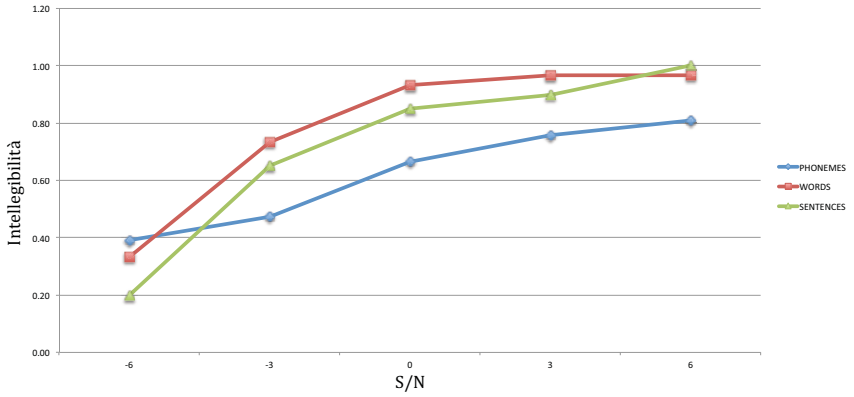


Figura 1: Risultati dei test soggettivi

Conclusioni e futuri sviluppi

I risultati riassunti in Figura 1 portano a ritenere che non ci sia sostanziale differenza tra l'intelligibilità misurata usando parole singole e quella stimata utilizzando brevi frasi. I fonemi invece sembrano effettivamente fornire una misura più sensibile dell'intelligibilità, in quanto passerebbero più lentamente da basse percentuali di corretta identificazione ad alte percentuali di corretta identificazione, fornendo maggiori informazioni qualora si voglia utilizzare tali risultati in funzioni diagnostiche. Ulteriori studi potrebbero meglio precisare questo aspetto ma resta aperto il problema di stabilire quale sia l'attendibilità di una determinata trascrizione di una frase di difficile interpretazione (disputed utterance). Si procederà quindi a individuare una procedura per stimare i rapporti di verisimiglianza di una determinata trascrizione.

Bibliografia

[1] ISO/TR 4870 Acoustics – The construction and calibration of speech intelligibilità tests, 1991
[2] Romito L., Il con testo, l'intelligibilità, il rapport segnale rumore, Atti 1° Convegno AISV, Padova 2-4 dicembre 2004.
[3] Herman J.M. Steeneken, "The Measurement of Speech Intelligibility ", TNO Human Factors, Soesterberg, the Netherlands.
[4] Ma J., Hu y., Loizou C.: "Objective measures for predicting speech intelligibility in moist conditions based on new band importance functions" JASA 125, May 2009.

RICONOSCIMENTO EMOTIVO NELLE APPLICAZIONI MULTIMODALI PER DISPOSITIVI MOBILI

Antonio Caso, Francesco Cutugno, Antonio Origlia

antocaso@gmail.com, cutugno@na.infn.it, antonio.origlia@unina.it

Tra le varie applicazioni dell'affective computing, la possibilità di progettare sistemi capaci di auto-valutare la propria performance durante l'interazione è uno dei più studiati. Questo perché le moderne interfacce delle applicazioni tentano sempre più frequentemente di proporre modalità di interazione quanto più possibile *naturali* all'utente. Questo comporta l'introduzione, per esempio, di sistemi di dialogo e interfacce gestuali il cui scopo è quello di consentire agli utenti di porre richieste in termini intuitivi e, appunto, naturali.
Il merito principale di questo recente orientamento allo sviluppo delle interfacce ricade nella volontà di rendere le macchine capaci di adattarsi al modo di comunicare degli esseri umani invece di richiedere agli utenti di adattare il modo di esprimere le proprie necessità all'interfaccia della macchina. Al contrario, il rischio principale che si corre è quello di indisporre più rapidamente gli utenti nel caso in cui la qualità dell'interazione non sia soddisfacente. Per questo motivo, è importante dotare questo tipo di sistemi di moduli di autovalutazione che consentano la rilevazione di segnali di impazienza da parte dell'utente. Questi segnali vengono emessi in maniera spontanea nel momento in cui il sistema pone l'interazione su un livello naturale ed è importante rilevarli per consentire l'avvio di procedure di "recupero" come la deviazione della chiamata ad un operatore umano, nel caso dei call center intelligenti (Herm et al., 2008). Per quanto riguarda gli smartphones, la disponibilità di servizi internet sui dispositivi mobili ha fatto sì che le interfacce naturali diventassero la scelta primaria soprattutto per quel che riguarda i servizi di ricerca (Ehlen, 2011).

In questo lavoro presentiamo una applicazione del riconoscimento dello stress emotivo mirato ad applicare questo secondo tipo di strategia in una app per smartphones e tablet Android. Il caso di studio proposto riguarda una applicazione multimodale per smartphone progettata per dare indicazioni in ambiente urbano riguardo il sistema di trasporti pubblici.

Fra le caratteristiche principali del sistema risalta la possibilità di effettuare una richiesta verbale di disponibilità di linee di trasporto mentre un'area su una mappa viene delimitata tramite un gesto sullo schermo touch. Al ricevimento della richiesta e dopo la sua interpretazione automatica, il sistema presenta dei dati in tempo reale circa i tempi di attesa alle fermate.

La modalità primaria di interazione è fornita da un sistema di dialogo che trasforma le richieste espresse a voce in richieste all'applicazione sottostante, mentre l'interazione gestuale interviene talvolta a rendere più complessa sia l'interazione che il livello di attesa dell'utente. Il sistema di dialogo si occupa di selezionare un task tra quelli disponibili, controllare che le informazioni necessarie a completare il task siano state fornite dall'utente e, nel caso in cui mancassero delle informazioni (sia per problemi di riconoscimento che per effettiva assenza di queste), richiederle tramite sintesi vocale. A questo si affianca un modulo di analisi dello stress vocale volto a far sì che, nel caso in cui l'utente appaia spazientirsi con il sistema, venga proposta una interfaccia grafica che proponga visivamente la scelta tra i task che appaiono essere stati richiesti con la maggiore probabilità. L'input vocale è quindi analizzato sia a livello semantico, per l'estrazione delle richieste di servizio, sia a livello intonativo, per la rilevazione dell'eventuale necessità di ricorrere a forme di interazione più semplici nel caso in cui non sia possibile comprendere la richiesta dell'utente. Il sistema è progettato seguendo le direttive CARE (Coutaz et al., 1995) e risponde alle caratteristiche imposte dal W3C ai sistemi multimodali (Larson et al. 2003).

L'estrazione delle features vocali consiste di uno script PRAAT che analizza la produzione vocale ed estrae informazioni spettrali da unità sillabiche rilevate automaticamente in base al profilo dell'energia. Tali features sono quelle che più frequentemente si trovano ad essere correlate, in letteratura, con l'asse dell'attivazione nei modelli dimensionali e con la contrapposizione dell'emozione *Rabbia* con l'emozione *Neutrale* nei modelli discreti. Il compito di distinguere tra un livello alto di stress ed uno basso è delegato ad una Support Vector Machine (SVM) residente su un server remoto (risultati del lavoro descritto sono pubblicati in altri lavori degli autori qui non riportati per evitare autocitazioni).

Contrariamente a quanto fatto in precedenza, invece di impiegare corpora emotivi generici per l'addestramento della SVM, si è raccolto un corpus di produzioni emotive tramite l'uso di una app per smartphones Android. Tale corpus consiste di una parte contenente parlato emotivo letto e di una parte contenente parlato emotivo spontaneo elicitato tramite l'uso di un gioco implementato su smartphones e tablet. L'andamento del gioco viene influenzato da un operatore della cui presenza l'utente è ignaro, operando quindi con un setup di tipo Wizard of Oz (WoZ). Il gioco prevede un tempo limite di due minuti durante i quali il giocatore deve indicare a voce la posizione di una forma geometrica che compare per un breve istante sul display assumendo una divisione in quattro quadranti. Mentre all'inizio il gioco si comporta in maniera "onesta", man mano che ci si avvicina allo scadere del tempo gli errori provocati intenzionalmente si sommano alla pressione imposta dal timer per provocare reazioni maggiormente stressate. Lo scopo di questo nuovo corpus è raccogliere materiale audio registrato tramite il microfono dello smartphones in ambienti non controllati in maniera tale da addestrare il classificatore su materiale quanto più possibile simile a quello che verrà in seguito inviato dall'applicazione. Il corpus raccolto consiste al momento di circa 400 enunciati da 10 parlanti di area campana (5 maschi e 5 femmine) per la parte elicitata attraverso il gioco e di altrettanti enunciati di parlato emotivo letto provenienti dagli stessi parlanti. Il corpus è stato utilizzato per addestrare e testare un classificatore automatico basato su Support Vector Machines specializzato nella distinzione binaria *neutro/rabbia* in questo specifico strato diamesico che presenta particolari peculiarità. Il classificatore, allo stato attuale, presenta una accuratezza di circa il 75% al momento in cui si scrive sono in corso ulteriori raffinamenti.

Bibliografia

Coutaz, J., Nigay, L., Salber, D., Blandford, A., May, J., Young, R.M.: Four easy pieces for assessing the usability of multimodal interaction: the care properties. In Proc. of INTERACT. pp. 115--120 (1995)

Bodell, M., Dahl, D., Kliche, I., Larson, J., Tumuluri, R., Yudkowsky, M., Selvaraj, M., Porter, B., Raggett, D., Raman, T., Wahbe, A.: Multimodal architectures and interfaces (2011): multimodal architectures and interfaces (2011), http://www.w3.org/TR/mmi-arch/

Larson, J.A., Raman, T.V., Raggett, D., Bodell, M., Johnston, M., Kumar, S., Potter, S., Waters, K.: W3C multimodal interaction framework (2003): http://www.w3.org/TR/mmi-framework/

Herm, O., Schmitt, A., Liscombe, J., 2008. When calls go wrong: How to detect problematic calls based on log files and emotions. In Proc. of Interspeech. pp. 463--466

Ehlen, P., Johnston, M.: Multimodal local search in speak4it. In Proc. of IUI. pp. 435--436 (2011)

UN'ARCHITETTURA ROBOTICA AFFETTIVA BASATA SU SILLABE FONETICHE

Antonio Origlia, Francesco Cutugno

antonio.origlia@unina.it, cutugno@na.infn.it

L'analisi della voce per l'estrazione di features relative allo stato emotivo di un parlante è stata oggetto di studi sempre più frequenti negli ultimi anni. Al crescere dell'interesse nei confronti dell'informazione contenuta nella componente intonativa del parlato, la possibilità di realizzare sistemi automatici che tenessero conto di questo tipo di informazione è andata aumentando di pari passo. Lo scopo della maggior parte di questi studi, essendo l'area ancora oggetto di ricerca di livello fondamentale, si concentra sulla classificazione di segmenti di parlato preregistrati ed annotati da giudici umani. Tuttavia, accanto alla ricerca di base relativa all'individuazione delle caratteristiche acustiche del parlato che meglio descrivono l'intento emotivo del parlante, si è recentemente avviato un processo di inclusione di moduli atti a valutare il contenuto emotivo di una frase in sistemi automatici allo scopo di aiutarne le decisioni. Un esempio di applicazioni del genere si è visto nel campo dei call center intelligenti (Herm et al., 2008), principalmente mirati a rilevare la rabbia (anger detection) negli utenti per dirottare la chiamata ad un operatore umano nel caso in cui il sistema non fosse in grado di fornire adeguata assistenza . Applicazioni dell'affective computing più in generale hanno riguardato inoltre l'interazione uomo-robot (Brazin, 2002; Arkin et al., 2003). In questo lavoro, presentiamo un'applicazione della rilevazione di emozioni dal parlato in tempo reale per produrre reazioni da parte di una piattaforma robotica.

La piattaforma in questione è il robot Pleo, un prodotto commerciale controllabile sia attraverso l'uso di un linguaggio di programmazione specifico che attraverso un collegamento con un computer remoto. A causa della complessità della procedura di analisi del segnale di seguito descritta, questa seconda opzione è quella implementata dal sistema presentato. Pleo ha la forma di un cucciolo di dinosauro, e questo aiuta l'interazione in quanto non fornisce un termine di paragone con un essere animato del quale si sia potuta fare esperienza. In Figura 1 viene mostrato il robot Pleo.

Figura 1: Il robot Pleo

Il metodo di estrazione di features e di produzione di stimoli emotivi, implementato in Simulink, è basato sul concetto di sillaba fonetica (talvolta indicata come pseudosillaba) seguendo la terminologia usata in D'Alessandro (1995), che definiva questa unità come *"*[...] *a continuous voiced segment of speech organized around one local loudness peak, and possibly preceeded and/or followed by voiceless segments"*. La definizione che tuttavia si adatta meglio ai segmenti che

vengono effettivamente individuati dal sistema è quella riportata in Roach (2000) che descrive la sillaba fonetica come *"*[...] *consisting of a centre which has little or no obstruction to airflow and which sounds comparatively loud; before and after that centre* [...] *there will be greater obstruction to airflow and/or less loud sound"*.

Il sistema sovrappone il concetto linguistico di sillaba fonetica a quello, tecnologico, di buffer a dimensione variabile. Tale buffer viene riempito da frames catturati in tempo reale finché non viene riconosciuto il template della sillaba fonetica. Una volta riconosciuto tale template, l'intero buffer viene passato al modulo di analisi che, in base alle informazioni acustiche, produce uno stimolo emotivo. La sillaba fonetica è quindi l'unità di analisi fondamentale del sistema di controllo del robot.

Per quanto riguarda la rappresentazione interna dello stato emotivo, il robot implementa un modello tridimensionale che prevede i tre assi relativi a Valenza, Attivazione e Dominanza (Grimm & Kroschel, 2005). L'architettura generale del sistema prevede che ogni modulo di registrazione di eventi da sensori diversi possa iscriversi ad una interfaccia emotiva inviando stimoli positivi o negativi lungo gli assi definiti. La composizione di questi stimoli produce il nuovo stato emotivo all'interno del robot, guidandone le azioni. Questa scelta di progettazione ha lo scopo di rendere modulare ed estendibile il sistema. Come esempio, in Figura 2 viene presentato uno schema dell'architettura che prevede l'uso di microfoni e sensori tattili.

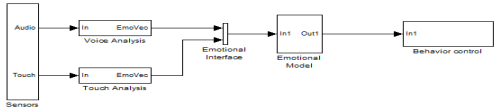


Figura 2.L'architettura del sistema

La prima sperimentazione del sistema realizzato, presentata in questo lavoro, consiste nel causare reazioni nel robot Pleo utilizzando il livello di energia trovato nei nuclei delle sillabe fonetiche. A livelli di energia al di sopra di una soglia del silenzio stabilita empiricamente, toni di voce bassi provocano un abbassamento del livello di attivazione fino a portare il robot ad "addormentarsi". Livelli di energia alti vengono, al contrario, interpretati come stimoli eccitatori e spingono il robot ad esibire il comportamento "giocare".
I dati raccolti forniscono indicazioni incoraggianti per quanto riguarda la possibilità di utilizzare i parametri acustici fondamentali della voce umana per ottenere un'interazione percepita il più possibile come "naturale" e la possibilità di realizzare architetture robotiche basate su una teoria linguistica. I comportamenti richiesti al robot appaiono emergere con facilità nonostrante il sistema di controllo utilizzato nei test tenga unicamente conto, di fattori legati alla sola energia, rendendo l'intelligenza artificiale sufficientemente credibile da parte degli utenti. Questo, comunque, assumendo che il tipo di intelligenza artificiale che si intende simulare è di tipo primitivo/animale sia per via dell'uso esclusivo delle emozioni sintetiche all'interno del sistema di controllo che dell'aspetto del robot.

Bibliografia

Breazeal, C., 2002. Designing sociable robots. MIT Press.

Arkin, R. C., Fujita, M., Takagi, T., R. Hasegawa, An ethological and emotional basis for human-robot interaction, in Robotics and Autonomous Systems, 2003, pp. 191--201.

D'Alessandro, C., Mertens, P., 1995. Automatic pitch contour stylization using a model of tonal perception. Computer Speech and Language 9 (3), pp. 257--288.

Grimm, M., Kroschel, K., 2005. Emotion estimation in speech using a 3D emotion space concept. In Proc. of IEEE Automatic Speech Recognition & Understanding Workshop, pp. 381--385.

Herm, O., Schmitt, A., Liscombe, J., 2008. When calls go wrong: How to detect problematic calls based on log files and emotions. In Proc. of Interspeech. pp. 463--466

Roach, P., 2000. English Phonetics and Phonology. A Practical Course. CUP.

Esperimenti di identificazione della lingua parlata in ambito giornalistico

Diego Giuliani, Roberto Gretter

Nell'ambito del riconoscimento automatico della voce in ambito giornalistico, usualmente si assume di conoscere la lingua in cui un dato canale, ad esempio televisivo, trasmette i suoi telegiornali. In effetti questa assunzione viene spesso disattesa in canali internazionali (l'Italia è un'eccezione da questo punto di vista), dove gran parte delle interviste a persone straniere iniziano con alcuni secondi dell'audio originale, che poi cala di volume quando interviene la traduzione nella lingua di riferimento. In alcuni casi, le interviste in lingua straniera vengono trasmesse direttamente con l'audio originale e vengono aggiunti dei sottotitoli per consentirne la comprensione. Applicare un riconoscitore automatico nella sola lingua di riferimento al flusso audio provoca quindi, inevitabilmente, lo sgradevolissimo effetto di introdurre una sequenza di errori ogniqualvolta compare del parlato in una lingua diversa.

Sorge quindi la necessità di far precedere il processo di riconoscimento automatico da un modulo di identificazione del linguaggio, capace di elaborare il flusso audio, dividerlo in segmenti ed associare ad ogni segmento di parlato la lingua identificata. Segmenti non appartenenti al linguaggio di riferimento possono quindi essere ignorati oppure essere elaborati da un riconoscitore appropriato.

Negli ultimi anni sta emergendo la tendenza ad acquisire risorse linguistiche a basso costo, come ad esempio dati testuali raccolti via web per costruire modelli del linguaggio aggiornati e capaci di seguire giorno per giorno l'evolversi delle varie lingue. Come dati audio sono accessibili diverse fonti: web, canali radio o televisivi. L'audio raccolto tramite alcuni di questi canali può essere utilizzato per addestrare modelli acustici in una nuova lingua con procedure completamente non supervisionate. Viene ad esempio effettuato un primo riconoscimento con modelli acustici derivati da altre lingue, e dall'allineamento risultante è possibile addestrare dei modelli acustici imperfetti che, per passi successivi, possono essere raffinati fino ad ottenere prestazioni ragionevoli.

Utilizzando questa procedura, negli anni scorsi abbiamo costruito dei riconoscitori in diverse lingue, ottenendo come sottoprodotto del materiale audio etichettato in maniera non supervisionata, con un'accuratezza di parola (Word Accuracy) che, a seconda della lingua, varia tra 70% e 90%. Tale materiale è stato utilizzato per creare dei corpora in diverse lingue, omogenei per tipologia di contenuto e dimensione, poi utilizzati per addestrare diversi sistemi di identificazione del linguaggio (Language IDentification, LID).

Dai test set predisposti negli scorsi anni per valutare le prestazioni dei riconoscitori vocali, ottenuti trascrivendo manualmente alcune ore di materiale della stessa tipologia in diverse lingue, abbiamo estratto del materiale utilizzabile per valutare le prestazioni di sistemi di LID.

In questo lavoro considereremo 6 lingue: italiano, turco, spagnolo, francese, tedesco, russo. Esperimenti preliminari considerano data set abbastanza contenuti, infatti come materiale di addestramento abbiamo utilizzato, per ognuna di queste lingue:

- 3 ore di materiale audio trascritto in maniera non supervisionata;
- testi raccolti da web pari a 10 milioni di parole;
- un lessico composto dalle 5000 parole più frequenti, trascritto foneticamente.

Per valutare le prestazioni dei vari sistemi implementati abbiamo definito 3 insiemi di test, costituiti da segmenti audio. Ogni segmento audio contiene parlato in una sola tra le 6 lingue considerate, ed è caratterizzato da una durata prestabilita (ad esempio, tra 3 e 7 secondi).

Le caratteristiche dei 3 insiemi di test sono:

identificativo	numero di segmenti	durata minima	durata massima
TF1_TT5	541	1 secondo	5 secondi
TF3_TT7	522	3 secondi	7 secondi
TF5_TT9	481	5 secondi	9 secondi

L'approccio più noto per l'identificazione del linguaggio parlato è quello basato sull'utilizzo di una mistura di Gaussiane (nota come "Gaussian Mixture Model", GMM) per modellare le proprietà acustiche di una data lingua. In questo caso, per ogni lingua da riconoscere uno specifico GMM viene addestrato separatamente sui dati disponibili per quella lingua. Questo approccio non richiede la trascrizione od annotazione dei dati acustici ma la sola conoscenza che i dati acustici siano istanze di una certa lingua. In fase di test, dato un segmento di parlato, i valori di verosimiglianza che rappresentano la plausibilità che il segmento di parlato sia stato generato da ciascuno dei GMM sono confrontati tra loro e vince il modello corrispondente al valore di massima verosimiglianza.

Un secondo approccio implementato fa uso di un vero e proprio sistema di riconoscimento (nel seguito verrà chiamato ASR), che considera anche informazioni lessicali e linguistiche. Sui dati audio disponibili viene addestrato un insieme di modelli acustici multilingua (ad esempio tutte le lingue considerate condividono il modello acustico del fonema /a/), mentre i dati linguistici vengono utilizzati per addestrare un modello del linguaggio statistico anch'esso multilingua. In questa fase ogni parola viene preceduta da un'etichetta che individua la lingua di origine (ad es. it:città de:nicht fr:attaque). Un riconoscitore che utilizzi tali modelli potrà emettere una sequenza di parole in lingue diverse. Negli esperimenti effettuati, dovendo emettere un'unica etichetta per ogni segmento dato, a valle del riconoscitore viene applicato un filtro a maggioranza.

Risultati preliminari in termini di accuratezza, ottenuti con le due tecniche, sono riportati nella tabella che segue:

identificativo	GMM	ASR
TF1_TT5	81.15%	90.57%
TF3_TT7	83.14%	94.06%
TF5_TT9	79.63%	92.52%

Nel lavoro completo verranno riportati risultati ottenuti con data set maggiori e con alcune variazioni dei metodi qui accennati.

Enhancing Emotion Recognition through Improved Frame-Level Features

Imen Trabelsi[1,2] Dorra Ben Ayed[1,2] Noureddine Ellouze[2]

[1] Institute of Computer Science of Tunis (ISI), Tunis, Tunisia
[2]National School of Engineer of Tunis (ENIT), Tunis, Tunisia
trabelsi.imen1@gmail.com, BenAyedDorra@gmail.com, N.ellouze@enit.rnu.tn

Abstract

The purpose of speech emotion recognition system is to classify speaker's utterances into different emotional states such as disgust, boredom, sadness, neutral, and happiness.
Speech features that are commonly used in speech emotion recognition (SER) rely on global utterance level prosodic features. In our work, we evaluate the impact of frame-level features extraction.
The speech samples are from Berlin emotional database and the features extracted from these utterances are energy, different variant of mel frequency cepstrum coefficients (MFCC), velocity and acceleration features. The idea is to explore the successful approach in the literature of speaker recognition, GMM-UBM, to handle with emotion identification task.

Index Terms: speech emotion recognition, MFCC, Energy, GMM Supervector.

1. Introduction

Speech emotion recognition (SER) is an extremely challenging task in the domains of human-robotics interfaces and affective computing and has various applications in call centers [1], intelligent tutoring systems [2], spoken language research [3].
and other research areas.

The primary channels for robots to recognize human's emotion include facial expressions, gesture and body posture. Among these indicators, the speech is considered as a rapid transfer of complex information. This signal provides a strong interface for communication with computers. There have been plenty of studies on speech emotion recognition.

Many kind of acoustic features have been explored to build the emotion model [4]. Various classification methods have been verified for emotional pattern classification such as hidden markov models [5], gaussian mixture, artificial neural network [6] and support vector machines [7].

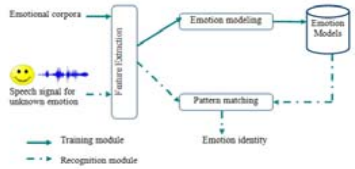
In our paper, we investigate the relationship between generative method based GMM and discriminative method based SVM [8].

The rest of paper is organized as follow: First, the description of the proposed speech emotion recognition system (SER). Second the experimental results of the system. Conclusion is drawn in the final section.

2. Emotion Recognition System

The proposed speech emotion recognition system contains three main modules (see figure 1) namely (1) extraction of feature, (2) learning the models using machine learning techniques and (3) evaluation of models.

First, suitable data sets for training and testing is collected. Second, relevant features are extracted. Third, the extracted features are modelled. Fourth, a set of machine learning techniques could be used to learn the training models. Finally, testing unknown emotional samples are used to evaluate the performances of models.



2.1. Feature extraction

The first problem that occurs when trying to build a recognition framework is the discrimination of the features to be used. Common acoustic features used to build the emotion model include pitch, intensity, voice quality features and formants [9]. Others included cepstral analysis [4].

In this paper, our feature extractor is based on: Mel Frequency Cepstral Coefficients (MFCCs), MFCC-low, energy, velocity and acceleration coefficients.

- MFCCs have been the most popular low-level features, they demonstrate good performance in speech and speaker recognition. We use the advantage of this representation for our emotion identification task.
- MFCC-Low is a variant of MFCC. Mel filter banks are placed in [20-300] Hz. Our reason for introducing MFCC-low was to represent pitch variation.
- Energy is an important prosodic feature of speech. It is, often referred to as the volume or intensity of the speech, is also known to contain valuable information [10]. Studies have shown that short term energy has been one of the most important features which provide information that can be used to differentiate different sets of emotions.
- Velocity (delta) and acceleration (delta-delta) parameters have been shown to play an important role in capturing the temporal characteristics between the different frames that can contribute to a better discrimination [11]. The time derivative is

approximated by differencing between frames after and before the current. It has become common to combine both dynamic features and static features.

2.2. Emotion modeling

GMMs have been successfully employed in emotion recognition [12]. The probability density function of the feature space for each emotion is modeled with a weighted mixture of simple Gaussian components.

This module is assured by the construction of a universal background model (UBM), which is trained over all emotional classes. There are a number of different parameters involved in the UBM training process, which are the mean vector, covariance matrix and the weight. These parameters are estimated using the iterative expectation-maximization (EM) algorithm [13].

Each emotion is then modeled separately by adapting only the mean vectors of UBM using Maximum A Posteriori (MAP) criterion [14], while the weights and covariance matrix were set to the corresponding parameters of the UBM.

To use a whole utterance as a feature vector, we transform the acoustic vector sequence to a single vector of fixed dimension. This vector is called supervector.

2.3. SVM Classification Algorithm

The support vector machines [15] are supervised learning machines that find the maximum margin hyperplane separating two classes of data.

SVMs solve non-linear problems by projecting the input features vectors into a higher dimensional space by means of a Mercer kernel.

This powerful tool is explored for discriminating the emotions using GMM mean supervectors. The reason for choosing the SVM classifier for this task is that, it will provide better discrimination even with a high dimension feature space.

In our research, we give each training supervector sample with the corresponding emotion class label. After that, we input them to the SVM classifier and gain a SVM emotionnal model. The output of the each model is given to decision logic. The model having the best score determines the emotion statue. The output of the matching step is a posteriori probability.

Our experiments are implemented using the LibSvm [16] with a linear inner-product kernel function using the one-against-one strategy for multi-class classification.

The whole speech emotion recognition is chown in figure 2.

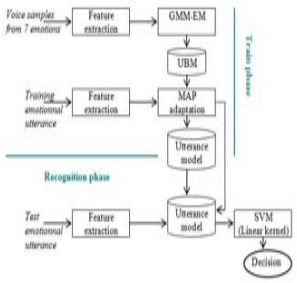


Figure 2 : *UBM-SVM based speech emotion recognition.*

3. Experiments and results

3.1. Emotional speech database

The database used in this paper is the Berlin database of emotional speech (EMO-DB) which is recorded by speech workgroup leaded in the anechoic chamber of the Technical University in Berlin. It is a simulated open source speech database.

This database contains about 500 speech samples proven from ten professional native German actors (5 actors and 5 actresses), to simulate 7 different emotions. The length of the speech samples varies from 2 seconds to 8 seconds.

Table 1 summarizes the different emotions.

Table1. *Number of utterances belonging to each Emotion Category*

Emotion	Label	Number
Anger	A	128
Boredom	B	81
Disgust	D	44
Fear	F	69
Happiness	H	71
Sadness	S	45
Neutral	N	62

3.2. System Description

The data were recorded at a sample rate of 16 KHZ and a resolution of 16 bits.

First, the signal is segmented into speech and silence. Then, silence segments are thrown away and the speech segments are pre-emphasized with a coefficient 0.95. From pre-emphasized speech, each feature vector was extracted from at 8ms shift using a 16 ms analysis window. A Hamming window is applied to each signal frame to reduce signal discontinuity.

Our baseline system consists of a 128 component GMM-UBM built using acoustic data of different emotional sentences. Individual speaker models are MAP-adapted; only mean vectors, with a relevance factor of 16.

3.3. Results and discussion

Table 2 presents the results conducted on different variants of MFCC in order to extract the most reliable feature.

Table2. *Recognition rate from different variant of MFCC*

Data	Range of filter banks	Recognition rate(%)
MFCC	300-3400	72.85
Low-MFCC	0-300	62
Combined MFCC	0-3400	81.42

Combination of MFCC and MFCC-low led to an accuracy of 81.42%.

MFCC-low features performed well in comparison with the small scale of filter banks used, it may be due to its ability to capture voice source quality variations.

For the rest of the paper, we choose the combined MFCC.

The table below (table 3) shows the full feature set used.

Table3. *Different speech feature vectors*

Data	Features	Size of features
Data1	Combined MFCC	12
Data2	Combined MFCC+Log Energy	13
Data3	Combined MFCC+ Δ+ Δ Δ	36
Data4	Combined MFCC+Log Energy+ Δ(MFCC) + Δ Δ(MFCC)	37
Data5	Combined MFCC+ Δ(MFCC) + Δ Δ(MFCC) + Δ(log energy)+ Δ Δ(log energy)	39

Table 4 presents the results from series of recognition experiments to determine the effect of different frame-level features performance.

Table 4. Recognition rate by using different features

Data Feature	Recognition rate(%)
Data1	81,42
Data2	**87,14**
Data3	84,28
Data4	81,42
Data5	71,42

Table 3 shows that the recognition rate is varied between (71.42%) and (87.4%).

This result demonstrates that the two kinds of data (Data2 and Data3) are important for emotion recognition.

The best performance comes from Data3 when MFCC are combined with log energy. The lowest recognition rate (71.42%) comes from Data5 when MFCC are combined with log energy and dynamics parameters.

We can conclude from these results is that we can get an accuracy of 81.42% with only 12 features comparing with an accuracy of 71.42% with the total 39 features.

Table 4 shows emotion recognition accuracy by analysis over all emotions. Averaged class recognition accuracy is given.

Table 5. *Recognition of emotion by category*

	Recognized emotion(%)						
Feature	A	B	D	F	H	N	S
Data1	90	90	80	90	70	50	100
Data2	80	90	90	90	80	80	100
Data3	100	100	100	90	70	60	70
Data4	100	90	100	70	70	40	100
Data5	100	80	90	70	60	20	70
RR(%)	85	87,5	86,6	82	70	50	88

We can observe that sadness has the highest recognition rate (88%) and boredom follows it with (87, 5%). The lowest rate was for the neutral synthesized speech at 50%, this cloud be explained by the fact that neutral speech doesn't contain specific emotional information.

Negative emotion (sadness, boredom, disgust) got the highest classification rate, this could be attributed to the exaggerated expression of emotion by the actors.

We also conclude that GMM SVM achieves higher recognition rate even when the training data size is small (45 utterances for sadness).

Table 5 shows the confusion matrix table that is achieved by the optimal experiment using Data3.

Table 5. *Confusion matrix of Data3*

	Recognized emotion category (%)						
	A	B	D	F	H	N	S
A	**80**	0	0	10	10	0	0
B	0	**90**	0	0	0	10	0
D	0	0	**90**	0	0	10	0
F	10	0	0	**90**	0	0	0
H	20	0	0	0	**80**	0	0
N	0	20	0	0	0	**80**	0
S	0	0	0	0	0	0	**100**

From these results, we can see that fear, happiness and neutral are the most frequently confused emotions.

It can be also found that sadness is easily classified. This matrix reveals that there are similarities between different categories of emotions that must be studied in further work.

4. Conclusion

Emotional speech recognition is gaining interest due to the widespread applications into various fields.
In our work, this task has been evaluated using spectral features, modeled by GMM-SVM on the frame level and tested on EMO-DB.

Results show that MFCC, with filter banks placed in [0-3400] combined with energy extracted at the frame level outperform the other features. The recognition rate is equal to 87.14%.

Automated recognizing emotion with high recognition accuracy still remains a challenge due to the lack of a full understanding of emotion in human minds. The problem is extremely complicated and thus, the researchers usually deal with acted emotions, just like in our paper. However, in real situations, different individuals show their emotions in a diverse degree and manner.

In Our future work, we will try to study the performance of the proposed system in a spontaneous emotional database. We will explore the possibilities of integrating other modalities such as manual gestures and facial expression and combine with the result of some other machine learning methods such as KNN, HMM or Random Forest.

5. References

[1] [1] DONN, M., RUILI, W. and LIYANAGE, C., (2007). Ensemble methods for spoken emotion recognition in call-centres. *Speech communication*, vol. 49.

[2] XIAO. L., YADEGAR J., AND KAMAT, N.,(2011). A Robust Multi-Modal Emotion Recognition Framework for Intelligent Tutorig Systems. In IEEE International Conference on Advanced Learning Technologies (ICALT).

[3] Forbes, K. and Litman, D., (2005). Using bigrams to identify relationships between student certainness states and tutor responses in a spoken dialogue corpus. In Proc. Of 6th SIGdial Workshop on Discourse and Dialogue, Lisbon, Portugal.

[4] VERVERIDIS, D. and KOTROPOULOS, C., (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, vol. 48, no. 9, pp. 1162–1181.

[5] NEW, T., FOO, S. W., and SILVA, L. D., (2003). Speech emotion recognition using hidden Markov models. *Speech Communication*, vol. 41, pp. 603–623.

[6] PAO, T.-L, CHEN, Y.-T and YEH, J.-H., (2006). Mandarin emotional speech recognition based on SVM and NN. In Proc of the 18th International Conference on Pattern Recognition (ICPR'06), vol. 1, pp. 1096-1100.

[7] LIN, Y. and WEI, G., (2005). Speech emotion recognition based on HMM and SVM. In Proc. of 2005 International Conference on Machine Learning and Cybernetics, vol. 8, pp. 4898-4901.

[8] TRABELSI, I., BENAYED, D., (2011). Evaluation d'une approche hybride GMM-SVM pour l'identification de locuteurs. *La revue e-STA*, 8(1), 61-65. *http://www.see.asso.fr/esta/?page=8&id_document=458&id_article=231.*

[9] FERNANDEZ, R. and PICARD, R.W., (2005). Classical and Novel Discriminant Features for Affect Recognition from Speech. In Proc. Of InterSpeech , pp. 1-4, Lisbon, Portugal.

[10] WRIGLEY, S. N., BROWN,G. J. WAN, V., , and Renals. S., (2005). Speech and crosstalk detection in multichannel audio. In IEEE Transactions on Speech and Audio Processing.

[11] Bouvrie, J., Ezzat, T. and Poggio, T., (2008). Localized Spectro-Temporal Cepstral Analysis of Speech. In Proc. ICASSP 2008, pp. 4733-4736.

[12] Vlasenko, B., Schuller, B., Wendemuth A. and Rigoll G., (2007). Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing. In Proc. of Affective Computing and Intelligent Interaction, pages 139–147, Lisbon, Portugal.

[13] Dempster, A. P., Laid, N. M. and Durbin, D., (1977). Maximum Likelihood from incomplete data via the EM algorithm. J. Royal Statistical Soc, vol. 39, pp. 1-38.

[14] Reynolds, D., Quatieri, T. and R. Dunn., (2000). Speaker verification using adapted gaussian mixture models. DSP, Vol. 10, No. 3, pp. 19–41.

[15] Vapnik, V., (2005). The nature of statistical learning theory. Spring-verlag, New York.

[16] Chang, C. C. and Lin, C. J., (2001). LIBSVM : a library for support vector machines. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Svjetllana TITINI, Ph.D
Universiteti "Aleksandër Moisiu", Durrës, ALBANIA

I problemi specifici della traduzione della terminologia giuridica nella lingua albanese

La traduzione della terminologia del diritto incontra varie difficoltà che incidono sulla comprensione e sulla chiarezza del testo giuridico nella lingua d'arrivo.

In questo articolo verranno evidenziati alcuni dei principali problemi relativi alla traduzione della terminologia del diritto e verranno messi in rilievo gli errori che sorgono durante questo processo facendo riferimento agli esempi nella lingua albanese.

I problemi che riguardano la traduzione della terminologia giuridica verranno analizzati suddividendoli in due gruppi.

In primo luogo verranno analizzati i problemi che nascono dalla specificità del diritto nazionale. L'esistenza dei vari ordinamenti giuridici diventa anche una delle cause fondamentali da cui derivano la divergenza dei concetti e degli istituti e in alcuni casi la loro assenza nei differenti ordinamenti. Ogni stato possiede un proprio sistema giuridico e di conseguenza una propria terminologia giuridica, ciò significa che numerosi referenti (concetti e istituti) non hanno nella lingua d'arrivo i corrispettivi termini. Da questo fatto derivano anche degli errori di traduzione che si presentano nella pratica giuridica e nella comunicazione orale in generale (in quanto i termini giuridici fanno parte della lingua standard) che consistono nella formazione di neologismi sbagliati o nell'uso di termini che esprimono solo parzialmente il significato dei termini nella lingua di partenza.

In secondo luogo verranno analizzati i problemi nascenti dalla lingua con riguardo al rapporto tra parola e concetto che cambia a seconda della lingua. In questo caso sarà evidenziato tramite gli esempi come alcuni termini abbiano connotazioni positive o negative che hanno inciso anche sulle scelte terminologiche in diversi casi e in diversi periodi. Inoltre sarà evidenziato tramite gli esempi come la molteplicità dei significati del termine giuridico incide sulla possibilità di traduzione ma nello stesso tempo sulla trasparenza linguistica.

Alla fine di questo lavoro saranno presentate le conclusioni attinenti alle problematiche evidenziate durante il procedimento traduttivo e alle possibili soluzioni per evitare il rischio di errori di traduzione al quale viene esposta la terminologia del diritto.

A CROSS-LINGUISTIC ANALYSIS OF JAZZ VOCAL IMPROVISATION

Alessandra De Martino

(Language Understanding and Speech Interfaces Lab, Dept. Of Physics, University of Naples "Federico II")

alessandr.demartino@studenti.unina.it

The relationship between language and music has always captured the interest of linguists. On one hand, we can find linguists who have demonstrated the deep connection between language and music composition, and, in particular to what extent the mother tongue influences the composition of classical music (Patel and Daniele, 2003; Daniele and Patel, 2004), on the basis of rhythmic measures (Grabe & Low, 2002). These results have brought to the heated debate on the typological distinction of the languages in syllable-timed vs. stress-timed. The same rhythmic measures have been applied to the jazz area (McDonough et al., 2007), showing the difficulties in establishing a clear typological distinction.

The goal of this study is to examine the vocal improvisation skills of jazz singers or musicians, in particular the kind of jazz improvisation represented by the *scat*, taken as an uncommon language. A sample of *scat* made by Anglophones has been compared to the *scat* produced by Italian-speakers, using a system (PRAAT) for cataloguing every phonetic component of each fragment extracted from the pieces: bars, accents, syllables, consonants, vowels and tones. The results show a statistically relevant difference in the consonant structures and vowels usage between the two groups of samples.

This analysis has been developed on a structural level, starting from the didactic approaches to *scat* (Aebersold, 1967; Stoloff, 1996; Madura, 1996; 2008a; 2008b), taking into account the studies on the cognition mechanisms at the origin of an improvisational output (Bauer 2007; Shaw 2008). This work represents an innovative turning point in this area, due to the fact that there are no computational approaches comparing improvisational outputs, considered as spontaneous productions, by subjects of different place of origin.

Samples of *scat* made by Italian-speakers and American-speakers have been collected. The corpora are composed of six American tracks and five Italian ones, equalized on the time level, and contain approximately one thousand of syllables. I defined the components I wanted to examine by producing an annotation on seven tiers that isolated:

- accents
- bars
- phonetic transcription of syllables (Albano Leoni & Maturi, 2002)
- structures
- type of consonants
- frequencies
- generic types of vowels.

This preliminary study has taken into account only part of the annotations listed, due to the structural nature of the analysis. The relevant data, in this case, are: bars, structures, type of consonants and type of vowels.

The selected Italian subjects are all males, aged between thirty and fifty years old, all with theoretical bases in music and specialized in jazz, at different levels, from intermediate to advanced, discarding the beginners. Following the same criteria, the American subjects were selected in a range of age between twenty and fifty years old, at the two different levels of achievement: intermediate and advanced.

The basic data collected are: simple core, closed, complex syllables and centralized vowels. Consequently, comprehensive percentages could be extracted and tests for the statistical relevance of the differences could be performed. From the comparison of the two groups, concerning the percentages of simple core syllables and complex syllables, it is clear that the behavior in the two languages is similar ($p>0.05$).

Different considerations must be done for closed syllables, where the percentage shows the different syllable's construction in the two languages under analysis, while the statistical test indicates that the difference is relevant. American subjects show a higher usage of closed syllables in comparison with Italian ones, and it can be justified by the mother tongue, that surely influences the production of improvisation, almost like the production in a second language. Even more relevant is the centralized vowel results, due to the different codification of the vowels. In detail, I simplified the classification of the vowels, distinguishing them in 'I' and 'D', that stand for centralized (in Italian 'indefinite') and peripheral (in Italian 'definite'), due to their degree of articulation from the target features.

Results demonstrate the great influence of native language on music production, that, on one hand, depends on the level of musical acquisition and experience and, on the other, manifests the features present in a determined spoken language, at level of cognition. This hypothesis has been supported by the values of closed syllables usage (14,44% for American subjects and 5,52% for Italian ones, with a p-value of 0,033), and centralized vowels usage (18,66% for American subjects and 5,35% for Italian ones, with a p-value of 0,005).

This first study establishes the first step of a research project aimed at understanding the phonetic processes guiding musical improvisation in a second language, and to what extent our imitative capability imposes its influence. Further analysis on this subject will study the data on pitch level too, recording samples on the same musical piece, in order to analyze the kind of notes sung on determined chords, and to specialize more precisely the suitable subjects.

References

Aebersold, J. 1967: *Jazz: How to play and improvise, Volume 1*. New Albany, IN. Jamey Aebersold Jazz.

Albano Leoni, F. A., & Maturi, P. 2002: *Manuale di fonetica*. Roma, Italy. Carocci.

Bauer, W. 2007: *Louis Armstrong's "Skid Dat De Dat": Timbral organization in an early scat solo*. Jazz Perspectives, 1(2). pp. 133–165.

Daniele, J. R., & Patel, A. D. 2004*: The interplay of linguistic and historical influences on musical rhythm in different cultures*. 8th International Conference on Music Perception and Cognition, August 3–7, Northwestern University, Evanston, IL. pp. 759–762.

Grabe, E. and Low, E. L. 2002: *Durational variability in speech and the rhythm class hypothesis*. Laboratory Phonology *7*. Edited by C. Gussenhoven and N. Warner Mouton de Gruyter. Berlin. pp. 515–546.

Madura Ward-Steinman, P. 2008: *Vocal improvisation and creative thinking by Australian and American university jazz singers: A factor analytic study*. Journal of Research in Music Education*, 56(1).* pp. 5-17.

Madura Ward-Steinman, P. 2008: *Vocal improvisation and creative thinking by Australian and American university jazz singers: Case studies of outliers' musical influence.* Journal of Research in Music Education*, 177*. pp. 29-43.

Madura, P. D. 1996: *Relationships among vocal jazz improvisation achievement, jazz theory knowledge, imitative ability, musical experience, creativity and gender.* Journal of Research in Music Education, 44 (3). pp. 252–67.

McDonough, J., H. Danko, and J. Zenz, 2007: *Rhythmic structure of music and language: An empirical investigation of the speech cadence of American jazz masters Louis Armstrong and Jelly Roll Morton*. In L.Wolter and J. Thorson (Eds.), University of Rochester. Working Papers in the Language Sciences, 3(I). pp. 45-56.

Patel, A. D. and Daniele, J. R. 2003: *An empirical comparison of rhythm in language and music*. Cognition. 87. pp. B35–B45.

Shaw, P. A. 2008: *Scat syllables and markedness theory*. Working Papers in Linguistics, Toronto, 27. pp. 145–191.

Stoloff, B. 1996: *Scat!: Vocal Improvisation Techniques.* Brooklyn, N.Y. : Gerard & Sarzin Publishing Company.

Sviluppo di un sistema *embedded* di distant speech recognition

Alessandro Sosi, Fabio Brugnara, Marco Matassoni,
Maurizio Omologo, Mirco Ravanelli

Fondazione Bruno Kessler, Trento

1 – Distant speech recognition

La voce rappresenta il sistema di comunicazione tra esseri umani più semplice e rapido. Il parlato può anche essere considerato una delle interfacce uomo-macchina più intuitive. Il mondo della ricerca scientifica e dell'industria ne è consapevole e investe in questa tecnologia ormai da diversi anni. Numerosi sono i dispositivi che hanno tentato di conquistare il mercato. Molti sistemi hanno fallito la propria missione a causa della poca robustezza al rumore e dell'inadeguatezza all'interazione a distanza. La *distant speech recognition* rappresenta, infatti, una delle sfide più ambiziose nell'ambito delle interfacce uomo-macchina intelligenti, soprattutto perché, raggiunto un buon livello di funzionamento, apre la strada a tutta una serie di nuove applicazioni precluse da un semplice sistema *close-talk.* Gran parte delle implementazioni immaginabili deve poter funzionare grazie ad un sistema di *distant speech recognition* completo, autonomo e facilmente integrabile. Sistemi di riconoscimento vocale basati su *cloud* sono efficaci solamente su dispositivi che possono vantare una connessione alla rete costante e affidabile. E' facile intuire come il *cloud* possa, in molti casi, essere un limite. Risulta dunque preferibile, per task di medio-bassa complessità, compiere le operazioni di *speech processing* in locale. Il mondo dei sistemi *embedded* rappresenta l'anello di congiunzione tra la richiesta di buona capacità computazionale e il sogno dell'integrazione all'interno di oggetti e scenari quotidiani. In questo senso il futuro del riconoscimento vocale sarà sempre più dipendente da piattaforme di piccole dimensioni, basso consumo, basso costo ma, allo stesso tempo, capacità di calcolo tali di riuscire a gestire un sistema completo di *distant speech recognition*.

2 – Hardware prototipo

Il prototipo sviluppato è un sistema completo di *distant speech recognition*: si parte dall'acquisizione sonora fino ad arrivare al comando da fornire ad un eventuale dispositivo di attuazione. Il sistema è attualmente formato da otto minuscoli microfoni MEMS e da un piccolo computer. Le dimensioni dei microfoni sono nell'ordine dei millimetri, mentre il computer, una piattaforma *embedded*, è costituito da un'unica singola scheda dotata di un processore ARM più piccolo di una moneta da un centesimo di euro.

I microfoni utilizzati sono microfoni digitali MEMS (*Micro Electro-Mechanical System*) prodotti da ST-Microelectronics. Questi microfoni hanno il vantaggio di avere dimensioni ridottissime, tali da permettere di immaginare un

facile inserimento degli stessi all'interno di qualsiasi oggetto. Nonostante le piccole dimensioni, i microfoni MEMS sono dotati di buone caratteristiche tecniche ed hanno un costo inferiore rispetto ad altre soluzioni derivanti dal settore audio professionale. I vantaggi non finiscono qui: questi microfoni sono dotati di una scheda dedicata che permette di interfacciarli tramite una semplice connessione USB.

Il piccolo computer utilizzato è un *Singol Board Computer*, composto da un'unica scheda di dimensioni 8.5x8.5cm. Questo computer è prodotto da Calao System ed è noto con il nome di "SnowBall". Il cuore di questa scheda è un processore ARM *dual-core* in grado di assicurare buone prestazioni con un basso consumo energetico. Sulla scheda è installata "Linaro", una versione di Linux adatta ai *System on Chip*.

3 – Software prototipo

Il software attualmente presente su Snowball è organizzato in due macro-blocchi. La prima riguarda il *front-end* multi-microfonico, la seconda parte consta in un riconoscitore vocale completo, basato sugli *Hidden Markov Models*.

Il riconoscimento vocale basato su microfoni *close-talk* ha raggiunto un buon livello di maturità tecnica. Una delle maggiori sfide consiste nell'ottenere lo stesso livello di *word accuracy* e robustezza in un contesto di *distant speech recognition*. In questo secondo caso le difficoltà aggiuntive sono numerose: riverberazioni ed eco ambientali, diminuzione del rapporto segnale rumore, interferenti generici, sorgenti sonore multiple, orientamento del parlatore. Gli algoritmi di *front-end* multi-microfonico implementano tecniche atte a far fronte a queste complicazioni. Grazie agli otto microfoni e a tecniche di *beamforming,* è possibile incrementare il rapporto segnale rumore e aumentare la *direct to reverberant ratio.* La *voice activity detection* consente, successivamente, di selezionare le parti di segnale contenenti attività vocale.

Il segnale proveniente dal blocco di *front-end* multi-microfonico passa in seguito ad un riconoscitore vocale *speaker independent* che è in grado di funzionare in *real time*. Grazie alla sua flessibilità e scalabilità, il sistema proposto può essere utilizzato in configurazioni differenti, capaci di svolgere compiti differenti: l'identificazione di una parola chiave (*keyword spotting*), il riconoscimento di uno o più comandi vocali, la trascrizione del parlato.

Il sistema nel suo complesso è utilizzabile in molti scenari differenti. Tra le possibili applicazioni rivestono particolare importanza quelle nel settore della domotica, ambiente nel quale si prefigge la possibilità di utilizzare la propria voce come un telecomando per controllare un'abitazione. Proprio nel settore della domotica si stanno concentrando gli sforzi attuali. Il sistema proposto prevede, infatti, di riuscire a ricevere dei comandi vocali per svolgere delle operazioni all'interno di una casa. L'accensione e lo spegnimento delle luci e la regolazione della temperatura delle stanze sono solo alcune delle possibilità offerte. Il sistema finale, anticipando quanto sarà presentato nell'articolo finale, è stato pensato come una macchina a due stati: il primo stato è costituito da un

keyword spotter sempre in ascolto. Il secondo stato, attivato a seguito della corretta identificazione della parola chiave, è in grado di ascoltare e comprendere un comando e di trasmettere un segnale di attuazione. I comandi possono raggiungere una certa complessità ed essere in numero sufficientemente alto da coprire un ampio elenco di operazioni. Con lo scopo di rendere più robusto il sistema, è stata inserita una rete di rigetto che agisce parallelamente sia al task di *keyword spotting,* sia al task di identificazione dei comandi.

4 – Prospettive future

Numerose appaiono le prospettive future. Da un lato la creazione di nuovi prototipi (anche sfruttando diverse piattaforme ARM presenti sul mercato), dall'altra il miglioramento e perfezionamento della parte software. In particolare il *front-end* multi-microfonico può essere ulteriormente sviluppato mediante l'utilizzo di tecniche di *enhancement* adattativo.

5 – Bibliografia

[1] – Matthias Woelfer, John McDonough (2009) – **Distant Speech Recognition** – John Wiley & Sons, UK.

Controllare la casa con la voce: il progetto DIRHA

Alessio Brutti, Luca Cristoforetti, Marco Matassoni, Francesco Nesta,
Maurizio Omologo, Mirco Ravanelli, Piergiorgio Svaizer

Fondazione Bruno Kessler, Trento, Italia

matasso@fbk.eu

Lo sviluppo della domotica ha recentemente aperto nuovi scenari per le tecnologie vocali in quanto rappresenta un ideale complemento per l'introduzione dell'interazione vocale per il controllo dei dispositivi che si trovano abitualmente nelle case. In particolare per un'utenza svantaggiata (ad esempio con disabilità motorie) risulta estremamente attraente poter pilotare alcuni dispositivi dell'abitazione con la propria voce, senza dover ricorrere a palmari o altri strumenti da tenere sempre a portata di mano.

Il progetto DIRHA, recentemente avviato, mira appunto all'esplorazione e alla progettazione di sistemi basati su microfonia distribuita nell'ambiente domestico per consentire un'interazione a mani libere con il sistema che controlla le apparecchiature della casa. La collaborazione con un'azienda che si occupa di domotizzazione punta all'obiettivo di convergere su soluzioni concretamente applicabili all'utente finale.

Il progetto vuole studiare a fondo una possibile interazione vocale in cui si utilizzi un linguaggio naturale per formulare le proprie richieste al sistema domotico. La caratteristica innovativa inoltre è rappresentata dall'impiego di microfoni distribuiti nell'ambiente che rendono possibile quindi un'interazione cosiddetta *hands-free* che non vincola l'utente a parlare in una determina posizione o a utilizzare un microfono vicino alla bocca. I temi scientifici coinvolti in questa ricerca sono molteplici: l'elaborazione acustica multi-canale, il riconoscimento vocale robusto rispetto ai possibili rumori di fondo, l'elaborazione e l'interpretazione del linguaggio naturale, l'identificazione e la verifica del parlatore, la gestione del dialogo tra l'utente e il sistema. Uno specifico obiettivo del progetto è inoltre lo studio di un nuovo tipo di dispositivo di acquisizione, rappresentato da microfoni digitali MEMS (Micro Electrical-Mechanical System).

Si prevede di realizzare un sistema capace di funzionare in quattro lingue (italiano, tedesco, greco, portoghese) e di installarlo nelle case di alcuni utenti reali, disponibili all'utilizzo e alla valutazione sul campo del prototipo. I soggetti selezionati per una prima sperimentazione di questa innovativa tecnologia sono dei disabili motori che rappresentano quindi una categoria molto motivata al suo utilizzo nella vita quotidiana in casa. Questi utenti sono stati coinvolti fin dall'inizio del progetto per individuare le più importanti funzionalità del sistema domotico in sviluppo, in modo da costruire degli scenari applicativi realistici.

Gli aspetti particolarmente innovativi del progetto DIRHA sono rappresentati da una modalità di interazione sempre attiva: il sistema sarà in grado di reagire in qualsiasi momento ad una richiesta dell'utente rimanendo peraltro inattivo a fronte di un qualsiasi evento acustico o parlato non pertinente. A tale scopo è necessario studiare e sviluppare algoritmi per l'elaborazione di segnali multi-canale in grado di rilevare accuratamente e eventualmente classificare posizione e natura delle varie possibili sorgenti acustiche nelle stanze della casa.

Uno dei problemi iniziali affrontati nel progetto è stata l'acquisizione di dati acustici rappresentativi dello scenario applicativo per addestrare e validare i vari componenti del sistema previsto. Un approccio particolarmente appropriato in questo caso prevede lo sviluppo di corpora simulati con considerevoli vantaggi rispetto alla tradizionale acquisizione di dati reali. L'acquisizione di estesi corpora di dati reali è,

infatti, un'attività estremamente costosa, in quanto ogni sequenza deve essere manualmente acquisita, segmentata ed etichettata. Lo sviluppo di tecniche per la simulazione dei dati riduce notevolmente il lavoro manuale e rende possibile la generazione di enormi quantità di dati utilizzando solo limitate misurazioni reali (risposta all'impulso, rumore di sottofondo) provenienti dall'ambiente di interesse. Attraverso le simulazioni è possibile inoltre variare le condizioni sperimentali a piacere, valutando così le tecnologie in condizioni specifiche difficilmente riproducibili in ambienti reali. Lo strumento per la creazione di corpora simulati, attraverso un linguaggio analogo a xml, è in grado di sintetizzare i dati attraverso la definizione di alcuni parametri stocastici (ad esempio, tipologia, posizione, orientazione e numero delle sorgenti attive, la probabilità di sovrapposizione fra sorgenti, SNR, ecc). Il principale punto di forza di questo simulatore risiede nella possibilità di generare dati multi-microfonici preservando anche il tempo di propagazione del segnale acustico nell'ambiente. Tale caratteristica rende i corpora generati utili in vari ambiti previsti nello sviluppo del sistema DIRHA: tecniche di localizzazione, beamforming, separazione delle sorgenti, enhancement del segnale, segmentazione e classificazione, cancellazione d'eco oltre che riconoscimento vocale.

Altro aspetto particolarmente sfidante del progetto riguarda la gestione dell'interazione con gli utenti, che virtualmente possono richiedere i servizi del sistema simultaneamente in diverse stanze della casa. Da qui la necessità di progettare e implementare un gestore di dialogo concorrente in grado di istanziare più sessioni di dialogo in accordo con le possibili richieste vocali acquisite in varie posizioni. Questo compito è reso più complesso dalla scelta di lasciare l'utente libero di esprimersi liberamente e quindi di non vincolare i comandi ad una lista predefinita e fissa: quello che il riconoscitore del parlato produce, talvolta con errori, deve essere interpretato opportunamente per permettere al sistema di dialogo di eseguire l'azione corretta richiesta dal parlatore. Accanto al più classico impiego di grammatiche progettate specificatamente per supportare i diversi stati del dialogo, è prevista quindi l'esplorazione di un approccio più innovativo basato su tecniche di machine learning per generalizzare da esempi il significato da associare alle possibili richieste dell'utente.

Dai risultati del progetto si prevede di ottenere indicazioni importanti per l'applicazione di queste tecnologie in altri scenari in cui l'utente non vuole o non può essere vincolato dai microfoni e in cui la voce (o l'audio in generale) rappresenta il mezzo più efficace per controllare dei dispositivi o ottenere informazioni dall'ambiente. Possibili esempi sono l'assistenza o l'ausilio per anziani, la robotica, la sorveglianza, l'automobile.

Nell'articolo completo si descriveranno le attività condotte nella prima parte del progetto: le caratteristiche del sistema in base alle richieste ed esigenze di possibili utenti intervistati, l'architettura hardware e software ipotizzata, alcuni risultati preliminari su dati acustici raccolti o generati appositamente per il dominio considerato.

Marco A. Piccolino-Boniforti
Department of Linguistics, University of Cambridge, United Kingdom
map55@cam.ac.uk

Linking the output of a computational model of prefix recognition to looks at targets and competitors from an eye-tracking experiment

The present research contributes to the investigation of the sound-to-grammar mapping by developing a novel computational model in which complex acoustic patterns can be represented conveniently, and exploited for simulating the prediction of English prefixes by human listeners. The implemented model, which accepts recordings of real speech as input, was compared in a simulation with the qualitative results of an eye-tracking experiment. The main purpose of this comparison was to check whether a computational model of this kind is able to provide any useful insight about the behaviour of listeners in specific tasks where subtle differences in phonetic detail can signal a grammatical distinction which, as is the case for true and pseudo prefixes in British English, is noticeable both at the acoustic (Smith et al. 2012) and perceptual (Baker 2008) levels. The computational model accounted for observed perceptual differences between true and pseudo prefixes.

The model presented here is rooted in the principles of rational analysis (Anderson, 1991) and Firthian prosodic analysis (Firth, 1948), and formulated in Bayesian terms. It is based on three core theoretical assumptions: first, that the goals to be achieved and the computations to be performed in speech recognition, as well as the representation and processing mechanisms recruited, crucially depend on the task a listener is facing, and on the environment in which the task occurs. Second, that whatever the task and the environment, the human speech recognition system behaves optimally with respect to them. Third, that internal representations of acoustic patterns are distinct from the linguistic categories associated with them.

In the current model it is assumed that listeners, by analysing fine-tuned, learned auditory patterns in the proper prosodic and grammatical context, can set prefix prediction as an intermediate task in order to fulfil higher-level goals. The model is first motivated in terms of acoustic analyses and behavioural experiments. The computational aspects of the model are dealt with, in terms of goal, environment and constraints. The model is then given a formal description with the aid of a Bayesian network. Finally, those model components that are implemented in the simulation are also described in terms of processes and representations.

The representational level exploits several tools and findings from the fields of machine learning and signal processing, and interprets them in the context of human speech recognition. Because of their suitability for the modelling task at hand, two tools are dealt with in particular: the relevance vector machine (Tipping, 2001), which is capable of simulating the formation of linguistic categories from complex acoustic spaces, and the auditory primal sketch (Todd, 1994), which is capable of extracting the multi-dimensional features of the acoustic signal that are connected to prominence and rhythm, and represent them in an integrated fashion.

The implemented architecture consists of a number of auditory feature extraction components and, in recognition mode, of a sequence of probabilistic binary classifiers that are based on the relevance vector machine.
In training mode, which simulates learning and memory, input to the system is a set of audio files

containing recordings of prefixes and pseudo prefixes, a set of corresponding category labels and, optionally, a set of vectors containing phonetic segmentation information for the audio files. Output is a set probabilistic RVM binary classifiers.
In recognition mode, which simulates probabilistic grammatical category assignment by listeners, input to the system are one or more audio files containing the recording of a prefix or pseudo prefix, a set of probabilistic RVM binary classifiers (which represent abstract linguistic categories) and, optionally, a vector containing phonetic segmentation information for the audio files. Output is a prefix probability score for each trained probabilistic binary classifier.

Goal of the simulation presented here was a qualitative comparison between model output and output from an eye-tracking experiment which is about to be published. In the eye-tracking experiment, subjects listened to a sentence that described one of two pictures which were presented to them. One of the pictures could be described by a sentence containing a true-prefixed word (such as "mistiming"), while the other could be described by a sentence containing a pseudo-prefixed word (such as "mysterious"). The sentences describing each pair of pictures were identical up to the critical prefix syllable, but differed after it. So, in the experiment, “target” referred to the case in which listeners, while hearing a sentence referring to one picture, also looked at that picture; while “competitor” referred to the case in which listeners, while hearing a sentence referring to one picture, looked at the other (the “wrong”) picture. In the match condition, sentences were spliced so that both the first part (up to the prefix) and the second part (after the prefix) of the sentence came from two different tokens of the same sentence. Conversely, in the mismatch condition, the first part of the sentence came from a sentence token in which the status of the prefix was different (a true prefix for a pseudo prefix, and vice versa), but the rest of the sentence was identical. Listeners had to click on the picture which corresponded to what they were hearing, and their looks to targets and prefixes over time were recorded.
A form of the eye-tracking experiment's output suitable for comparison with model output was the one provided in terms of proportion of looks to targets for the match and mismatch conditions. Proportion of looks represents average fixations to targets and competitors for the match and mismatch conditions, measured at time slices of 4 ms, and plotted by aligning all stimuli at word (prefix) onset.

The simulation consisted of training different sets of probabilistic binary classifiers and comparing qualitatively the resulting curves of recognition probabilities with the curves representing the proportion of looks in the match and mismatch target conditions from the eye-tracking experiment.
The model does account for observed perceptual differences between true and pseudo prefixes, which in the eye-tracking experiment are manifested in differences in proportion of looks to targets for the match and mismatch conditions for the time window that goes from 200 to 400 ms. These qualitative results are encouraging, and provide a further backing of the evidence that acoustic information is exploited by listeners at levels of linguistic analysis that go beyond the phonemic level and encompass grammatical distinctions.

Smith, R.; Baker, R. & Hawkins, S. (2012) 'Phonetic detail that distinguishes prefixed from pseudo-prefixed words'. *Journal of Phonetics*, 40, 689-705
Baker, R. (2008) *The production and perception of morphologically and grammatically conditioned phonetic detail*. PhD thesis, University of Cambridge
Anderson, J. (1990) *The adaptive character of thought*. Lawrence Erlbaum
Firth, J. (1948) 'Sounds and Prosodies'. *Transactions of the Philological Society*, 47, 127-152
Tipping, M. (2001) 'Sparse Bayesian Learning and the Relevance Vector Machine'. *Journal of Machine Learning Research*, 1, 211-244
Todd, N. (1994) 'The auditory "Primal Sketch": a multiscale model of rhythmic grouping'. *Journal of New Music Research*, 23, 25-70

UNA VALUTAZIONE OGGETTIVA DEI METODI PIÙ DIFFUSI PER L'ESTRAZIONE AUTOMATICA DELLA FREQUENZA FONDAMENTALE

Fabio Tamburini
FICLIT – Università di Bologna

Il pitch, e in particolare la frequenza fondamentale - F0 - che rappresenta la sua controparte fisica, è uno dei parametri percettivi più rilevanti della lingua parlata e uno dei fenomeni fondamentali da considerare attentamente quando si analizzano dati linguistici a livello fonetico e fonologico. L'estrazione automatica di F0 è di conseguenza oggetto di studio da lungo tempo e in letteratura esistono numerosissimi lavori che si pongono come obiettivo lo sviluppo di algoritmi in grado di estrarre in modo affidabile F0 dalla componente acustica degli enunciati, algoritmi che vengono comunemente identificati come PDA (*Pitch Detection Algorithm*).

Tecnicamente, l'estrazione di F0 è un problema tutt'altro che banale e, la grande varietà di metodologie applicate a questo problema ne dimostra l'estrema complessità, specialmente se si considera che difficilmente è possibile predisporre un PDA che funzioni in modo ottimale per le differenti condizioni di registrazione considerando che parametri come il tipo di parlato, il rumore, le sovrapposizioni, ecc. sono in grado di influenzare pesantemente le prestazioni di questo tipo di algoritmi. Gli studiosi impegnati sul versante tecnologico si sono spinti alla ricerca di tecniche sempre più sofisticate per questi casi estremi, ancorché estremamente rilevanti per la costruzione di applicazioni reali, considerando risolto, o magari semplicemente abbandonando, il problema dell'estrazione di F0 per il cosiddetto "*clean speech*". Tuttavia, chiunque abbia utilizzato i più comuni programmi disponibili per l'estrazione automatica di F0 è ben cosciente che errori di *halving* o *doubling* del valore di F0, per citare solo una tipologia di problemi, sono tutt'altro che rari e che l'identificazione automatica delle zone *voiced* all'interno dell'enunciato pone ancora numerosi problemi.

Ogni lavoro che propone un nuovo metodo per l'estrazione automatica di F0 ha ormai da anni il dovere di eseguire una valutazione delle prestazioni in rapporto agli altri PDA, ma, di solito, queste valutazioni soffrono delle tipiche mancanze che derivano da sistemi di valutazione approssimativi: ci si limita a esaminare un insieme molto limitato di algoritmi, spesso non disponibili nella loro implementazione, tipicamente considerando corpora non distribuiti, relativi a lingue particolari e/o che contengono specifiche tipologie di lingua parlata (parlato patologico, parlato disturbato da rumore, ecc.). A mio parere, due sono gli studi, tra i più recenti, che hanno eseguito valutazioni piuttosto complete e basate su corpora scaricabili liberamente (de Cheveigné, Kawahara 2002; Camacho, 2007). Questi studi utilizzano nella valutazione una singola metrica che misura un unico tipo di errore, non considerando o considerando parzialmente l'intero panorama di indicatori sviluppati a partire dal pionieristico lavoro di Rabiner e colleghi (1976), e quindi, a mio avviso, i risultati ottenuti sembrano essere piuttosto parziali, anche se questa metrica è diventata di fatto lo standard per chi esegue una valutazione dei PDA.

Ci sembra quindi rilevante, effettuare una valutazione completa della maggior parte dei PDA, con particolare attenzione per quelli disponibili liberamente e quelli frequentemente utilizzati dalla comunità scientifica, misurando le prestazioni di questi sistemi con un'ampia gamma di misure quantitative. In particolare analizzeremo le misure definite in (Rabiner, et. al 1976; Chu, Alwan, 2009; Lee, Ellis, 2012). Non abbiamo la possibilità di sviluppare in questo abstract una completa disamina dei pro e contro delle varie metriche che utilizzeremo nella valutazione, segnaliamo unicamente che, per varie ragioni, ci sembra più opportuno introdurre una nuova misura di performance che sia in grado di catturare, con un unico indicatore, tutte le tipologie di errore possibili. Definiamo quindi il *Pitch Error Rate* come:

$$PER = (E_{f0} + E_{voi \rightarrow unv} + E_{unv \rightarrow voi})/N_{frame}$$

dove $E_{voi \rightarrow unv}$ e $E_{unv \rightarrow voi}$ rappresentano il numero di frame erroneamente classificati tra *voiced* e *unvoiced*, mentre E_{f0} rappresenta il numero di frame *voiced* nei quali il PDA differisce dal gold standard per più di 10 campioni (come definito in (Rabiner, et. al 1976)), ovvero, per quanto riguarda i corpora considerati, 16Hz.

La valutazione si è avvalsa di due corpora considerati come *gold standard*, entrambi disponibili liberamente e largamente utilizzati in letteratura nella valutazione dei PDA:

- *Keele Pitch Database* (Plante, *et al*. 1995): è composto da 10 locutori, 5 maschi e 5 femmine, che leggono, in ambiente controllato, un piccolo brano bilanciato in lingua inglese ('*North Wind story*'). Il corpus contiene anche l'output di un laringografo, dal quale è possibile stimare con precisione il valore di F0.
- *FDA* (Bagshaw, *et al.* 1993): è un piccolo corpus contenente 5' di registrazione divisi in 100 enunciati, letti da due locutori un maschio e una femmina, particolarmente ricchi di fricative sonore, nasali, liquide e glide, suoni particolarmente problematici da analizzare da parte dei PDA. Anche in questo caso il gold standard per i valori di F0 è stimato a partire dall'output del laringografo e la lingua di riferimento è l'inglese.

La tabella seguente elenca gli algoritmi compresi nella valutazione e l'implementazione considerata. Nella scelta, oltre a includere i programmi maggiormente utilizzati, si è scelto di privilegiare quelli disponibili gratuitamente. Per la valutazione sono stati utilizzati i parametri standard per ogni algoritmo considerato, imponendo unicamente uno shift tra i frame di 0.01 sec.

ALGORITMO	IMPLEMENTAZIONE	Rif. BIBLIOGRAFICO
FXANAL	SFS v4.8/win	(Secrest, Doddington, 1983)
ESRPD	Edimburgh Speech Tools (pda)	(Bagshaw, *et al.* 1993; Medan, et al. 1991)
PRAAT	Praat v5.105 (To Pitch (ac)+Kill octave jump)	(Boersma, 1993)
RAPT	ESPS get_f0, Snack/Wavesurfer, SFS v4.8/win, e altri…	(Talkin, 1995)
YIN	http://www.ircam.fr/pcm/cheveign/sw/yin.zip	(de Cheveigné, Kawahara, 2002)
WU	http://www.cse.ohio-state.edu/pnl/shareware/wu-tsap03/	(Wu, *et al*. 2003)
SWIPE'	SPTK, v3.5	(Camacho, 2007)
YAAPT	http://ws2.binghamton.edu/zahorian/yaapt.htm	(Zahorian, Hu, 2008)
PEFAC	VoiceBox per Matlab	(Gonzalez, Brookes, M. 2011)
SAcC	http://labrosa.ee.columbia.edu/projects/SAcC/	(Lee, Ellis, 2012)

Le due tabelle seguenti mostrano i valori di performance ottenuti dai vari algoritmi rispetto alle metriche considerate, ordinati rispetto alla nuova metrica proposta (**PER**):

FDA corpus

PDA	PER	GPE20	RabGPE	RabVDE	PTE	VE	UE
RAPT	0.07128	0.01642	0.03958	0.05723	0.06523	0.06824	0.06222
SWIPE'	0.07517	0.00241	0.02158	0.06614	0.07765	0.12483	0.03047
YAAPT	0.07929	0.02102	0.05184	0.06153	0.06841	0.05890	0.07792
PRAAT	0.08401	0.07095	0.08961	0.05070	0.08645	0.13286	0.04004
SAcC	0.08541	0.00626	0.02573	0.07723	0.09327	0.14046	0.04609
WU	0.10327	0.01087	0.03518	0.09117	0.08843	0.05324	0.12363
YIN	0.11228	0.01674	0.03715	0.10019	0.10990	0.13009	0.08972
FXANAL	0.11657	0.02881	0.05672	0.09791	0.11675	0.14586	0.08765
ESRPD	0.11801	0.06126	0.07140	0.09760	0.14673	0.28213	0.01132
PEFAC	0.14273	0.05188	0.09448	0.10896	0.12072	0.10295	0.13849

KEELE corpus

PDA	PER	GPE20	RabGPE	RabVDE	PTE	VE	UE
RAPT	0.07441	0.01792	0.03283	0.05866	0.06811	0.07117	0.06505
SWIPE'	0.08097	0.00290	0.00885	0.10864	0.11057	0.19941	0.02173
YAAPT	0.08139	0.01948	0.03307	0.06548	0.07462	0.06828	0.08096
SAcC	0.09836	0.01377	0.02067	0.08981	0.09538	0.14503	0.04574
YIN	0.12689	0.01431	0.02271	0.11710	0.12501	0.14651	0.10352
WU	0.12801	0.02791	0.03598	0.11190	0.12572	0.11393	0.13751
FXANAL	0.14714	0.04124	0.05870	0.12097	0.13907	0.13365	0.14450
ESRPD	0.18417	0.04690	0.05545	0.16225	0.17789	0.34164	0.01413
PRAAT	0.20324	0.25643	0.26486	0.08894	0.20158	0.34324	0.05991
PEFAC	0.29194	0.10602	0.18897	0.21376	0.25881	0.28114	0.23647

Nella comunicazione si presenterà una discussione articolata sulla complessa analisi dei risultati. Qualitativamente, le migliori prestazioni si evidenziano per gli algoritmi RAPT, SWIPE' e YAAPT che risultano essere più stabili e performanti rispetto a vari indicatori.

BIGLIOGRAFIA

Bagshaw P. C., Hiller S. M., Jack M. A. (1993). “Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching”. In *Proc. Eurospeech '93, Berlin*, 1003-1006.

Boersma P. (1993), "Accurate short-term analysis of the fundamental and the harmonics-to-noise ratio of a sampled sound.", in *Proceedings of the Institute of Phonetic Sciences*, University of Amsterdam , 17, 97–110.

Camacho A., (2007). “SWIPE: A sawtooth waveform inspired pitch estimator for speech and music”, PhD Thesis, University of Florida.

Chu W., Alwan A. (2009), “Reducing F0 frame error of F0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend”, in Proc. ICASSP2009.

de Cheveigné A., Kawahara H. (2002), "YIN, a fundamental frequency estimator for speech and music”, JASA, 111(4):1917-30.

Gonzalez S., Brookes, M. (2011), “A pitch estimation filter robust to high levels of noise (PEFAC)”, in *Proc EUSIPCO* 2011.

Lee B.S., Ellis D. (2012), "Noise Robust Pitch Tracking by Subband Autocorrelation Classification", In *Proc. Interspeech* 2012, Portland (OR).

Medan Y., Yair E., and Chazan D. (1991). "Super resolution pitch determination of speech signals", *IEEE Trans. Sig. Proc*. 39, 40–48.

Plante F., Ainsworth W.A., Meyer G. (1995) "A Pitch Extraction Reference Database", in *Proc. Eurospeech'95*, Madrid, 837-840.

Rabiner L.R., Cheng M.J., Rosenberg A.E., McGonegal C.A. (1976), “ A Comparative Performance Study of Several Pitch Detection Algoritms”, *IEEE Trans. Ac., Sp. Sig. Proc*., 24(5).

Secrest B., Doddington G. (1983), "An integrated pitch tracking algorithm for speech systems", in Proc. *ICASSP-83*, 1352-1355.

Talkin D. (1995), "A robust algorithm for pitch tracking (RAPT)", in W. B. Kleijn & K. K. Paliwal (eds.) *Speech Coding and Synthesis*, New York: Elsevier.

Wu M., Wang D.L., Brown G.J. (2003), “A multipitch tracking algorithm for noisy speech”. *IEEE Transactions on Speech and Audio Processing*, 11, 229-241.

Zahorian S.A., Hu H. (2008), "A Spectral/temporal method for Robust Fundamental Frequency Tracking". *JASA*, 123 (6).

Analisi qualitativa del modello C2H per il controllo del contrasto fonetico nella sintesi del parlato

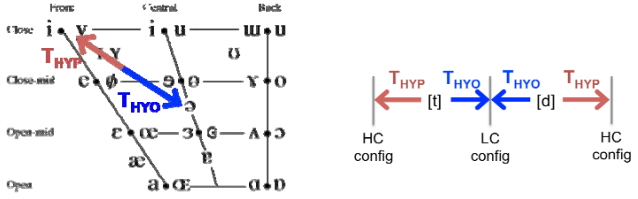
Mauro Nicolao, Roger K. Moore

Speech and Hearing Group, Dept. Computer Science, University of Sheffield, UK
m.nicolao@dcs.shef.ac.uk , r.k.moore@dcs.shef.ac.uk

In questo lavoro è presentata una valutazione attraverso due tipi di analisi oggettiva del modello *C2H (Computational model of the H&H theory)* proposto in (Moore & Nicolao, 2011) e (Nicolao et al., 2012) e basato sull’ipotesi che esistano dei punti di attrazione a basso contrasto fonetico verso cui la produzione orale tende a convergere. Le differenze acustiche tra le unità fonetiche e questi attrattori identificano alcune speciali direzioni lungo le quali ogni punto rappresenta un diverso grado contrasto. La possibilità di muoversi lungo questa direzione rappresenta quindi un sistema per controllare lo sforzo nell’articolazione del parlato.

Il modello C2H prende spunto dall’osservazione che gli esseri umani, quando parlano, adattano la loro produzione in base al contesto in cui la comunicazione ha luogo, *effetto Lombard* (Lombard, 1911), e in risposta a diverse esigenze del destinatario della comunicazione (Moore, 2007). Un denominatore comune a molte osservazioni è il controllo costante del parlante sulla sua produzione e il rapido adattamento in risposta all’evoluzione delle condizioni. La comunicazione verbale può quindi essere descritta come un processo di ottimizzazione che massimizza il trasferimento di concetti dal parlante all’ascoltatore, minimizzando, lo sforzo coinvolto nel gesto. Una formalizzazione di questo concetto si trova nella teoria *H&H (Hyper and Hypo)* di Lindblom, (Linblom, 1990) dove i termini *ipo* e *iper* si riferiscono al grado di articolazione con cui il parlato è prodotto.

Seguendo il metodo in (Nicolao et al., 2012), è stato realizzato un sistema di sintesi automatica del parlato basato su modelli statistici (*TTS-HTS*) per la lingua inglese. E’ stato addestrato, inoltre, un insieme di trasformazioni lineari tali da incrementare o diminuire le distanze dei principali parametri acustici (forma dello spettro, frequenza fondamentale e durata) di ogni fonema rispetto ai punti di attrazione. Una schematizzazione del metodo è mostrata in Figura 1.



(a) Controllo produzione vocali (CPV) (b) Controllo produzione consonanti (CPC)

Figura 1: Rappresentazione grafica (frecce blu, T_{HYO}) delle trasformazioni verso l’ipo-articolazione o configurazione a basso contrasto (LC) tali per cui (a) la produzione di ogni vocale viene ridotta verso [ə] e (b) ogni consonante viene fatta muovere verso il fonema con cui più facile confonderla. Le trasformazioni (frecce rosse, T_{HYP}) verso l’iper-articolazione o configurazione ad altro contrasto (HC) fonetico agiscono in direzione opposta alle precedenti.

L’analisi dell’audio generato da questo sistema è stata effettuata con indici di valutazione oggettiva differenti rispetto a (Nicolao et al., 2012) in modo verificare ulteriormente la validità delle trasformazioni proposte.

La prima analisi sull’audio prodotto dal sintetizzatore è stata di tipo acustico. Sono stati analizzati dei campioni audio prodotti con tre diversi gradi di articolazione: a basso contrasto fonetico (*HYO*), audio standard del sistema TTS-HTS (*STD*) e ad alto contrasto fonetico (*HYP*). Alcuni dei più comuni parametri acustici sono stati misurati:

- durata media delle parole (*MWD*) e delle frasi (*MSD*),
- misure sullo spettro medio: energia nell’intervallo di frequenze 1-3 kHz (*LTAS13*), inclinazione (*Spectral Tilt*) e baricentro (*CoG*),
- frequenza fondamentale media (*F0 mean*) e intervallo (*F0 range*),
- estensione dello spazio delle principali formanti delle vocali (*F1F2 area*).

I risultati di questa analisi sono riportati in Tabella 1 per entrambe le trasformazioni.

	CPV			CPC		
	HYO	*STD*	*HYP*	*HYO*	*STD*	*HYP*
MWD (s)	0.27	0.318	0.356	0.311	0.318	0.33
MSD (s)	2.98	3.501	3.91	3.43	3.501	3.592
LTAS13 (dB SPL)	33.6	36.2	41.1	35.4	36.2	38.4
Spectral Tilt (dB/dec)	-6.1	-5.8	-4.9	-6.1	-5.8	-5.1
CoG (Hz)	712	821	1024	547	821	1156
F0 mean (Hz)	172.6	174.2	174.7	174.1	174.2	173.4
F0 range (Hz)	146-185	151-183	145-190	144-185	151-183	150-184
F1F2 area (Hz^2)	1014	29021	70509	41824	29021	56103

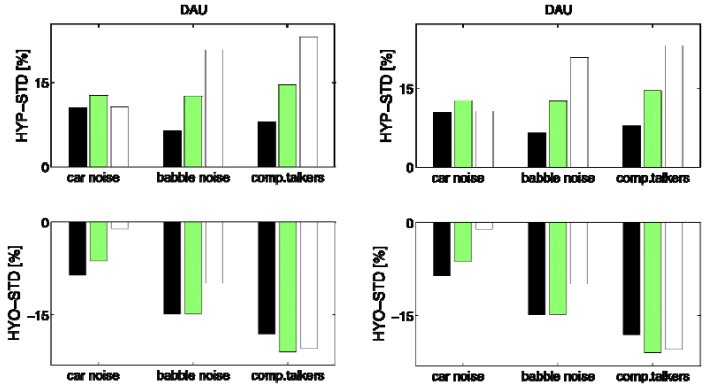
Tabella 1: Confronto dei risultati dell'analisi acustica sui campioni prodotti dal sistema TTS-HTS a diversi gradi di contrasto fonetico con trasformazione delle vocali (CPV, parte sinistra) e delle consonanti (CPC, parte destra).

Dai dati riportati emerge abbastanza chiaramente che queste trasformazioni, addestrate unicamente per incrementare o diminuire la distanza acustica tra fonemi facilmente confondibili, producono dei cambiamenti nel parlato simili a quelli osservati nell'effetto Lombard. Si notano, infatti:

- un allungamento della durata media delle parole, maggiore nel caso CPV,
- un cambiamento della distribuzione dell'energia dello spettro con un evidente spostamento del baricentro in CPC,
- una conferma per CPV che la trasformazione controlla efficacemente l'estensione dello spazio delle vocali (F1F2 area).

In nessuna delle due trasformazioni si rilevano sostanziali modifiche della frequenza fondamentale poiché il metodo di addestramento utilizzato non ha permesso di registrare modifiche relative a questo parametro.

Il secondo tipo di analisi si basa sulla considerazione che l'aumento del contrasto fonetico influisce sull'intelligibilità del parlato. La qualità della trasformazione è stata quindi valutata misurando la variazione d'intelligibilità, stimata attraverso l'indice *DAU* (Dau et al., 1996), al variare del grado di contrasto fonetico del parlato prodotto e dell'ambiente rumoroso.



(a) (b)

Figure 2: Differenze percentuali tra l'indice DAU per l'audio STD e, rispettivamente, quello per HYP e HYO. Tre differenti disturbi sono stati considerati: rumore all'interno di un auto (*car noise*), molti parlatori distanti (*babble noise*) e uno/due parlatori vicini (*comp. talkers*). Tre diverse intensità di disturbo: SNR = 6 dB (colonna nera), SNR = 0 dB (verde) e SNR = -6 dB (bianca). L'intensità media del segnale è stata mantenuta constante: RMS = -24 dB FS.

Il grafico, riportato in Figura 2, mostra come, diminuendo il contrasto fonetico (*HYO*) si ottiene una diminuzione dell'indice di intelligibilità rispetto alla normale produzione del sistema di sintesi (*STD*), mentre agendo nel senso opposto (*HYP*), il valore dell'indice aumenta per tutti i tipi di rumore e tutti i differenti livelli di SNR.

I risultati proposti in questo lavoro confermano quindi che alcune delle caratteristiche proprie del parlato in ambiente rumoroso possono essere modellate tramite il sistema C2H, partendo da semplici considerazioni di aumento/diminuzione del contrasto fonetico.

Bibliografia

Dau, T., Puschel, D. & Kohlrausch, A., "A quantitative model of the effective signal processing in the auditory system. I. Model structure," JASA, vol. 99, no. 6, pp. 3615–3622, Jun. 1996

Lindblom, B, "Explaining phonetic variation: a sketch of the H&H theory," Speech production and speech modelling, vol. 55, pp. 403–439, 1990

Lombard, E., "Le Signe del'Elevation dela Voix – The sign of the rise in the voice", Ann. Maladiers Oreille, Larynx, Nez, Pharynx - Annals of diseases of the ear, larynx, nose and pharynx, vol. 37, pp. 101–119, 1911

Moore, R. K., "PRESENCE: A Human-Inspired Architecture for Speech-Based Human-Machine Interaction," IEEE Transactions on Computers, vol. 56, no. 9, pp. 1176–1188, Sep. 2007

Moore, R. K. & Nicolao, M., "Reactive Speech Synthesis: Actively Managing Phonetic Contrast Along an H&H Continuum," in ICPhS 2011, Hong Kong, China, Aug. 2011, pp. 1422–1425

Nicolao, M., Latorre, J. & R. K. Moore, "C2H: A Computational Model of H&H-based Phonetic Contrast in Synthetic Speech," in INTERSPEECH 2012, Portland, OR, Sep. 2012

Decodifica di vocali percepite, immaginate e articolate tramite segnale elettroencefalografico

Anna Dora Manca, Mirko Grimaldi
Centro di Ricerca Interdisciplinare sul Linguaggio (CRIL), Dipartimento di Studi Umanistici, Università del Salento
annadoramanca@gmail.com, mirko.grimaldi@unisalento.it

Introduzione

L'apparente facilità con cui percepiamo il linguaggio contrasta con la complessità degli atti motori necessari per produrre concatenazioni di suoni e con la complessità del segnale acustico generato (Halle 2003). Studi pionieristici come quello di Libermann et al. (1985) hanno posto l'accento sulla relazione tra sistema percettivo e motorio nei processi di percezione e produzione linguistica, postulando un modulo cerebrale predisposto all'elaborazione diretta dei gesti articolatori implicati nella produzione dei suoni linguistici.

Recenti evidenze neurocognitive supportano l'ipotesi che la stessa area della corteccia cerebrale che controlla l'apparato fonatorio si attivi sia quando produciamo sia quando articoliamo un suono, ma anche quando immaginiamo di articolarlo, attraverso un processo d'integrazione sensorimotoria (Guenther 2006; Tian & Poeppel 2010). Da un'altra prospettiva sono state fornite evidenze neurofisiologiche, non ancora del tutto definite, circa l'elaborazione dei cosiddetti *tratti distintivi* (Philips et al. 2000; Obleser et al. 2003, 2004). Nello stesso tempo altri studi hanno iniziato a chiarire la correlazione fra l'attività cerebrale e differenti aspetti dei processi di percezione e produzione o d'immaginazione di percepire e produrre suoni (Janata, 2001; Mitchell et al. 2008; Suppes & Han 2000; Formisano et al. 2008; Meyer et al. 2007; DaSalla et al. 2009; Deng et al. 2010; Kellis et al. 2010). In particolare, alcuni studi, grazie a registrazioni EEG intracorticali con elettrodi collocati direttamente sulla corteccia cerebrale, hanno dimostrato correlazioni dirette fra vocali e consonanti prodotte o immaginate e alcune aree cerebrali deputate all'elaborazione dei suoni linguistici (Pey et al. 2001; Tankus et al. 2012). Tuttavia, le registrazioni intracorticali sono invasive e possono essere utilizzate solo in soggetti che necessitano interventi neurochirurgici specifici.

Obiettivi

Questo studio, utilizzando una metodica EEG non invasiva, ovvero i Potenziali Evento Correlati (ERP), si propone di indagare i correlati neurofisiologici della percezione vocalica (**P**) rispetto a tre livelli di produzione: (1) articolare vocali con suono (**AVS**); (2) articolare vocali senza suono (**AV**); (3) immaginare di articolare vocali (**IA**). L'ipotesi è che il sistema uditivo sia coinvolto nella produzione e che il sistema motorio sia criticamente coinvolto nella percezione del parlato.

Metodo sperimentale

Tredici soggetti italiani (7 femmine; età media 25 ± 3), ai quali è stata montata una cuffia EEG a 64 canali (software di acquisizione BCI200), seduti di fronte allo schermo nero di un computer (19"), hanno ascoltato tramite altoparlanti le vocali /a, i/ (durata dello stimolo 300ms) in modo randomizzato (3 blocchi di 80 ripetizioni randomizzate x ogni stimolo acustico). Ascoltata la vocale, una croce bianca sullo schermo, seguita da uno schermo nero (intervallo di tempo randomizzato 200-500ms) segnalava ogni volta al soggetto di prepararsi per eseguire i successivi compiti sperimentali alla comparsa di uno schermo bianco (durata 2sc) nella sequenza: (1) **AVS**; (2) **AV**; (3) **IA** (cfr. Fig. 1).

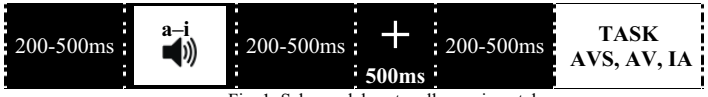


Fig. 1: Schema del protocollo sperimentale

I dati sono stati filtrati offline e gli artefatti oculari e oro-facciali sono stati rimossi con l'ICA (Stone 2002). Le epoche sono state estratte con due tipi di segmentazione in base all'onset dello stimolo d'interesse: stimolo acustico per analisi del compito percettivo, schermo bianco per l'analisi dei diversi compiti di produzione. N100 e P200 sono le componenti osservate, una deflessione del segnale EEG, negativa la prima e postiiva la seconda. N1/P2 appaiono rispettivamente 100ms e 200ms dopo lo stimolo e sono ben note per essere risposte ERP obbligatorie che riflettono rappresentazioni uditive centrali senza la partecipazione attiva dei soggetti (Hillyard et al. 1983; Näätanen & Picton 1987).

Risultati

L'ispezione visiva delle onde ERP ottenute dal Grand-average delle registrazioni dei 13 soggetti ha evidenziato la presenza dei classici picchi correlati alle componeti N1/P2 in tutti e quattro i compiti sperimentali (confermata da un t-test against zero, p<.005, negli elettrodi Cz e Fz). La comparazione dell'attività media delle due vocali nelle quattro condizioni è avvenuta considerando piccole finestre temporali di 100ms, a partire da 50ms dopo l'onset dello stimolo di interesse fino a 250ms, tramite un'ANOVA a misure ripetute con fattori le vocali (2: /a, i/) e gli elettrodi (3: FCZ-CZ-FZ). Di seguito sono riportati solo i risultati significativi (cfr. Figg. 2 e 3):

- 50-150ms: /i/ > /a/ sia nel compito **P** (F(1, 12)=7,367; p=0.019)) sia in quello **AVS** (F(1, 12) = 6,576; p=0.02)). Il post-hoc mostra che tale differenza è in CZ.
- 150-250ms: /i/ = /a/ in **P** (F(1, 12) = ,095; p=763)). Si rileva differenza per il fattore elettrodo (F(1,110 13,324) = 6,887; p=,019)). Il post-hoc mostra che /a/ è distribuita in maniera significativa in tutti e tre gli elettrodi, mentre /i/ è maggiore in Cz (p=0.036). Nel compito **IA** /i/ = /a/ (F(1, 12) = 3,224; p=,098)), ma tende alla significatività per il fattore elettrodo (F(1,300 15,596) = 3,924; p=,057)). Il post-hoc rivela che /i/ > /a/ in CZ (p=,014).

Un'ANOVA a misure ripetute ha anche indagato la distribuzione topografica dell'attività neurale rispetto a ogni compito sperimentale con fattore 2 vocali e 3 regioni di interesse (ROI): fronto-centrale (Fcz-Cz-Fz), emisfero sinistro (C1-C3-C5) emisfero destro (C2-C4-C6). I risultati significativi mostrano che:

- 50-150ms: /i/ > /a/ (p<,005) in **P** (F(1, 38) = 32,843; p=0.000)), **AVS** (F(1, 38)=4,344; p=0.044)) e **AV** (F(1, 38)=4,655; p=0.037). In particolare, in **P** il post-hoc mostra che la regione fronto-centrale è più attiva per entrambe le vocali (p=001) e che non c'è lateralizzazione emisferica (p=,108).
- 150-250ms: /i/ > /a/ in **IA** (F(1, 38)=10,666; p=0.002)). Il post-hoc mostra che in **IA** /i/ > /a/ nella regione fronto-centrale dello scalpo (p=0.005).

Discussione e conclusioni

Questo studio dimostra per la prima volta che le componenti ERP N1/P2 si elicitano non solo durante processi percettivi acustici, ma anche nell'atto di articolare suoni con segnale acustico, nell'atto di articolare suoni senza emissione di suono e, fatto ancora più importante, nell'atto d'immaginare di articolare suoni (cfr. Ganuschchak et al. 2011). Dal momento che nelle tre condizioni articolatorie era assente ogni stimolo uditivo, l'interpretazione più coerente è che la produzione dei suoni linguistici sia mediata dagli stessi sostrati neurali che sovrintendono alla percezione attraverso una sorta di copia efferente dei processi somatosensoriali (cfr. Tian & Poeppel 2010). In sintesi, i sistemi neurali percettivi sono anche coinvolti durante i processi di produzione, inclusi quelli non vocali (**AV** e **IA**).

Un altro risultato interessante di questo studio è la conferma che le caratteristiche spettro-acustiche delle vocali si riflettono nell'attività neurofisiologica. Infatti, le componenti N1/P2 risultano modulate in funzione delle peculiari caratteristiche formantiche di /a, i/: l'attività corticale di /i/ è maggiore di /a/ (Ohl et al. 1997; Obleser et al. 2003). In modo interessante, i risultati dimostrano che tali modulazioni avvengono sia in percezione che in produzione, e soprattutto anche per **AV** e **IA**, nell'area fronto-centrale. Infine, tempi di modulazione sono precoci nei compiti di **P**, **AVS** e **AV** e più lenti per il compito di **IA**, il cui processo di copia efferente somatosensoriale richiede probabilmente un tempo maggiore.

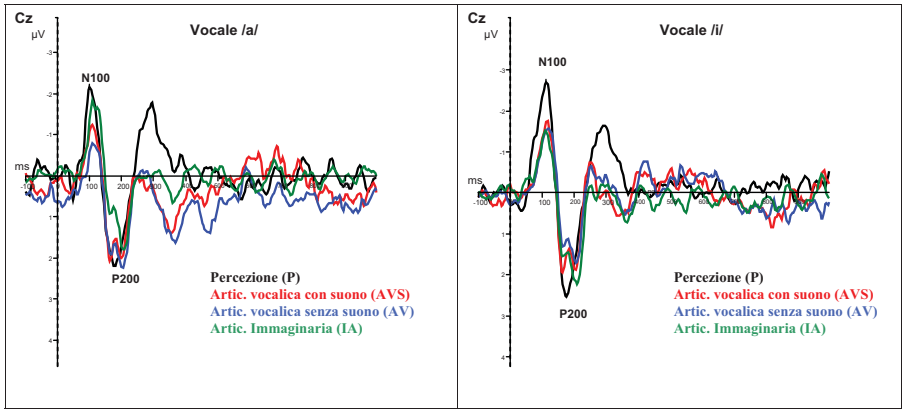


Fig. 2: Grand-average dell'attività neurale in risposta agli stimoli /a, i/ nelle quattro condizioni sperimentali.

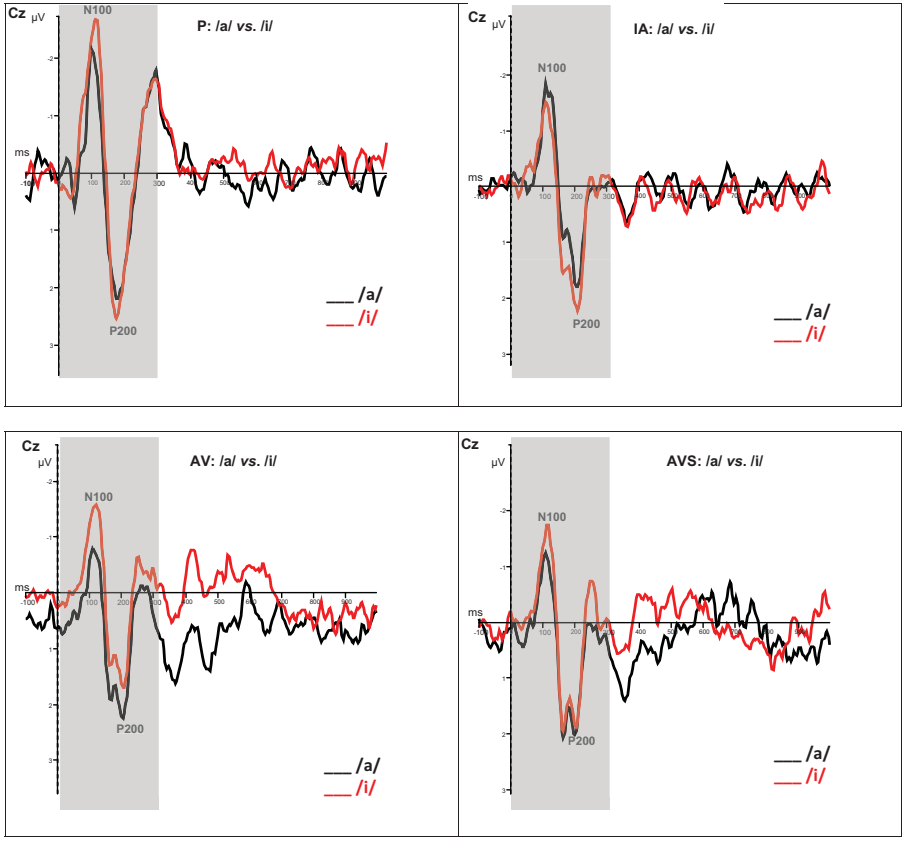


Fig. 3: Confronto del Grand-average /a/ vs. /i/ nelle quattro condizioni sperimentali.

BIBLIOGRAFIA

DaSalla C S et al 2009 Single-trial classification of vowel speech imagery using common spatial patterns. *Neural Netw., 22, 1334–9.*

Deng S et al 2010 EEG classification of imagined syllable rhythm using Hilbert spectrum methods. J. *Neural Eng. 7.*

Janata P. (2001). Brain electrical activity evoked by mental formation of auditory expectations and images. *Brain Topography, 13, 169-193.*

Formisano E. et al 2008 'Who' is saying "what"? Brain-based decoding of human voice and speech, Science, 322, 970-973.

Ganushchak L. I., Christoffels, I. K., Schiller, N. O., 2011, The use of electroencephalography in language production research: a review, Frontiers in Psycholoy, September, 2, 208.

Guenther F.(2006). Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders 39. 350–365.*

Halle, M. (2003), *From Memory to Speech and Back*, Berlin, Mouton.

Kellis S et al 2010 Decoding spoken words using local field potentials recorded from the cortical surface J. Neural Eng. 7 056007.

Liberman A.M., Mattingly I.G. (1985). The motor theory of speech perception revised. *Cognition 21, 1*-36.

Mitchell T.M., Shinkareva S.V., Carlson A., Chang K.M., Malave V.L., Mason R.A., Just M.A .(2008). Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science 320, 1191-1195.*

Meyer M., Elmer S., Baumann S., Jancke L. (2007). Short-term plasticity in the auditory cortex: Differential neuronal responses to perception and imagery of speech and music. *Restorative neurology and neuroscience 25. 411-431.*

Naatanen, R., & Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: A review and analysis of the component structure. *Psychophysiology, 24, 375–425.*

Nunez L.P., Srinivasan R., (2006). Electric field of the brain. *Oxford University Press.*

Ohl F.W., Scheich H. (1997) Orderly cortical representationof vowels based on formant interaction. *Proc. Natl. Acad Sci. USA. 9440-9444.*

Obleser J., Elbert T., Lahiri A., Eulitz C. (2003). Cortical representation of vowels reflects acoustic dissimilarity determined by formant frequencies. *Cognitive Brain Research, 15, S: 207-213.*

Obleser J., Lahiri A., Eulitz C. (2004). Magnetic Brain response mirrors extraction of phonological features from speakers vowels. *J. of Cognitive Neuroscience, 16: 31-39.*

Obleser J., Elbert T., Eulitz C. (2004). Attentional influences on functional mapping of speech sounds in human auditory cortex. *BMC Neuroscience, 5-24*

Phillips C., Pellathy T., Marantz A., Yellin E., Wexler K., Poeppel D., McGinnis M., & Roberts T. (2000). Auditory cortex accesses phonological categories: an MEG mismatch study. *Journal of Cognitive Neuroscience, 12*, 1038-1045

Roberts T. D., Ferrari P., Stufflebeam S.M., P., Poeppel (2000). Latency of the auditory evoked neuromagnetic field components: stimulus dependence and insights towards perception. *J. Clin. Neurophysiol. 17, 114-129.*

Stone J.V. (2002). Independent component analysis: an introduction. *TRENDS in Cognitive Sciences Vol.6 N.2 59-64*

Suppes P and Han B 2000 Brain-wave representation of words by superposition of a few sine waves. *Proc. Natl Acad. Sci. USA 97: 8738–43.*

Tankus A., Fried I., Shoham S., 2012, Structured neuronal encoding and decoding of human speech features, Nature, 3, 10015.

Tian X., Poeppel D. (2010). Mental imagery of speech and movement implicates the dynamics of internal forward models. *Frontiers in Psychology. Volume 1, Article 166.*

Vihla M., Eulitz C. (2003). Topography of the auditory evoked potential in human reflects differences between vowels embedded in pseudo words. *Neuroscience Letters 338, 189-192.*

The mapping between prosody and information structure in German and in Italian L2 learners. Who tranfers what?

C. Avesani, G. Bocci, M. Vayra, A. Zappoli

1. Romance and Germanic languages are claimed to differ in their use of the prosodic marking of discourse-related properties. Germanic languages massively deaccent Given information (e.g. Ladd, 1996); the pervasiveness of such a property originally induced Cruttenden (1993) to claim it to be a cognitive universal. However, it has been observed that Romance languages fail to deaccent Given information (e.g. Ladd, 1996; Swerts et al., 2002; Avesani et al., 2005). The picture is not so clear-cut: English and German can accent Given entities (e.g., Terken and Hirschberg, 1994; Bauman, 2008), and Italian requires deaccenting in some configurations (e.g., post-focal elements). Crucially, though, deaccenting in Italian has been shown to be void of any role in marking the information status of an entity and to be only driven by phonological requirements on the prosodic structure (Bocci, in press); while in German items can be deaccented by virtue of being Given in the discourse or by virtue of the syntactic configuration in which the constituent occurs (Truckenbrodt, 2011).

2. The present work addresses the question of how the information status of a discourse entity is prosodically realized by romance learners of a germanic language and by germanic learners of a romance language. We will address the issue of whether differential learning patterns emerge in two groups of speakers, Italians learning L2-German, and Germans learning L2-Italian, by examining how given, New and constrative information is intonationally realized in their interlanguage compared to their source language and their target language. Our hypothesis is that speakers of a "plastic" language such as German (Vallduvì, 1992), in which deaccenting cues aspects of both pragmatic structure and syntactic structure, will have less difficulty in learning the intonational patterns of Italian, a "non-plastic" language in which deaccenting is ruled by phonological constraints, compared to speakers of L1-Italian learning L2-German.

3. In our production study we adopted the experimental setting previously used by Swerts et al. (2002), where the New, Given and constrative pragmatic status of an Adjective and of a Noun was systematically changed within the same DP. Accent patterns for L1- and L2-Italian and for L1- and L2-German are obtained via a simple dialogue game played by 4 pairs of Italian speakers and by 2 pairs of German ones. Each pair of speakers played the game twice: first in his/her L2, then in his/her L1. The setting aimed at eliciting a (semi)spontaneous conversation between the two players of a card game; such a game was essentially an alignment task of figures played by the two subjects in 64 moves. In each game, both players had an identical set of eight cards to their disposal, each card showing the picture of a fruit (a banana or a melon) in a particular colour (lilac or green). We obtained a set of spoken DPs (N+Adj) that allow an unambiguous operationalization of the relevant contexts. The whole set of pragmatic combinations in which the target Ns and Adjs could occur are the following: New-New (beginning of game); Contrastive-Given; Given-Contrastive; Contrastive-Contrastive.

The target items are all-voiced and matching by stress position and by segment composition as much as possible between the languages. 10 speakers participated in the game: 6 Italians, fluent in L2-German, and 4 Germans, fluent in L2-Italian. Among the Italian speakers, 5 out of 6 are learners of L2-German at a B2 or C1 level of proficiency and have been studying German for at least 8 years. The 4 German speakers declare an equivalent level of proficiecy but they have been studying L2-Italian for much a shorter period of time (2 years on average). We obtained 24 occurrences of New and Given items and 48 occurrences of Contrastive items for L1-Italian and L2-German, and 16 occurrences of New and Given items and 32 occurrences of Contrastive items for L1-German and L2-Italian (total=160 items). Each DP was ToBI transcribed and measures for syllable and vowel duration, pitch accent (PA) alignment and scaling were calculated.

4. The distributional analysis of PA association as a function of the DP's pragmatic status (see fig.1) shows that: 1) in L1-Italian Given information is pitch accented as much as Contrastive and New information, confirming previous data. 2) In L1-Italian, word1 - in contrast to word2 - can be optionally left unaccented. We argue that the lack of PAs on word 1 must be imputed only to phonological reasons: only rightmost elements in phrasal prosodic constituents are mandatorily accented. In fact, the lack of PAs is unrelated to the pragmatic status of word1. 3) In L1-German Given infomation is deaccented 100% of the times only in nuclear position (word 2), while it is mostly accented in prenuclear position (contra Bauman, 2008); however, if deaccenting occurs in word 1, it only occurs on Given items, differently from Italian. 4) Germans do always accent Given items in their L2-Italian, while Italian speakers fail to deaccent them in nuclear position in L2-German. 4) Duration is not a significant acoustic correlate of information status in neither language.

Overall, our results show that at the level of the mapping between prosody and information structure, Italians transfer their L1 intonation onto their L2-German, while Germans master the Italian intonational patterns. These results confirm those obtained by Rasier and Hiligsman (2007) on French and Dutch, and support Eckman's Differential Markedness Hypothesis (Eckman, 1977). However, looking at the structural components of the speakers' interlanguage intonation, Germans, although mastering the Italian PA distribution, do tranfer onto L2-Italian their PAs phonological inventory. Analogously, Germans transfer onto L2-Italian the "phonetic details" of their L1 pitch accent alignment, which appears to be hard-wired in the segmental string via a language-specific tone-segment coordination. We will discuss the learning mechanisms at different levels: the pragmatic-prosody interface, the phonological and the phonetic level.

References

Avesani C. & Vayra M. (2005). Accenting deaccenting and information structure in Italian dialogues. In L. Dybkjaer e W. Minker (eds), *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*, Lisbona, 2-3 settembre 2005, pp. 19-24.

Baumann S. (2008). Degrees of Givenness and their Prosodic Marking. In Riehl, C.M. & As. Rothe (eds.): *Was ist linguistische Evidenz?*, pp. 35-55.

Bocci, G. (in press). *The syntax-prosody interface from a cartographic perspective:*

evidence from Italian, John Benjamins.
Cruttenden, A. (1993). The de-accenting and reaccenting of repeated lexical items. In *Proceedings of the ESCA workshop on Prosody, Lund*, pp. 16-19.
Ladd R. (1996). *Intonational Phonology*, Cambridge, Cambridge University Press.
Rasier L. & Hiligsman P. (2007). Prosodic transfer from L1 to L2. Theoretical and methodological issues, *Nouveaux cahier de linguistique française*, *28*, 41-66.
Swerts M., Krahmer E. & Avesani C (2002). Prosodic marking of information status in Dutch and Italian: a comparative analysis, *Journal of Phonetics*, *30*, 4:629-65
Vallduvì E. (1992). *The Informational Component*, New York-London, Garland Publishing.
Eckman, F.R. (1977). Markedness and the Contrastive analysis hypothesis, *Language Learning,* 27, pp. 315 – 330.
Terken, J. and Hirschberg, J. (1994) Deaccentuation of words representing Given information: effects of persistence of grammatical function and surface position, *Language and Speech*, 37(2), 125—145.
Truckenbrodt, H. (2011). The syntax-phonology interface. In J. Goldsmith, J. Riggle, & A. Yu (eds.), *The Handbook of Phonological Theory*, Cambridge University Press, pp. 196-196).

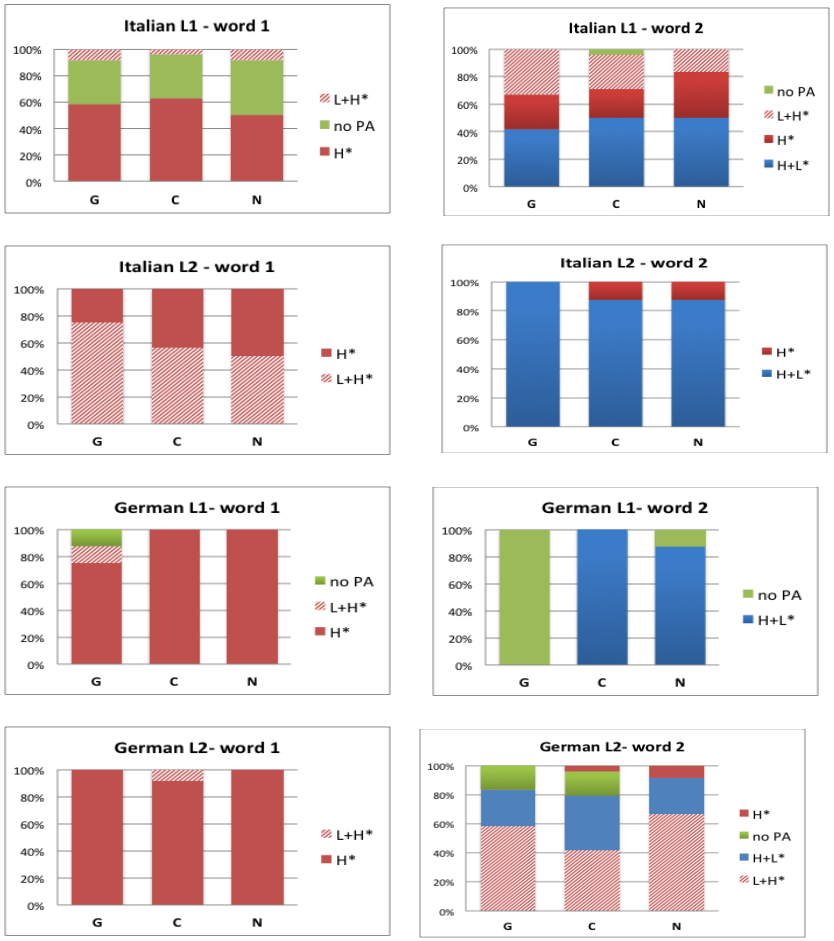


Figure 1. Pitch accent distribution in Italian L1 and L2 and in German L1 and L2 according to the information status of word 1 (in Italian: Adjective; in German: Noun) and word 2 (in Italian: Noun; in German: Adjective). G= Given, C = Contrastive, N = New.

XXX*: il Corpus dei Sottotitoli Multilingue degli Interventi alle Conferenze TED

Sommario

I dati giocano un ruolo chiave nell'apprendimento automatico – noto in letteratura come Machine Learning – essendo essi la principale sorgente di informazione da cui inferire i valori dei parametri dei modelli matematici in uso.

Nella traduzione automatica statistica (statistical machine translation, SMT), l'apprendimento viene compiuto su testi paralleli, ovvero documenti, frasi o anche semplici frammenti di frasi accoppiati alle loro rispettive traduzioni in una o più lingue. È tipico che per addestrare adeguatamente i modelli di traduzione e di riordinamento di un sistema SMT sia necessario impiegare una grande quantità di dati paralleli, possibilmente nel dominio semantico di interesse.

Purtroppo, i dati paralleli sono una risorsa scarsa, disponibile solo per alcune coppie di lingue e per pochi domini, spesso molto specifici. Ad esempio, MultiUN [1] fornisce una quantità notevole di dati paralleli, ma per sole sei lingue; Europarl [2] include la traduzione nella maggior parte delle lingue europee degli atti del Parlamento Europeo (fino a 50 milioni di parole); JRC-Acquis[1] comprende l'intero corpo della legislazione dell'Unione Europea che si applica agli Stati membri, tradotta completamente o parzialmente in 22 lingue (da 30 a 60 milioni di parole per ciascuna lingua); altri corpora paralleli più piccoli per domini molto specifici si trovano in OPUS [3] per alcune decine di lingue.

D'altro canto, è impensabile per i laboratori di ricerca coprire ogni possibile esigenza in termini di corpora paralleli ricorrendo a traduttori professionisti, dato il loro alto costo.

I dati disponibili sul sito di TED[2] risultano quindi particolarmente preziosi per la comunità della traduzione automatica. TED è una organizzazione nonprofit che invita "gli intellettuali ed i professionisti più brillanti a tenere il discorso della loro vita". Il sito rende disponibili, con licenza Creative Commons BY-NC-ND, le registrazioni audiovideo dei migliori interventi con tanto di sottotitoli in inglese e la loro traduzione eseguita da volontari. L'insieme dei sottotitoli rappresenta pertanto una risorsa parallela multilingue di valore inconfutabile, giacché cresce continuamente nel tempo (ad oggi il sito mette a disposizione le registrazioni di oltre 1200 interventi), include le traduzioni in decine e decine di lingue (vi sono interventi tradotti in 92 idiomi diversi), italiano incluso, e copre argomenti che spaziano su tutto lo scibile umano, rendendo la risorsa potenzialmente utile per qualsiasi applicazione.

Con l'obiettivo di rendere questo corpus di fruibilità immediata presso la comunità scientifica, abbiamo sviluppato XXX – `acronimo cancellato per anonimizzazione` – un sito Web che ospita una versione pronta all'uso di questa risorsa multilingue, dei benchmark di riferimento per la traduzione automatica e degli strumenti software per la gestione e manipolazione dei suoi testi.[3]

Oltre che di per sé, il sito di XXX svolge un importante ruolo per IWSLT, il workshop internazionale per la traduzione del linguaggio parlato (International Workshop on Spoken Language Translation).[4] A partire dall'edizione 2012, infatti, i dati per l'addestramento, lo sviluppo e la valutazione di sistemi per la traduzione automatica dei discorsi TED, uno dei problemi proposti alla campagna di valutazione di IWSLT, vengono rilasciati attraverso il sito Web di XXX. Accanto alle risorse linguistiche e agli strumenti software, il sito rende disponibili anche le traduzioni automatiche generate da sistemi di base ed i relativi punteggi di due delle metriche più comuni in uso nella traduzione automatica (BLEU e TER); in questo modo vengono forniti non solo ai partecipanti ma alla comunità intera dei risultati di riferimento con cui validare le prestazioni dei propri sistemi.

In caso di esito positivo della procedura di revisione di questo sommario, la presentazione al convegno e la versione finale dell'articolo includeranno una descrizione dettagliata del corpus dei discorsi di TED, formato dei file e procedura per ottenere l'allineamento a livello di frasi compresi; verranno anche fornite delle statistiche sul corpus, con particolare riferimento all'italiano e ai dati paralleli che si possono ottenere tra l'italiano e tutte le altre lingue; verrà inoltre proposta un'analisi quantitativa della difficoltà di tradurre automaticamente i sottotitoli di TED. La relazione tra XXX e IWSLT sarà oggetto di una sezione specifica, che includerà una panoramica delle caratteristiche salienti dei sistemi di base che abbiamo sviluppato quali riferimenti per la campagna di valutazione di IWSLT 2012, con l'aggiunta di quello per la

*Nome cancellato per anonimizzazione.

[1] `http://langtech.jrc.ec.europa.eu/JRC-Acquis.html` (attivo al 22 ottobre 2012).

[2] `http://www.ted.com` (attivo al 22 ottobre 2012).

[3] `indirizzo Web` cancellato per anonimizzazione.

[4] `http://hltc.cs.ust.hk/iwslt` (attivo al 22 ottobre 2012).

traduzione tra l'inglese e l'italiano. La presentazione del sito Web di XXX concluderà sia la relazione al convegno sia l'articolo.

Riferimenti bibliografici

[1] Andreas Eisele and Yu Chen. MultiUN: A Multilingual Corpus from United Nation Documents. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

[2] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, pages 79–86, Phuket, Thailand, September 2005.

[3] Jörg Tiedemann. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing (vol V)*, pages 237–248. John Benjamins, Amsterdam/Philadelphia, 2009.

LA STANDARDIZZAZIONE DEL TRATTO DI LUNGHEZZA DELLE CONSONANTI AFFRICATE NELLE VARIETÀ DI ITALIANO CONTEMPORANEO

DONATELLA CARUCCI / RENATA SAVY

Una delle caratteristiche salienti dell'italiano standard e rara nelle lingue del mondo è il valore distintivo della lunghezza consonantica, analizzata sul piano fonetico-acustico attraverso il correlato della durata dei segmenti. Altra peculiarità dell'italiano standard è la presenza di quattro fonemi affricati nel proprio sistema fonologico, suoni di natura complessa e composita, generalmente poco presenti nelle lingue del mondo (Celata 2004). L'italiano abbina il tratto [+/- lungo] a quindici consonanti, di queste consonanti con lunghezza distintiva fanno parte le due affricate (pre-)palatali italiane, mentre le (alveo-)dentali vengono definite lunghe, non distintivamente, in taluni contesti (Mioni 1993, Savy, Crocco,Giordano 2005). La correlazione fonologica di lunghezza per le affricate (pre-)palatali trova quindi giustificazione sul piano fisico-fonetico, ma non altrettanto vale per le (alveo-)dentali, le cui misurazioni sono quindi meno rilevanti, in quanto, la loro realizzazione è sempre lunga in posizione intervocalica e le corrispondenti brevi si hanno solo in posizione iniziale assoluta e in posizione postconsonantica. Come è noto, tuttavia, rispetto allo standard, le descrizioni tradizionali dei dialetti e delle varietà regionali d'italiano (si vedano tra gli altri Rohlfs 1966, Muljačič 1972, Mioni 1993, Canepari 1999, Schmid 1999, Bertinetto e Loporcaro 2005, D'Achille 2006) concordemente affermano che alcune varietà del settentrione non mostrano l'opposizione di lunghezza per le consonanti, mentre in alcune varietà centrali e meridionali, una generale tendenza all'allungamento colpisce anche fonemi previsti scempi nello standard. Precedenti lavori sulla durata delle consonanti rafforzate italiane in contesto intervocalico hanno evidenziato che esse non costituiscono una classe omogenea di suoni (Endo & Bertinetto 1999, Celata 2004). La presente ricerca si propone di approfondire l'analisi del tratto di lunghezza abbinato alle affricate, con l'obiettivo specifico di verificare l'entità delle realizzazioni in quindici varietà di italiano colto semi-spontaneo per scoprire se ci sono delle regolarità per le quali è possibile parlare di processi di ristandardizzazione (Giordano, Savy 2012).

Metodologia

A tale scopo è stato analizzato un corpus di parlato dialogico elicitato, contenente 3651 segmenti affricati, etichettati a livello fonetico (ʧ=2105, tʃ=282, ʤ=400, dʒ=211, ʦ=204, ts=268, ʣ=181). Sono state misurate le durate medie dei foni in questione e per un parziale bilanciamento delle occorrenze, la catalogazione e la successiva analisi ha seguito una suddivisione di tipo distribuzionale, secondo cinque contesti:

1) V_V: posizione intervocalica, nella quale si possono osservare meglio tutte le questioni relative alla durata (1290);
2) W_W: posizione intervocalica tra parole, come la precedente, risulta essere significativa per l'analisi (1104);
3) RF: posizione eventualmente interessate da raddoppiamento fono sintattico (140);

4) C_: posizione postconsonantica, raggruppa gli eventuali fonemi scempi (325);
5) _W: posizione iniziale di parola, dopo pause o ad inizio di turno, sempre scempi (792);

Il limite del campione risiede comunque nella mancanza di bilanciamento tra le diverse occorrenze e in particolare nella totale indisponibilità del fonema (alveo-)dentale sonoro.

Risultati

Il primo dato significativo emerso dalle misurazioni sia delle affricate (pre-)palatali che delle (alveo-)dentali, riguarda tutte le varietà, dove le sonore sono realizzate sempre più brevi delle loro corrispettive sorde. Le affricate (pre-)palatali sorde mostrano, quanto a durata, una generale tendenza alla standardizzazione nelle varietà del Nord, quindi all'acquisizione dell'opposizione di lunghezza consonantica, tranne che per alcuni casi. La situazione risulta opposta per quel che riguarda le sonore, che ancora non sono in opposizione contrastiva nella maggior parte delle varietà; in questo caso le varietà centrali standardizzanti mostrano il giusto grado di opposizione, fatta eccezione per la varietà Romana, che continua a non seguire la standardizzazione, non distinguendo la sonora breve, da quella lunga, mentre altre varietà le distinguono solo marginalmente. Per quel che riguarda le affricate (alveo-)dentali, notiamo che in posizione intervocalica tutte le varietà hanno l'effettiva realizzazione lunga (rafforzata), anche se le misurazioni mostrano un'alta deviazione standard, da collegare a variabili quali contesto distribuzionale e diversa velocità di eloquio. Inoltre, nei contesti in cui è previsto il fonema scempio, ossia in posizione iniziale o postconsonantica, alcune varietà (es:Roma,Milano) presentano un extra-allungamento della sorda, ma in conclusione possiamo sostenere che il tratto di lunghezza intrinseca si realizza foneticamente in tutte le varietà con durate confrontabili.

Riferimenti bibliografici

Bertinetto P.M., Loporcaro M., The sound pattern of Standard Italian, as compared with the varieties spoken in Florence, Milan and Rome, «*Journal of the International Phonetics Association* », 2005, pp.131-151.

Canepari L., *Manuale di fonetica*, Zanichelli, Bologna, 1999.

Canepari L., *Il Dizionario di Pronuncia Italiana*, Zanichelli, Bologna, 1999.

Celata C., *Acquisizione e mutamento di categorie fonologiche – Le affricate in italiano*, Franco Angeli, Milano, 2004.

Celata C., Kaeppeli B., Affricazione e rafforzamento in italiano: alcuni dati sperimentali, *Quaderni del Laboratorio di Linguistica della Scuola Normale Superiore* 4, 2003, pp 43-59.

D'Achille P., *L'italiano contemporaneo*, Il Mulino,Bologna, 2006.

Endo, R. & Bertinetto, P.M., Caratteristiche prosodiche delle così dette rafforzate italiane, in: Delmonte, R. & Bristot, A. (curr.*), Aspetti computazionali in fonetica, linguistica e didattica delle lingue: modelli e algoritmi.* Atti delle None Giornate di Studio del Gruppo di Fonetica Sperimentale (A.I.A.), Roma, Esagrafica, 1999, pp. 243-255.

Giordano R., Savy R., Sulla standardizzazione del consonantismo italiano: consonanti geminate, rafforzate e fricative alveolari in contesto intervocalico, *Atti dell'XI Congresso Internazionale Silfi,* 2012, pp. 31-45.

Mioni A.M., *Fonetica e fonologia, in Introduzione all'italiano contemporaneo*, a c. di Sobrero A., Roma, Laterza, 1993.

Muljačič Z., *Fonologia della lingua italiana*, Il Mulino, Bologna, 1972.

Rohlfs G., *Historische Grammatik der Italienischen Sprache und ihrer Mundarten. Vol. 1. Lautlehre*. Bern, Francke, 1949, (ed. it. *Grammatica storica della lingua italiana e dei suoi dialetti. Vol. 1. Fonetica,)* Einaudi, Torino, 1966.

Savy R., Crocco C. & Giordano R., Geminate e geminazioni tra codifica fonologica e codifica fonetica: esempi dal corpus AVIP. In *Atti del VI Convegno Internazionale SILFI,* E. Burr, Firenze, 2005, pp. 179-197.

Schmid S., *Fonetica e fonologia dell'italiano,* Paravia, Torino, 1999.

Context-based Language Model Adaptation for Lecture Translation

Nick Ruiz, Marcello Federico

FBK - Fondazione Bruno Kessler
Povo (TN) Italy

1 Introduction

Generally, Statistical Machine Translation systems are trained on general-purpose corpora, such as legislative proceedings or newswire texts. For SMT systems to be useful in the real world, it is necessary that SMT systems are robust with respect to the form or genre of new, untranslated texts. In many cases, *domain adaptation* is applied by adapting the probabilistic models of a SMT system (e.g. translation and language models) to statistically represent an entire translation task. However, in other cases, such as lecture translation, each document or discourse can vary widely from one another and can even consist of topical changes that cannot be accurately accounted for in a birds' eye perspective. In such scenarios, it is preferable to employ *topic adaptation*, which seeks to adapt a discourse based on small contexts of information that neighbor a given sentence or utterance.

In this paper, we focus primarily on topic adaptation for language modeling to improve the fluency of translations, both through word choice and small reordering decisions. We present crosslingual topic adaptation methods which adapt a language model (LM) based on the topic distribution of an adaptation context during translation. We construct a topic model on trained a collection of bilingual documents to model both topic and unigram distributions which are later used to adapt general purpose LMs on the fly, given only source language texts. In particular, we explore adaptation techniques based on the theory of Minimum Discrimination Information (MDI) [1]. Since MDI adaptation cannot be computed in real-time for scenarios such as lecture translation, we present a lazy log-linear approximation that can be efficiently computed during translation decoding.

2 Topic Adaptation using MDI

2.1 MDI-based Adaptation

MDI adaptation was proposed in [2] as a technique to adapt LMs based on small bag-of-word features drawn from an adaptation text. MDI adaptation scales the probabilities of a background LM, $P_B(w \mid h)$, on word w with n-gram history h by a ratio of unigram statistics observed in an adaptation text A against those observed in the background corpus B:

$$\alpha(w) = \left(\frac{\hat{P}_A(w)}{P_B(w)}\right)^{\gamma}, \quad 0 < \gamma \leq 1. \tag{1}$$

As such, the adapted LM probablities $P_A(w \mid h)$ are constructed and normalized as follows:

$$P_A(w \mid h) = \frac{P_B(w \mid h) \cdot \alpha(w)}{\sum_{w'} P_B(w' \mid h_j) \cdot \alpha(w')}. \tag{2}$$

There are two general setbacks to using MDI adaptation for LMs in SMT. First, the unigram statistics from small adaptation contexts are not reliable enough to accurately reestimate all entries within a language model. Secondly, LMs model the target language in SMT; thus the adaptation features cannot be computed directly from a source text.

2.2 Inferring unigrams via bilingual topic modeling

Topic modeling provides a robust solution to infer unigram distributions from small documents. One general topic modeling approach is Probabilistic Latent Semantic Analysis (PLSA) [3], which computes the probability of unigrams in a document d by marginalizing over a collection of latent topics T:

$$P(w \mid d) = \sum_{t \in T} P(w \mid t) P(t \mid d). \tag{3}$$

Topic modeling approaches decompose the problem of assigning probability to words in a document by modeling the probability of a topic occurring in a document, $p(t \mid d)$ and the probability that a word exists within that topic, $p(w \mid t)$. With a model that learns $p(w \mid t)$, the unigram features in a bilingual model can be reconstructed from a small adaptation text by computing its topic distribution.

Various bilingual topic modeling approaches have been proposed (e.g. [4] [5]) to infer target-language unigram features from source-language texts using bilingual corpora. Rather than constructing complex graphical model structures to accommodate bilingual topic modeling, we treat the problem of bilingual topic modeling as an extension of classic monolingual topic modeling. As such, we transform the problem of bilingual topic modeling by combining source and target parallel sentences into "monolingual" documents with vocabulary $V_{FE} = V_F \cup V_E$. During topic model inference, we infer unigram probabilities of V_{FE} using only documents containing only the source language, which is possible because the source language provides enough tokens to determine the topic distribution of a document. Removing words $f \in V_F$ from the probability distribution and normalizing yields a probability distribution for all words in V_E.

2.3 Drawbacks

While bilingual topic modeling resolves the problem of insufficient target language unigram statistics for MDI adaptation, MDI adaptation requires all of the n-gram probabilities in a LM to be restimated. Since state-of-the-art LMs employ back-off and interpolation, a full reestimation which requires probabilistic normalization is computationally infeasible in scenarios such as continuous speech translation that seek to adapt n-gram counts based on a sliding context window in real-time.

3 Lazy MDI Alternative for SMT

We exploit general properties of MDI adaptation to provide a fast alternative. The goal of MDI adaptation is to construct an adapted LM that minimizes its Kullback-Leibler divergence from the background LM, which is performed by unigram ratio scaling as described in (1) and (2). We loosely approximate this KL divergence in statistical machine translation by adapting only n-grams that appear as translation options for a given sentence without computing a normalization term that requires observing the probabilities of all high- and lower-order n-grams in the LM.

However, unbounded ratios have unpredictable effects on n-gram probabilities, so in place of normalization, we apply a smoothing function on the unigram ratio to constrain the effects of large differences in unigram observations in our adaptation context. We apply transformations to a *fast sigmoid* approximation that was originally proposed in [6]:

$$f(x, a) = \frac{ax}{a + \|x\| - 1}, a > 1, \tag{4}$$

which has the following properties:

$$f(0) = 0; \lim_{x \to +\infty} f(x) = a$$
$$f(1) = 1; \lim_{x \to -\infty} f(x) = -a.$$

In particular the $f(1) = 1$ constraint ensures that background LM probabilities remain fixed when the ratio is balanced.

Since we are no longer normalizing n-gram probabilities, we can consider the smoothed unigram probabilities as a function that rewards or penalizes translation options based on the likelihood that the words composing the target phrase should appear in the translation. The smoothed unigram probability ratio is added as a new feature in the discriminative log-linear model of the SMT decoder. While our new feature is independent from any LM features, we can logically consider the adaptation of a background LM as a log-linear combination of the LM feature and the Lazy MDI feature. By rearranging terms, our unnormalized log-linear approximation of (2) is:

$$\hat{P}_A(w \mid h) = P_B(w \mid h)^{\gamma_1} \cdot \hat{\alpha}(w)^{\gamma_2}; \quad \hat{\alpha}(w) = f\left(\frac{P_A(w)}{P_B(w)}\right) \tag{5}$$

Since only the translation hypotheses suggested by the translation model are scored by the LM, only a subset of unigram ratios are considered during adaptation.

4 Experiments

4.1 Lowercased Evaluation

We compare classic MDI against Lazy MDI for LM adaptation on 5-line contexts using a PLSA model with 250 topics, using the data set of English-French translations of TED talks according to the IWSLT 2010[1] evaluation. The TED training transcripts consist of approximately 84k sentences and the test set consists of 2.4k sentences.

Lowercased SMT systems are trained from the TED corpus using the Moses SMT toolkit [7]. One 5-gram background LM was constructed with the IRSTLM toolkit [8] on the French training data (with improved Kneser-Ney smoothing) [9]. The weights of the log-linear model were optimized via minimum error rate training (MERT) [10] on the TED development set, using 200 best translations at each tuning iteration.

We ran 3 MERT instances for each system and evaluated using MultiEval 0.3 [11]. Evaluation results in terms of BLEU, METEOR (French), TER, and segment length are listed in Table 1. We observe similar results between MDI and smoothed Lazy MDI.

System	BLEU	$\bar{s}_{\text{sel}}$	s_{Test}	p-value
Baseline	28.0	0.5	0.3	-
MDI	28.2	0.5	0.2	0.01
Lazy MDI	28.3	0.5	0.1	0.00

Table 1. Lowercased evaluation of MDI and Lazy MDI adaptation techniques on the IWSLT 2010 TED test set, averaged across three MERT runs with p-values relative to the baseline. $\bar{s}_{sel}$ indicates the variance due to test set selection. Significant improvements are observed for both MDI and Lazy MDI.

5 Conclusion

We have outlined language modeling techniques suited for topic adaptation on small contexts of lecture transcripts, using the premises of Minimum Discrimination Information and topic modeling. We explore the utility of bilingual topic modeling with MDI and overcome the time restrictions of full language model reestimation by approximating MDI adaptation through a log-linear feature function that rewards or penalizes unigrams based on smoothed unigram ratios between an adaptation context and the background LM. Our Lazy MDI adaptation approach performs comparably to the classic MDI adaptation scenario, but has the advantage of faster performance due to the loose coupling of smoothed unigram ratios and the background LM.

[1] http://iwslt2010.fbk.eu/

References

1. S. A. Della Pietra, V. J. Della Pietra, R. Mercer, and S. Roukos, "Adaptive language model estimation using minimum discrimination estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. I, San Francisco, CA, 1992, pp. 633–636.
2. M. Federico, "Efficient language model adaptation through MDI estimation," in *Proceedings of the 6th European Conference on Speech Communication and Technology*, vol. 4, Budapest, Hungary, 1999, pp. 1583–1586.
3. T. Hofmann, "Probabilistic Latent Semantic Analysis," in *Proceedings of the 15th Conference on Uncertainty in AI*, Stockholm, Sweden, 1999, pp. 289–296.
4. Y.-C. Tam, I. Lane, and T. Schultz, "Bilingual LSA-based adaptation for statistical machine translation," *Machine Translation*, vol. 21, pp. 187–207, December 2007. [Online]. Available: http://portal.acm.org/citation.cfm?id=1466799.1466803
5. D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum, "Polylingual Topic Models," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, August 2009.
6. G. M. Georgiou, "Parallel distributed processing in the complex domain," Ph.D. dissertation, Tulane University, New Orleans, LA, USA, 1992, uMI Order No. GAX92-29796.
7. P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic, 2007, pp. 177–180. [Online]. Available: http://aclweb.org/anthology-new/P/P07/P07-2045.pdf
8. M. Federico, N. Bertoldi, and M. Cettolo, "IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models," in *Proceedings of Interspeech*, Melbourne, Australia, 2008, pp. 1618–1621.
9. S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech and Language*, vol. 4, no. 13, pp. 359–393, 1999.
10. F. J. Och, "Minimum Error Rate Training in Statistical Machine Translation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, E. Hinrichs and D. Roth, Eds., 2003, pp. 160–167. [Online]. Available: http://www.aclweb.org/anthology/P03-1021.pdf
11. J. Clark, C. Dyer, A. Lavie, and N. Smith, "Better hypothesis testing for statistical machine translation: Controlling for optimizer instability," in *Proceedings of the Association for Computational Lingustics*, ser. ACL 2011. Portland, Oregon, USA: Association for Computational Linguistics, 2011, available at http://www.cs.cmu.edu/ jhclark/pubs/significance.pdf.

The psychological reality of the metrical structure: the role of heads and boundaries in sentence comprehension

Giuliano Bocci & Cinzia Avesani

1. Introduction. Italian is a SVO language, but it massively allows phrasal displacement and the word order is quite flexible. Subjects (Ss) can appear either preverbally or postverbally (provided the appropriate morpho-syntactic and semantic conditions), and Objects (Os) can be fronted with or without a resumptive clitic, as in case of topicalization or focus fronting, respectively. It has been proposed in the literature (a.o. Szendröi 2001, but also Vallduví 2002) that Italian has a rigid prosodic template, since *Rightmostness* (of the prosodic heads) must fulfilled at every level of the prosodic hierarchy above the word level and this would account for focus-related word order alternations. Right-Dislocation would be exploited to align focus with the main prominence: all the post-focal constituents, being Given and part of the background are right-dislocated and prosodically non-prominent. This line of analysis, however, appears problematic in light of the experimental research. Unlike what observed in Germanic languages, it has been empirically shown (see Grice et al. 2005) that post-focal constituents in many varieties of Italian associate with compressed pitch accents: they cannot be analyzed as extra-prosodic. In this paper, we provide an analysis of the metrical structure of Italian based on a production and a comprehension experiment. We show that *Rightmostness* is violated in case of non-final focus, since post-focal constituents are not extra-prosodic neither non-prominent, being phrased and endowed with phrasal metrical prominence. We show that the distribution of phrasal heads and boundaries in the post-focal context plays a crucial role in driving syntactic parsing and allowing speakers to disambiguate ambiguous word orders.

2. The starting point: a production experiment. In order to ascertain the alleged inviolability of Rightmostness in Italian, in a previous study we addressed the issue of the metrical representation of post-focal constituents in Tuscan Italian by means of a production experiment on read speech (10 speakers, 436 utterances). The stimuli, exemplified in Table 1, presented an infinitival verb form (the target word) occurring in three conditions: A] in a broad focus sentence (BF); P] following a contrastively focused (CF) S; H] following a CF S and preceding a RDed topic. The rationale was the following (see Table 2). In A], the infinitive was expected not to qualify as a phrasal head, since the head should be assigned to the O. Similarly, the infinitive in P], being followed by its O, should not qualify as a phrasal head, regardless of the metrical status of post-focal material. In H], instead, we expected the RDed O to be phrased into an independent intonational phrase (ι). The infinitive, thus, was supposed to be wrapped between the φ-boundary closing the initial focus and the ι-boundary setting apart the RDed object. If phrasal prominences were assigned in post-focal context by virtue of default mapping rules, the infinitive in H] should qualify as φ-head, being the rightmost element within its φ-phrase. Conversely, if post-focal elements were extra-prosodic, it should bear only a word-level prominence as in A] and P]. The results clearly showed that metrical phrasal heads are assigned to post-focal material in H]. The infinitive (though Given and part of the background) bore a higher degree of metrical prominence than in A] (i.e. non-Given but in a structurally weak position) and P] (i.e. Given, and in a weak position): the infinitive's stressed vowel in H], for instance, was characterized by significantly longer durations than in A] and P], more extreme formant trajectories and higher spectral emphasis. Our findings, thus, led us to conclude that Rightmostness in H] is violated (at least) at the ι-level and to reject the existence of a rigid prosodic template in Italian.

3. A comprehension experiment. According to our analysis of Italian prosody, all elements undergo phrasing and every prosodic constituent must be headed. In the present contribution, we further confirm the validity of our conclusions and support the psychological reality of the proposed prosodic structure by means of a comprehension experiment on manipulated stimuli, based on the assumption that RDed topics mandatorily call for an ι-boundary separating them from the verb and thus the insertion of the head on the preceding element. If this is the case, we expect that by deleting the clitic from a sentence in H], we obtain a sentence in which the prosodic cues still signal the second DP (DP2) as a RDed topic, but the lack of the object clitic (whose occurrence is mandatory with RDed Os) prevents DP2 from being interpreted as RDed O. The resulting sentence is hence expected to be interpreted as an inverted structure $O_{CF}VS_{RDed}$, which is the only interpretation compatible both with the prosodic and the morpho-syntactic properties (since RDed Ss do not involve any overt resumptive pronoun). Furthermore, we predict that this sentence – in which the clitic is deleted – could be turned into a $S_{CF}VO$ sentence with O *in situ* (i.e. like the sentences in P]), if we further

manipulate the verb by shortening the duration of the stressed syllable ('σ) and of the final syllable (σ#), i.e. deleting the durational correlates of the φ-head and ι-boundary). Analogously, we predict that if we take a $S_{CF}VO$ sentence produced in P], and lengthen the duration of 'σ of the infinitive (i.e. adding the φ-head) and of its σ# (i.e. adding the ι-boundary), we obtain a sentence in which DP2 is endowed with the prosody of a RDed topic. Since no clitic doubles DP2, we expect the obtained sentence to be interpreted as $O_{CF}VS_{RDed}$.

We tested these hypotheses by means of a forced-choice comprehension experiment in which 12 native speakers of Italian were asked to identify the Subject in 64 experimental sentences obtained by manipulating the sentences (transitive and semantically reversible) of our production experiment: 16x2 from H] + 16x2 from P]. We presented twice 16 sentences originally produced in H]. In one case, we simply deleted the object clitic and the expected interpretation was $O_{CF}VS_{RDed}$, while in the other we additionally shortened both 'σ and σ# in order to restore a $S_{CF}VO_{in\ situ}$ interpretation. Moreover, we presented twice 16 sentences produced in P]. In one case, the sentences were not manipulated at all, while in the other we manipulated the verb by increasing the duration of both 'σ and σ#, so as to induce a RDed interpretation of DP2 and thus a $O_{CF}VS_{RDed}$ interpretation. The durations were manipulated with Praat by applying the (phone-based) coefficients calculated after the production experiment, while pitch contours were not manipulated. The results (see Figure 1) fully support all our hypotheses: a mixed logit model showed that the interpretation of DP2=Subject highly significantly correlates only with the factor "prosodic properties of the verb", while the factor "condition in production" (i.e. originally P] vs. H]) was not significant, nor was their interaction.

4. Our findings show the central role of the prosodic phonology, which mediates between the phonetic realization of an utterance and its abstract syntactic representation: small duration differences in relevant positions lead to a specific metrical representation and this, in turn, leads to a specific syntactic representation. The provided evidence show that *Rightmostness* is violable in Italian and thus cannot account for focus-related word-order alternations. We argue that Italian prosody is rigid only in the sense that it fails to destress Given information, and phrasing and headedness must apply exhaustively: the occurrence of post-focal phrasal prominences in H] is not due to specific discourse-properties (i.e. second occurrence of focus), but only to default mapping rules. We discuss the consequences of this analysis for a model of prosody, and its interface with syntax and information structure.

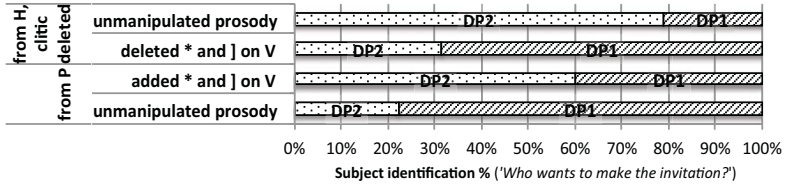
Table 1: Experimental sentences.

Table 1.	DP1 (=proper name)	Vfinite	Infinitive	DP2 (=proper name)
A	[Germanico	vorrebbe	**invitare**	Pierangela]$_{BF}$
	Germanico	*would like to*	*invite*	*Pierangela*
P	[Germanico]$_{CF}$	vorrebbe	**invitare**	Pierangela
	Germanico	*would like to*	*invite*	*Pierangela*
H	[Germanico]$_{CF}$ la	vorrebbe	**invitare**	[Pierangela]$_{RDed}$
	Germanico	*her$_{clitic}$ would like to*	*invite*	*Pierangela*

Table 2 Metrical structures.

Table 2.	Metrical structures
A BF	{ * }$_{\upsilon}$ [*]$_{\iota}$ (*)$_{\varphi}$ (*)$_{\varphi}$ Germanico vorrebbe **invitare** Pierangela
P initial CF	{ * }$_{\upsilon}$ [*]$_{\iota}$ (*)$_{\varphi}$ (*)$_{\varphi}$ [Germanico]$_{CF}$ vorrebbe **invitare** Pierangela
H initial CF, O RDed	{ * }$_{\upsilon}$ [*]$_{\iota}$ [*]$_{\iota}$ (*)$_{\varphi}$ (*)$_{\varphi}$ (*)$_{\varphi}$ [Germanico]$_{CF}$ la vorrebbe **invitare** [Pierangela]$_{RDed}$

Figure 1. Results of the comprehension experiment.



Grice, M., D'Imperio, M., Savino, M. and Avesani, C. 2006. Towards a strategy for labelling varieties of Italian. In: S.-A. Jun (ed.), *Prosodic Typology*. Oxford: Oxford University Press.

Szendröi, K. 2001. *Focus and the syntax-phonology interface*. Ph.D. Dissertation, UCL.

Vallduví, E. 1992. *Informational Component*. New York: Garland.

La percezione vocalica in bambini italiani ipoacusici con impianto cocleare: studio comportamentale ed elettrofisiologico

L. Garrapa[1,2], D. Bottari[3], M. Grimaldi[1], A. Calabrese[1,4], F. Pavani[5], M. De Benedetto[6], S. Vitale[6], P. Monastero[6], M. Greco[6]

[1] CRIL, Università del Salento, [2] Università di Padova,
[3] Universität Hamburg (Germania), [4] University of Connecticut (USA),
[5] CIMeC e DiSCoF, Università di Trento, [6] ORL, ASL/LE Ospedale "Fazzi", Lecce

luigia.garrapa@unisalento, davide.bottari@uni-hamburg.de, mirko.grimaldi@unisalento.it, andrea.calabrese@uconn.edu, francesco.pavani@unitn.it, micheledebenedetto@hotmail.it, vitasi@libero.it

Introduzione

L'impianto cocleare (IC) permette ai soggetti ipoacusici di accedere alla lingua orale. La discriminazione di contrasti linguistici nei bambini con IC è stata indagata (prevalentemente per le consonanti), mediante registrazioni elettrofisiologiche (preattentive). Questa metodica è stata combinata con test comportamentali (attentivi) in bambini inglesi, olandesi e finlandesi, ma, a nostra conoscenza, mai in bambini italiani.

Per valutare se il sistema uditivo dei bambini con IC discrimini suoni linguistici a livello preattentivo, precedenti studi hanno monitorato le componenti P1, N2 e *Mismatch Negativity* (MMN) dei Potenziali Evocati Evento-Correlati (ERPs). P1 è un correlato della detezione dell'inizio del suono; N2 è un correlato del contenuto del suono la cui ampiezza è massima per suoni linguistici (Sussman et al. 2008). MMN è un correlato della capacità di rilevare una devianza fra due suoni, indica l'accuratezza con cui essi vengono discriminati ed è adatta a studiare la rappresentazione astratta e l'elaborazione dei suoni linguistici (Naatanen et al. 2007). Singh et al. (2004), Sharma et al. (2005) e Gilley et al. (2008), hanno rilevato che i bambini con IC presentano P1, ma non sempre N2 e MMN, e che P1, N2 e MMN evocate da suoni linguistici in bambini con IC, soprattutto se impiantati tardivamente (> 3.5), hanno maggiore latenza e minore ampiezza rispetto ai bambini normo-udenti (NH). Quindi, i processi cognitivi di detezione del suono, identificazione del suo contenuto e discriminazione fra due suoni possono essere ritardati e meno accurati nei bambini con IC. Tuttavia, studi recenti, incrociando metodi comportamentali ed ERPs, hanno evidenziato che i bambini con IC a volte discriminano contrasti linguistici solo a livello preattentivo (Beynon et al. 2002) e altre solo a livello attentivo (Henkin et al. 2008).

Obiettivi

Questo lavoro si propone di chiarire se la latenza, l'ampiezza e la distribuzione sullo scalpo di P1, N2 e MMN dei bambini con IC differiscano rispetto a un gruppo di bambini NH. Inoltre ci proponiamo di appurare se, nei bambini con IC, i livelli preattentivo e attentivo di discriminazione si sviluppino parallelamente o meno.

Soggetti, materiali e metodi

8 bambini NH (età media = 6.9) e 11 bambini con IC (età media = 7.5; uso medio dell'IC = 4.5) residenti in provincia di Lecce hanno partecipato allo studio.

Le vocali analizzate sono /i/ e /u/, due vocali alte, realizzate con la radice della lingua in posizione avanzata, che differiscono per il luogo di articolazione e l'arrotondamento delle labbra.

A livello comportamentale, i bambini hanno identificato /i/ e /u/ presentate in isolamento (20 vocali in tutto) e hanno discriminato i contrasti /i/-/i/, /u/-/u/, /i/-/u/ e /u/-/i/ (40 coppie in tutto). Gli esemplari di /i/ e /u/ sono stimoli di parlato naturale (Eulitz & Lahiri 2004), in modo da introdurre variazione acustica naturale negli stimoli (Phillips et al. 2000), e sono stati prodotti in cabina silente da un giovane uomo e normalizzati per durata, F1, F2, intensità, volume e *rise/fall times* per renderli omogenei (Näätänen et al. 2001).

A livello preattentivo, la discriminazione di /u/-/i/ e /i/-/u/ (distanza acustica = 847 Mel) è stata indagata registrando gli ERPs evocati con un protocollo *oddball* articolato in 2 blocchi ($/u/_{std}$-$/i/_{dev}$ e $/i/_{std}$-$/u/_{dev}$) aventi 680 stimoli standard e 120 devianti. I parametri ERPs sono stati computati sugli elettrodi fronto-centrali. P1 e N2 sono state calcolate sull'ERP di $/i/_{std}$ e $/i/_{dev}$; MMN, invece, è stata

calcolata sottraendo l'ERP di $/i/_{std}$ dall'ERP di $/i/_{dev}$, per ottenere una MMN "pura", elicitata dal cambiamento di "ruolo" ricoperto da /i/ (Näätänen et al. 2007).

Risultati

I dati comportamentali si riferiscono ad entrambi i gruppi oggetto di indagine. Sia i bambini con IC che i bambini NH identificano correttamente /i/ e /u/ con una percentuale comparabile: /i/, 85% per i bambini con IC vs. 100% per quelli NH ($t(10)=1.638$, $p=.132$); nel caso di /u/, 95% vs. 99% ($t(15)=1.361$, $p=.193$). Parallelamente, i due gruppi discriminano correttamente con una percentuale comparabile le coppie /i/-/i/ (97% vs. 99%; $t(17)=.105$, $p=.918$), /u/-/u/ (98% vs. 100%; $t(17)=.846$, $p=.409$), /i/-/u/ (95%; $t(17)=.137$, $p=.893$) e /u/-/i/ (99% vs. 100%; $t(17)=.846$, $p=.409$).

I dati ERPs sono relativi a 4 bambini NH e 5 con IC (le sedute ERPs sono ancora in corso). Per quanto riguarda P1, l'ampiezza e la latenza per $/i/_{std}$ non differiscono fra i due gruppi ($t(79)=.098$, $p=.922$ per ampiezza; $t(61)=.764$, $p=.448$ per latenza). L'ampiezza per $/i/_{dev}$ è maggiore nei bambini NH ($t(79)=2.181$, $p<.05$), mentre la latenza è comparabile nei due gruppi ($t(63)=.993$, $p=.324$), cf. Figg. 1b e 2a.

Per quanto concerne N2, l'ampiezza per $/i/_{std}$ e $/i/_{dev}$ è maggiore nei bambini NH ($t(79)=3.241$, $p<.001$ per $/i/_{std}$; $t(79)=2.544$, $p<.01$ per $/i/_{dev}$), ma la latenza appare minore nei bambini con IC ($t(79)=3.096$, $p<.0025$ per $/i/_{std}$; $t(79)=6.175$, $p<.001$ per $/i/_{dev}$), cf. Figg. 1b e 2a.

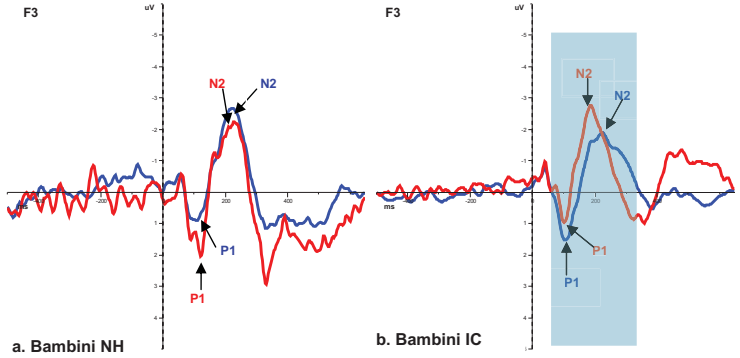


Fig. 1: Grand Average dell'ERP di $/i/_{std}$ (blu) e di $/i/_{dev}$ (rosso) in F3 nei due gruppi (dati filtrati a 0.3-20Hz ai fini grafici).

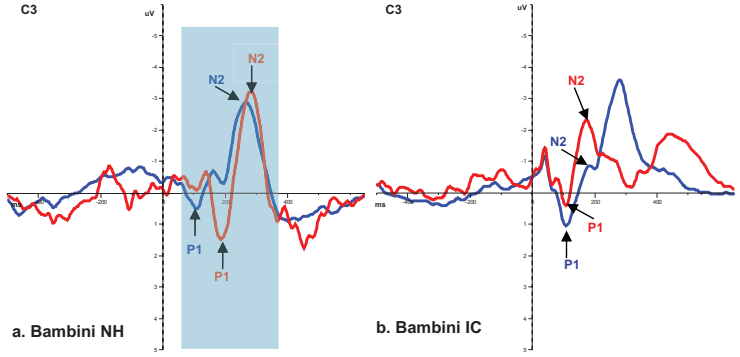


Fig. 2: Grand Average dell'ERP di $/i/_{std}$ (blu) e di $/i/_{dev}$ (rosso) in C3 nei due gruppi (dati filtrati a 0.3-20Hz ai fini grafici).

MMN è presente in entrambi i gruppi, ma è più robusta nelle regioni frontali (F3, Fz, F4) dei bambini con IC ($F(2)=.715$, $p=.495$) e nelle regioni centrali (C3, Cz, C4) dei bambini NH ($F(2)=.269$, $p=.766$). Essa non è lateralizzata né nei bambini con IC ($F(2)=.582$, $p=.563$) né in quelli NH ($F(2)=.661$, $p=.523$). L'ampiezza di MMN è comparabile nei 2 gruppi ($t(79)=.974$, $p=.21$), mentre la latenza appare minore nei bambini con IC ($t(54)=4.719$, $p< .001$), cf. Fig. 3.

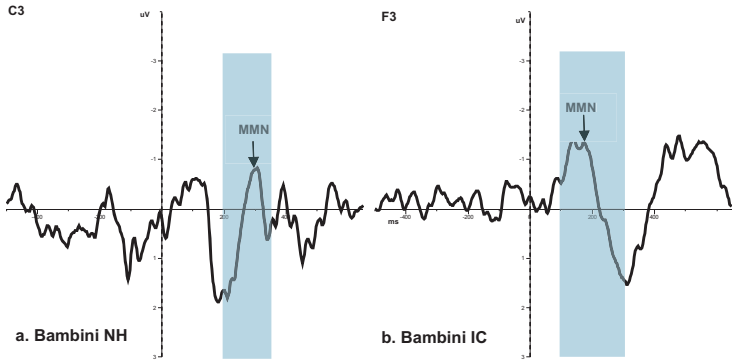


Fig. 3: MMN nella *difference wave* dei due gruppi (dati filtrati a 0.3-20Hz ai fini grafici).

Conclusioni

I risultati, in accordo con Henkin et al. (2008), dimostrano che a livello attentivo i bambini italiani con IC identificano e discriminano /i/ e /u/, che differiscono per il luogo di articolazione e l'arrotondamento delle labbra, con un'accuratezza comparabile a quella dei bambini NH.

Tuttavia, a livello preattentivo, i processi di identificazione e discriminazione sembrano essere meno accurati nei bambini con IC rispetto ai bambini NH. Infatti, in linea con Singh et al. (2004), Sharma et al. (2005), Gilley et al. (2008), e Torppa et al. (2012), i nostri dati evidenziano che, nei bambini con IC, P1 e N2 hanno una minore ampiezza rispetto ai bambini NH. Questo dato suggerisce delle differenze tra i due gruppi nelle prime fasi di elaborazione dei suoni linguistici da parte del sistema uditivo.

Contrariamente a quanto riscontrato in letteratura, la latenza di P1 appare comparabile nei due gruppi, mentre quella di N2 appare minore nei bambini con IC. Inoltre, MMN sembra avere un'ampiezza comparabile nei due gruppi e una latenza minore nei bambini con IC. Questi risultati possono essere imputati a diversi fattori. In particolare: 1) al fatto che i soggetti con IC che hanno preso parte a questo studio avevano già ricevuto da tempo l'IC e seguito un congruo periodo di rieducazione, maturando quindi tracce mnestiche robuste dei due fonemi; 2) al fatto che i dati ERPs sono ancora parziali, per cui il calcolo della latenza e dell'ampiezza di alcune componenti è ancora incompleto.

L'analisi completa dei dati ERPs stabilirà se, nonostante il periodo di sordità iniziale, il sistema uditivo dei bambini italiani con IC elabora e processa le differenze tra le vocali con una latenza e un'ampiezza comparabili a quelle dei bambini NH.

Bibliografia

Beynon et al. (2002): Evaluation of cochlear implant benefit with cortical auditory evoked potentials. *IJA* 41: 429-35.
Eulitz & Lahiri (2004): Neurobiological evidence for abstract phonological representations in the mental lexicon during speech recognition. *JCN* 16,4: 577-83.
Gilley et al. (2008): Cortical reorganization in children with cochlear implants. *BR* 1239: 56-65.
Henkin et al. (2008): Phonetic processing in children with cochlear implants: an auditory ERP study. *ER* 29,2: 239-49.

Näätäanen et al. (2001): The perception of speech sounds by the human brain as reflected by the Mismatch negativity (MMN) and its magnetic equivalent (mMMN). *Psychophysiology* 38:1-21.
Näätäanen et al.(2007):The Mismatch Negativity (MMN) in basic research of central auditory processing: a review. *CL* 118: 2544-90.
Phillips et al. (2000): Auditory cortex accesses phonological categories: An MEG mismatch study. *JCN* 12:1038-1105.
Sharma et al. (2005): P1 Latency as a biomarker for central auditory development in children with hearing impairment. *JAAA* 16: 564-73.
Singh et al. (2004):Event-related potentials (ERPs) in pediatric cochlear-implant patients. *EH* 25,6: 598-610.
Sussman et al. (2008): The maturation of human event-related potentials to sounds presented at different stimulus rates. *HR* 236: 61-79.
Torppa et al. (2012): Cortical processing of musical sounds in children with cochlear implants. *CN* 123,10:1966-79.

Studio acustico ed articolatorio di sequenze di sibilanti nella lingua francese

Sonia d'Apolito, Barbara Gili Fivela, Francesco Sigona
Centro di Ricerca Interdisciplinare sul Linguaggio – Università del Salento

Lo studio riguarda la realizzazione di sequenze eterosillabiche di sibilanti in francese da parte di un parlante nativo e di un apprendente italofono con un'elevata competenza in francese. In particolare, si osserva, dal punto di vista acustico e cinematico, come la realizzazione di eventuali processi fonologici, quali cancellazione e/o assimilazione di luogo, possa essere influenzata dalla velocità di eloquio e dalla presenza di un confine prosodico. Si tratta di due fattori importanti, poiché una velocità di eloquio elevata facilita la coarticolazione [3], e quindi le assimilazioni, e la presenza di un confine prosodico può invece interferire con il processo riducendo la sovrapposizione tra i gesti [4].

In francese, nessi di sibilanti sono stati osservati in modo sistematico in uno studio acustico sull'assimilazione all'interno di nessi al confine di parola a velocità normale [11]. I risultati mostrano l'effettiva presenza di assimilazioni di luogo in francese, benché in letteratura questo fenomeno non sia attestato [14]. In italiano, le sequenze di sibilanti non ci risulta siano state oggetto di studi specifici, probabilmente per il fatto che sono poco frequenti soprattutto al confine di parola. A fine parola /s/, /z/, /ʃ/ e /ʒ/ si trovano solo in prestiti [7;10], benché alcuni siano ormai di uso comune. Combinando questi prestiti con parole della lingua italiana che inizino con sibilante, è possibile ottenere un nesso di fricative al confine di parola, benché la realizzazione del nesso appaia comunque abbastanza difficoltosa e sia qui considerata come non appartenente alla competenza del parlante nativo di italiano, o come una sequenza fonotattica per lui molto marcata.

L'obiettivo di questo lavoro è osservare: 1) come vengono realizzati i nessi di sibilanti al variare dello stile di eloquio e della condizione prosodica, ossia se una maggiore velocità di eloquio e la presenza di confine prosodico possano interferire con la realizzazione di processi fonologici; 2) come l'apprendente italofono realizzi questi nessi e se si differenzi dal parlante francese nativo.

Diverse sequenze di sibilanti sono state osservate al confine di parola, proposte all'interno di una frase cornice nel contesto vocalico /a_i/. Il corpus è stato letto sia a velocità normale che sostenuta; inoltre, i contesti sono stati inseriti all'interno di due differenti condizioni prosodiche: le consonanti erano parte dello stesso sintagma intonativo oppure si trovavano in due sintagmi intonativi diversi. Un italofono con buona competenza in francese L2 (PI1) ed un parlante francofono (PF4) hanno letto le frasi per 7 volte. I materiali acustici e articolatori sono stati etichettati e misurati in PRAAT e Matlab, ed analizzati statisticamente con test non parametrici.

L'etichettatura acustica ha riguardato i segmenti della sequenza $V_1C_1\#C_2V_2$, inclusi un possibile schwa e/o pausa. Le misurazioni acustiche effettuate sono: picco di frequenza; quattro momenti spettrali; F0 vocali adiacenti; durata dei segmenti e durata normalizzata; due linee di regressione; ampiezza rms normalizzata. Per i dati cinematici, sono state osservate le traiettorie dei seguenti articolatori, sull'asse orizzontale (x) e verticale (z): punta della lingua (TT), poiché entrambe le consonanti sono coronali; labbro inferiore (solo asse x, LLx) per la protrusione della postalveolare; e dorso della lingua (TD) per il passaggio vocalico /a__i/. L'etichettatura cinematica ha previsto l'individuazione degli eventi articolatori corrispondenti all'apertura, alla chiusura e ai

picchi di velocità per le due fricative e vocali adiacenti, osservando le traiettorie di posizione e di velocità. Le misurazioni cinematiche effettuate sono: durata (ms) ed ampiezza (mm) del gesto di chiusura della fricativa in posizione C_1 e C_2; differenza temporale (ms) e di ampiezza (mm) tra il raggiungimento del target per le due fricative; fase relativa dei target per le due fricative, rispetto alla durata normalizzata del passaggio vocalico /a__i/, calcolata su TDz [13]; C-center rispetto al target della vocale /i/ su TDz [2], poiché dà indicazioni sulla struttura sillabica e ci permette di osservare se un segmento appartiene all'offset di una sillaba (V_1C_1) o all'onset di quella successiva (C_2V_2) e di capire se l'assimilazione si è accompagnata a eventuale risillabificazione.
I risultati acustici mostrano che, a velocità normale, PI1 e PF4 inseriscono uno schwa e, in caso di confine, anche una pausa; solo PF4 inserisce una pausa senza schwa davanti a fricativa sorda. Nessun parlante realizza assimilazioni. A velocità sostenuta, PI1 e PF4 inseriscono uno schwa solo in presenza di confine. Per la sequenza alveolare-postalveolare, PF4 non realizza l'alveolare in assenza di confine prosodico e, generalmente, davanti alla fricativa sorda in presenza di confine. PI1, invece, realizza sempre il gesto alveolare e inserisce uno schwa in pochi casi. Per la sequenza postalveolare-alveolare, entrambi realizzano assimilazioni progressive di luogo; PI1, solo in assenza di confine, PF4 anche in presenza di confine, ma solo davanti a /ʃ/ senza schwa. In generale, i casi di assimilazione progressiva di luogo si presentano come un lungo segmento postalveolare sordo, per cui /ʒ/ è sempre desonorizzato. Le misurazioni acustiche più robuste, quali quelle relative al picco di frequenza, CoG, skewness, kurtosi e linee di regressione, distinguono le assimilazioni progressive di luogo dall'alveolare, poiché mostrano caratteristiche più simili a quelle della postalveolare.

Dal punto di vista articolatorio, le realizzazioni a velocità normale presentano un *pattern* molto stabile poiché l'inserimento di schwa e/o pausa permette di identificare un'apertura intermedia tra le due fricative. A velocità sostenuta, invece, si ha una maggiore coarticolazione, per cui molto spesso sulle traiettorie si osserva un solo gesto la cui natura dipende dall'ordine nel quale compaiono le fricative. La maggiore coarticolazione è data dalla minore durata del gesto di chiusura, dell'intervallo tra picchi e della fase relativa. Inoltre, le assimilazioni progressive di luogo mostrano un intervallo tra picchi e una fase relativa minore rispetto alle altre realizzazioni e, anche nei casi in cui il target dell'alveolare continui ad essere identificabile (ad es. in PI1), l'assimilazione di luogo si realizza grazie ad una diversa relazione di fase tra il gesto di TT e LL. Per le assimilazioni di luogo, infatti, il target dell'alveolare si trova in corrispondenza di un plateau di LL, che inizia con la postalveolare in C_1 e termina dopo il target dell'alveolare, poiché la protrusione è mantenuta per tutta la durata del nesso. La durata del gesto di chiusura di C_1 è minore per PI1 e maggiore per PF4, come se per PF4 C_1 fosse parte della sillaba C_2V_2, e, infatti, una differenza tra PI1 e PF4 riguarda il C-center: per PI1, il target di C_1 mostra un C-center anticipato rispetto a quello riscontrato per C_2 (e quindi C_1 è nell'offset della sillaba V_1C_1); per PF4, invece, il C-center di C_1 è posticipato, addirittura leggermente ritardato rispetto al C-center di C_2, ad indicare che possa esserci stata risillabificazione e C_1 faccia parte dell'onset – divenuto complesso - della sillaba C_2V_2..

In conclusione, i processi fonologici si realizzano a velocità sostenuta, sebbene in misura maggiore per il francofono, e generalmente in assenza di confine, confermando le nostre ipotesi iniziali circa l'influenza dei fattori considerati. Inoltre i due parlanti realizzano assimilazioni progressive di luogo per la sequenza postalveolare-alveolare in modo differente in termini di durata del gesto di chiusura e di C-Center, e i dati

suggeriscono che per il francofono potrebbe esserci stata una risillabificazione, mentre per l'italofono le due fricative continuano ad appartenere a due sillabe differenti e l'assimilazione sembra essere data soprattutto dalla protrusione di LL. In ogni caso, non si tratta di fenomeni sistematici, né per il francofono né per l'italofono, e questo fa pensare che in realtà, più che di fenomeni fonologici, si tratti di eventi fonetici.

Riferimenti bibliografici

[1] Browman P. C., Goldstein, L. 1980. Articulatory gestures as phonological units, *Phonology*, 6, 201-251.
[2] Browman P. C., Goldstein, 1988. Some notes on syllable structure in Articulatory Phonology. *Phonetica*, 45, 140-155
[3] Byrd D., Tan C.C. 1996. Saying consonant clusters quickly. *Journal of Phonetics* 4, 263-282.
[4] Byrd D., Choi S., 2006. At the juncture of prosody, phonology, and phonetics – The interaction of phrasal and syllable structure in shaping the timing of consonant gestures, *Proc. Conference on Laboratory Phonology*, Paris.
[5] Davidson L. 2006. Phonology, phonetics or frequency: influences on the production of non-native sequences. *Journal of Phonetics* 34, 104-137.
[6] Evers V., et al. 1998. Crosslinguistic acoustic categorization of sibilants independent of phonological status, *Journal of Phonetics*, 26, 345-370.
[7] Farnetani E., Busà M.G. 2004. Italian clusters in continuous speech. *Proc. ICSLP*, 1, 359-362, Yokohama, Japan.
[8] Jesus M. T., Shadle C. H., 2002, A parametric study of the spectral characteristics of European Portuguese fricatives, *Journal of Phonetics*, 30, 437-464.
[9] Maniwa K., Jongman A. 2009. Acoustic characteristics of clearly spoken English fricatives, *JASA*, 125,6, 3962-3973.
[10] Muliacic Z. 1973. *Fonologia della lingua italiana*. Ed. Il Mulino, Bologna.
[11] Niebuhr O., et al.. 2008. On place assimilation in French sibilant sequences. *Proc. ISSP*, 221-224, Strasbourg, France.
[12] Oh E. 2008. Coarticulation in non-native speakers of English and French: an acoustic study. *Journal of Phonetics*, 36, 361-384.
[13] Tiede M., et al. 2007. Gestural phasing in /kt/ sequences contrasting within and cross word contexts. Proc. ICPhS, 521-524, Saarbruken, Germany.
[14] Walker D.C. 1982. *On a phonological innovation in French*. Ed. Cambridge University Press, 12, 72-77.

La percezione di varianti allofoniche condizionate: uno studio neurofisiologico

Sandra Miglietta[a, b], Mirko Grimaldi[b], Andrea Calabrese[c, b]

[a] Dipartimento Antichità, Medioevo e Rinascimento, Linguistica, Università di Firenze, Italy

[b] Centro di Ricerca Interdisciplinare sul Linguaggio (CRIL), Università del Salento, Italy

[c] Department of Linguistics, University of Connecticut, USA

Introduzione

Il processo di assimilazione è uno dei principali fenomeni che genera variazione allofonica, inducendo un fonema a modificarsi e ad assumere alcune delle caratteristiche del suono vicino (Kiparsky, 1995).

Questo lavoro si concentra su un processo di assimilazione vocalica presente in una varietà dell'Italia meridionale (XXXXX), con un sistema tonico a cinque vocali (/i, ɛ, a, ɔ, u/), dove /ɛ/ diventa [e] quando è seguita da una vocale atona -i: [ˈmɛte]/[ˈmeti] *io/tu mieto/i*; [ˈdɛnte]/[ˈdenti] *dente/i*, ecc. (XXXX). Il processo di assimilazione produce, quindi, la variazione allofonica [ɛ-e].

Studi comportamentali come Peperkamp et al. (2003) per il francese e Boomershine et al. (2008) per l'inglese e lo spagnolo hanno osservato che i parlanti hanno difficoltà a percepire la variazione allofonica consonantica. Da questi studi emerge che i parlanti riescono solo a processare parametri del segnale acustico correlati a contrasti fonemici.

Recentemente, la percezione di contrasti fonemici e di variazioni allofoniche è stata anche indagata con tecniche neurofisiologiche, come la Mismatch Negativity (MMN), un Potenziale Evento Correlato (ERP) indice robusto di tracce mnestiche connesse con l'elaborazione di fonemi (Näätänen et al. 2007).

Per esempio, Hacquard et al. (2007) dimostrano che la MMN prodotta da parlanti francesi e spagnoli nella elaborazione della coppia vocalica [ɛ-e] è identica, nonostante che per gli spagnoli si tratti di una variazione allofonica. Kazanina et al. (2006), invece, studiando parlanti coreani e russi, trovano che la MMN è presente solo per i secondi nella elaborazione uditiva della coppia [t-d], che è allofonica in coreano e fonemica in russo.

Tuttavia, come fa notare (XXXXX, 2012), se le rappresentazioni percettive computate attraverso il segnale acustico contenessero solo le informazioni sui contrasti fonemici, le variazioni fonetiche presenti nel segnale acustico non sarebbero percepibili. Così, i processi allofonici sia della L1 che della L2 non potrebbero essere acquisiti, e le varianti fonetiche dovute a differenze dialettali, sociolinguistiche, di registro, ecc., non sarebbero percepite, come invece avviene normalmente.

Obiettivi

Visti i risultati contrastanti in letteratura, e in particolare la carenza di studi sulla allofonia condizionata, questo lavoro si prefigge di indagare la percezione della variazione allofonica [ɛ-e] e del contrasto fonemico

[e-i] presenti nel dialetto di XXXXXX (XXXXXXXX), utilizzando tecniche comportamentali (attentive) e neurofisiologiche (preattentive). La nostra ipotesi è che le variazioni allofoniche prodotte da un processo di assimilazione siano discriminate dai parlanti in cui il fenomeno è attivo.

Metodi

12 soggetti, 7 donne, età media 21.2, hanno preso parte alle sessioni sperimentali, tutti parlanti nativi del dialetto di XXXXX. Un parlante nativo del dialetto di XXXXX ha prodotto gli stimoli [ɛ, e, i] utilizzati in questo studio (registrazione in una camera anecoica con CSL 4500 e microfono Shure SM58-LCE, campionamento a 44.1 kHz, risoluzione di ampiezza a 16 bits). Per ogni vocale sono stati scelti tre esemplari diversi ma selezionati con valori Hz di F0, F1 e F2 simili, in modo tale da introdurre variazione acustica naturale negli stimoli discriminati dai soggetti sperimentali.

Un test di discriminazione AX ha verificato la discriminazione attentiva della variazione allofonica [ɛ-e]. Gli ERP con un paradigma oddball (85% di stimoli standard, 15% di stimoli devianti) hanno indagato il processo di discriminazione preattentivo del contrasto fonemico [e-i] e della variazione allofonica [ɛ-e] analizzando in particolare la MMN. È stata utilizzata una cuffia a 64 canali actiCAP. Montaggio della cuffia, acquisizione, filtraggio e analisi del segnale EEG sono stati eseguiti secondo le linee guida di Picton et al. (2000).

Nella selezione delle coppie di contrasti, abbiamo tenuto conto delle distanze acustiche fra gli stimoli. Abbiamo quindi deciso di usare la variante allofonica [ɐ] sia per la condizione allofonica che per quella fonemica, così da ridurre le differenze acustiche fra le condizioni. In particolare, il fonema /i/ e stato accoppiato con [e] ottenendo una distanza acustica di 88mel, molto vicina a quella della coppia allofonica [e-ɛ] (130mel). L'utilizzo di [ɛ] per il contrasto fonemico avrebbe invece prodotto una distanza acustica di 212mel. Ciò ha permesso di tenere costante per quanto possibile le distanze acustiche fra le coppie di stimoli, poiché è ben noto come questo parametro possa influenzare l'ampiezza della MMN (Näätänen et al. 1997; 2011).

Infine per ridurre gli effetti sulla componente MMN dovuta alle caratteristiche acustiche degli stimoli, abbiamo applicato l'approccio della *identitiy* MMN (Pulvermüller & Shtyrov, 2006).

Risultati

Il test *attentivo* di discriminazione ha rilevato che i soggetti discriminano accuratamente il contrasto allofonico (*d'*= 2.55). Un'analisi ANOVA dei dati ERP ha evidenziato la presenza significativa di MMN nelle due condizioni (condizione allofonica: F (1,66) = 14.592, $p < 0.001$; condizione fonemica F (1,66) = 6.047, $p < 0.05$). Ciò significa che i soggetti hanno discriminato sia il contrasto fonemico che la coppia allofonica.

Un'analisi ANOVA dell'ampiezza e della latenza della componente MMN ha rilevato che l'*ampiezza* non è significativamente diversa fra le due condizioni (F (2,66) = 0.283, p = 0.76), mentre la *latenza* è più precoce per il contrasto fonemico (F (1,66) = 6.017, $p < 0.05$).

Nel complesso questi dati indicano che la latenza della MMN è un indice dello status fonologico della coppia vocalica: benché entrambe le coppie di stimoli elicitino in ampiezza la stessa MMN, il contrasto fonemico è computato più velocemente del contrasto allofonico (vd. Fig. 1).

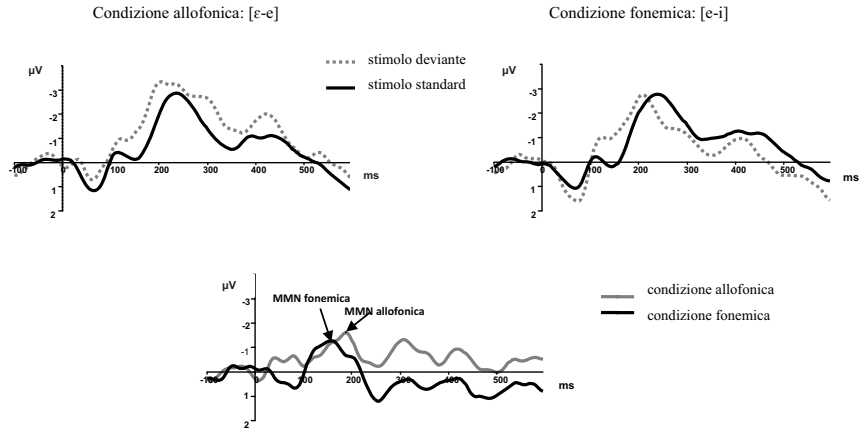


Fig.1. *Sopra:* ERP uditivi per gli stimoli devianti (linea grigia tratteggiata) e per gli stimoli standard (linea nera continua). Destra: variazione allofonica; sinistra: contrasto fonemico. *Sotto*: MMN per la condizione allofonica (linea grigia) e fonemica (linea nera).

Conclusioni

I risultati del nostro studio sono in accordo con Hacquard et al. (2007), e dimostrano che nella percezione uditiva, alternanze allofoniche predicibili (generate da un processo fonologico) condividono proprietà dei contrasti fonemici. Ne deriva un modello fonologico in cui l'acquisizione di categorie fonemiche avviene insieme all'apprendimento di pattern fonetici e delle relazioni (regole) che ci sono fra di loro.

D'altra parte la latenza precoce della MMN per il contrasto fonemico suggerisce la presenza di due distinte modalità di percezione: una *modalità fonologica* più precoce e una *modalità fonetica* più lenta. L'ipotesi è che entrambe le modalità attivino una analisi dei parametri acustici e quindi tracce *mnemoniche a breve termine.* Con la modalità fonologica i parlanti identificano differenze di significato tra parole, e solo le proprietà contrastive dei suoni sono decodificate e computate. Tale restrizione spiega la facilitazione della discriminazione fonemica. La modalità fonetica si attiva per processare sia le proprietà contrastive che quelle non contrastive. Ciò richiederebbe un surplus computazionale e un rallentamento del processo.

Bibliografia

Anonimo (2012). XXXXXXXXXXXXX

Anonimo (2006). XXXXXXXXXXXXX

Anonimo (2010). XXXXXXXXXXXXX

Boomershine, A., Hall, K. C., Hume, E., & Johnson K. (2008). The impact of allophony versus contrast on speech perception. In P. Avery, E. Dresher, & K. Rice (Eds.), *Contrast in Phonology* (pp. 143–172). Berlin: Mouton de Gruyter.

Hacquard, V., Walter, M. A., & Marantz, A. (2007). The effects of inventory on vowel perception in French and Spanish: an MEG study. *Brain. Lang., 100*, 295–300.

Kazanina, N., Phillips, C., & Idsardi, W. J. (2006). The influence of meaning on the perception of speech sounds. *In Proc. Natl. Aca. Sci. U S A* (pp. 1138–1186).

Kiparsky, P. (1995). The Phonological Basis of Sound Change, in Goldsmith J. A., The Handbook of Phonological Theory, Cambridge MA, Blackwell, 640-669.

Näätänen, R., Kujala, T., & Winkler, I. (2011). Auditory processing that leads to conscious perception: A unique window to central auditory processing opened by the mismatch negativity and related responses. *Psychophysiology, 48*, 4–22.

Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of: A review. *Clin. Neurophysiol., 118,* 2544–2590.

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Vainio, M., Alku, P., Ilmoniemi, R. J., Luuk, A., Allik, J., Sinkkonen, J., & Alho, K. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature, 385,* 432–434.

Peperkamp, S., Pettinato, M., & Dupoux, E. (2003). Reinterpreting loanword adaptations: the role of perception. In B. Beachley, A. Brown, & F. Conlin (Eds.), *Proceedings of the 27th Annual Boston University Conference on Language Development* (pp. 650–661). Somerville, MA: Cascadilla Press.

Pulvermüller, F., & Shtyrov, Y. (2006). Language outside the focus of attention: The mismatch negativity as a tool for studying higher cognitive processes. *Prog. Nueurobiol., 79,* 49–71.

Transfer intonativo in inglese L2: analisi fonetico-fonologica delle produzioni di parlanti padovani

Antonio Stella, Maria Grazia Busà
Università degli Studi di Padova
antonio.stella@unipd.it; mariagrazia.busa@unipd.it

La variazione degli accenti tonali nelle diverse condizioni di focus rappresenta un ostacolo difficile da realizzare per gli apprendenti di una lingua straniera, i quali sono influenzati dalle strategie intonative utilizzate nella lingua nativa. Ueyama [1] mostra che in parlanti giapponesi che apprendono l'inglese americano come lingua straniera le caratteristiche fonologiche dell'intonazione sono acquisite prima di quelle fonetiche, apprese correttamente solo da parlanti con alti livelli di competenza. Inoltre Mennen [2] mostra che il dettaglio fonetico di categorie fonologiche presenti sia nel sistema nativo che in quello non-nativo è difficilmente appreso anche da parlanti con alti livelli di competenza nella lingua straniera.

In questo contributo il nostro obiettivo è quello di analizzare le produzioni in inglese non-nativo prodotto da parlanti italiani di Padova con differente livello di competenza della lingua straniera. Lo scopo è quello di comparare le strategie di implementazione fonetica degli accenti tonali usate nell'italiano nativo e nell'inglese nativo e il grado di influenza dell'italiano nativo sulla produzione dell'inglese come lingua straniera per parlanti con alto e basso livello di competenza. L'analisi in produzione, insieme a verifiche percettive da svolgere successivamente, è necessaria per comprendere l'apprendimento dell'intonazione di una lingua straniera e quindi per poter sviluppare delle strategie didattiche.

L'analisi fonetica è condotta su accenti tonali in posizione iniziale di enunciati con focalizzazione differente: focalizzazione larga (BF), nella quale tutto l'enunciato è in focus, e focalizzazione stretta contrastiva (CF), nella quale invece il focus è ristretto solo alla parola target in posizione iniziale, che rappresenta l'elemento oggetto di contrasto. I materiali sperimentali sono elicitati utilizzando una serie di mini-dialoghi nei quali il soggetto, rispondendo a due domande, produce prima un BF e poi un CF su enunciati identici.

Tutte le produzioni sono state etichettate identificando sia l'onset e l'offset di ogni sillaba, che i tre target tonali dell'accento: L1, che rappresenta il target basso all'inizio dell'ascesa di F0; H, che rappresenta il picco accentuale; L2, che rappresenta il target basso alla fine della discesa di F0. Si è quindi provveduto a misurare: 1) l'allineamento dei target tonali dall'onset e offset della sillaba tonica; 2) l'altezza tonale dei target, per misurare le variazioni del campo di frequenze e dell'escursione degli accenti; 3) la durata della sillaba tonica.

Una prima serie di risultati è già stata raccolta sulla base delle produzioni in italiano e in inglese da parte di 3 parlanti nativi di Padova e sulle produzioni in inglese da parte di 3 parlanti nativi di Londra. I parlanti padovani che finora hanno preso parte agli esperimenti di produzione fanno parte del gruppo con basso livello di competenza. Il nostro obiettivo è di raccogliere i dati di 5 parlanti per ognuno dei 2 livelli di competenza e di 5 parlanti provenienti dalla zona di Londra. La raccolta dei dati è attualmente in corso.

I risultati preliminari mostrano che nella varietà di italiano parlata a Padova il picco tonale è allineato in anticipo in CF rispetto a BF, mentre i target bassi tendono ad avvicinarsi al picco tonale; questo spostamento si riflette soprattutto sulla durata dell'ascesa tonale che è sistematicamente più corta, sebbene sia sempre allineata con la sillaba tonica. La discesa tonale invece è realizzata sulla sillaba post-tonica e mostra una modificazione variabile tra i parlanti. Per quanto riguarda il campo di frequenze e l'escursione tonale dell'accento, i parlanti abbassano sistematicamente il valore di F0 dei tre target tonali nella produzione del CF; l'abbassamento maggiore interessa il picco dell'accento tonale che si abbassa quasi sistematicamente di un valore doppio rispetto ai target bassi, con una media di circa 50 Hz. In tal modo sia il campo di frequenze che l'escursione tonale di ascesa e discesa risultano più basse nella produzione di CF, in linea con quello che succede in altre varietà di italiano [3] [4]. La sillaba tonica sembra invece essere sistematicamente allungata di circa 60 ms nel CF rispetto al BF. Un esempio dei contorni intonativi prodotti dai parlanti nativi di Padova è riportato in Fig. 1. Data l'implementazione fonetica, l'accento tonale prodotto in posizione iniziale di enunciato potrebbe essere etichettato come L*+H nelle produzioni in BF, e L+H* in caso di CF. Le motivazioni di questa scelta derivano dalla posizione del picco, il quale si trova oltre la sillaba tonica nei due tipi di focalizzazione; in caso di CF infatti l'accento rimane comunque ascendente e il picco è estremamente vicino all'offset della sillaba tonica, facendo propendere per un cambio nel tono associato alla sillaba accentata. Tale trascrizione trova parziale riscontro nella descrizione del dialetto trevigiano effettuata in [5].

Per quanto riguarda le produzioni in inglese nativo, l'allineamento sembra essere utilizzato in maniera simile all'italiano: vi è sempre una ritrazione del picco dell'accento ed una maggiore variabilità nella posizione dei

target bassi. Differentemente dall'italiano, la variazione dell'altezza tonale non sembra essere un correlato prosodico utilizzato dai parlanti inglesi per differenziare i due tipi di focus: dal confronto tra BF e CF risulta infatti solo una diminuzione di una media di 10 Hz per tutti i target tonali in tutti e tre i parlanti. La durata sillabica è un correlato utilizzato dai parlanti inglesi per differenziare i due tipi di focus: essa è allungata di circa 60 ms in CF. Un esempio dei contorni intonativi prodotti dai parlanti nativi di Londra è riportato in Fig. 2.

In un'ottica comparativa le differenze fonetiche tra l'italiano parlato a Padova e l'inglese di Londra nell'implementazione dei due accenti tonali si riscontrano soprattutto in un diverso uso dell'altezza tonale, mentre l'allineamento sembra essere utilizzato in maniera simile, con una costante ritrazione del picco accentuale in CF. Nelle produzioni in inglese come lingua straniera, i parlanti padovani trasferiscono quasi completamente le caratteristiche prosodiche del sistema nativo, differenziando i due tipi di focus attraverso una forte diminuzione di F0 nel caso di CF rispetto a BF. Si ricorda che tali dati provengono da parlanti con bassa competenza; rimane da verificare se tale correlato sia correttamente modulato da parlanti con alto livello di competenza.

Le considerazioni sulla realizzazione fonetica degli accenti, benché siano il risultato dell'analisi di soli tre parlanti per lingua, forniscono un primo panorama sull'uso dei correlati prosodici nella variazione del contenuto pragmatico di un enunciato. E' da evidenziare comunque che le modificazioni fonetiche a livello di un singolo evento tonale non sono le uniche a determinare la differenza tra CF e BF: nell'inglese infatti si registra anche la presenza di una cesura prosodica dopo la parola in focus seguita da una marcata compressione degli eventi tonali post-focali. Dai risultati preliminari, tali fenomeni non trovano un riscontro sistematico nelle produzioni dell'italiano parlato a Padova. Con l'ultimazione della raccolta dei dati si potrà delineare un quadro completo delle differenze prosodiche che possono rappresentare un ostacolo nell'apprendimento dell'inglese come lingua straniera, sia a livello del singolo evento tonale che dell'intero contorno intonativo.

Figure

Figura 1: Esempi della prima parte dell'enunciato "*La Melania verrà domani mattina*" in condizione di BF (sinistra) e CF (destra), prodotti dal parlante S4, parlante nativo di Padova. La parola sottolineata rappresenta la parola target.

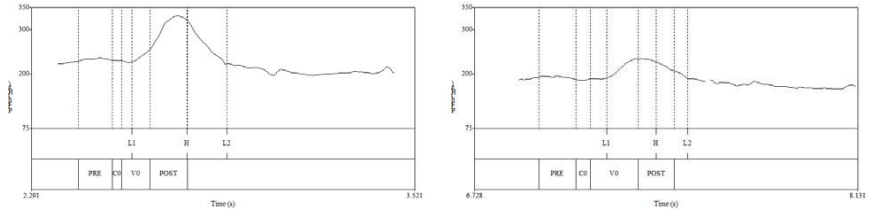
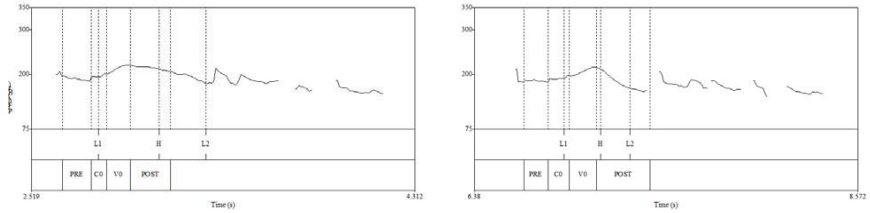


Figura 2: Esempi dell'enunciato "*The memorial will be built this year*" in condizione di BF (sinistra) e CF (destra), prodotti dal parlante S2, parlante nativo di Londra. La parola sottolineata rappresenta la parola target.



Bibliografia

[1] Ueyama, M. (1997), The phonology and phonetics of second language intonation: the case of "Japanese English", in *Proceedings of the 5th European Speech Conference*, 2411-2414.

[2] Mennen, I. (2007), Phonological and phonetic influences in non-native intonation, in *Non-native Prosody: Phonetic Descriptions and Teaching Practice*, The Hague: Mouton De Gruyter, 53-76.

[3] D'Imperio, M. (2002). Italian intonation: An overview and some questions. *Probus* 14, 37-69.

[4] Gili-Fivela, B. (2008). *Intonation in Production and Perception: The Case of Pisa Italian*. Alessandria: Edizioni dell'Orso.

[5] Payne, E. (2005) Rises and rise-plateau-slumps in Trevigiano, in *Cambridge Occasional Papers in Linguistics*, *2*, 173-186.

DiphthongClassification: un algoritmo per la classificazione dei dittonghi

Luciano Romito, Tarasi, Vigè, Rosita Lio

Abstract

L'obiettivo di questo lavoro è sviluppare un algoritmo in grado classificare la natura ascendente o discendente dei dittonghi. Questa idea nasce da un lavoro precedente (Romito, Tarasi & Renzelli 2010) in cui, insieme ad altri parametri, è stata verificata la presenza/assenza della metafonia per dittongazione in alcuni centri della Calabria. Come è stato dimostrato nel lavoro sopra citato, lo studio di questa variabile risulta molto interessante per due motivi: la sua resistenza ai cambiamenti sociolinguistici che interessano il territorio calabrese e la sua diffusione in alcuni centri a sud dell'attuale isoglossa che la identifica.

Tra i processi fonetici e fonologici che si sviluppano all'interno di una lingua, i fenomeni di armonizzazione tra due o più foni rivestono un ruolo molto importante. Tra questi vi è la *metafonia*, un fenomeno che interessa non solo la maggior parte delle lingue, ma anche molti dialetti italiani. Questo fenomeno viene annoverato nel processo fonologico dell'assimilazione. Essa è un tipo di assimilazione che riguarda vocali non adiacenti, in quanto comporta l'assimilazione progressiva della vocale accentata di una parola alla vocale seguente di un suffisso.

La metafonia, in genere, è descritta come fenomeno diacronico. Tale fenomeno consente di spiegare la derivazione di alcune parole attuali del lessico italiano come, ad esempio, 'uscio' che deriva dal latino OSTIUM. In questo caso, per effetto di metafonia, si assiste al passaggio della vocale [o] tonica ad [u] per effetto della vocale atona seguente [i].

Come è noto la metafonia può avvenire per **innalzamento** (la '*e'* e la '*o'* passano rispettivamente a '*i'* e '*u'*), per **dittongazione** (la '*e'* e la '*o'* dittongano generalmente in '*ie'* e in '*uo'* ma anche '*ua'* ascendenti o discendenti) e, nelle zone in cui cade il dittongo, può manifestarsi per **monottongamento**. In quest'ultimo caso, il dittongo è ritratto sul primo componente e l'esito metafonetico diventa, in alcuni dialetti calabresi, [ie] >[i:] ([$pi^{e}di$] > [pi:di]), [uo]>[u:] ([$cu^{o}ttu$] > [cu:ttu]).

La metafonia è il fenomeno più rilevante tra i processi di assimilazione a distanza fra vocali e, per questo motivo, occupa uno spazio privilegiato nelle grammatiche delle lingue e soprattutto dei dialetti italiani. Questo fenomeno non è presente nel toscano, ma i suoi effetti persistono in molti dialetti dell'Italia centro-meridionale. In questa area, il fenomeno si manifesta sia da -Ī che da -Ŭ. In particolare, ne possono essere distinti due tipi:

- *napoletana*, chiude le vocali toniche /e/ (< Ē, Ĭ) ed /o/ (< Ō, Ŭ) in /i/ e /u/ quando le vocali finali sono, o erano, -Ī ed –Ŭ; nelle stesse condizioni le toniche /ɛ/ (< Ĕ) ed /ɔ/ (< Ŏ) danno luogo a dittonghi vari;
- *ciociaresca* o *arpinata*, concorda con il tipo napoletano per il trattamento delle vocali medie chiuse, ma non riduce le vocali /ɛ/ ed /ɔ/ a dittonghi, bensì il dittongo è formato dalle vocali /e/ ed /o/.

La metafonia delle vocali medio-alte comporta l'assimilazione totale al grado di apertura delle vocali che attivano il processo. Il conseguente sviluppo storico di questi esiti metafonizzati segue di solito quello delle [i] ed [u] originarie, mentre sulle vocali medio-basse la metafonia agisce provocando dittonghi come [je], [wo], [jɛ], [wɔ], ['iə], ['uə].

Materiali e Metodi

Il corpus utilizzato in questa ricerca è composto da materiale sonoro contenuto nell'Archivio Sonoro Calabrese realizzato dal Laboratorio di Fonetica dell'Università della Calabria e da nuove registrazioni che interesseranno ulteriori centri della Regione. Le registrazioni sono state effettuate, e saranno condotte, secondo diverse modalità:

- interviste basate su brani letti;
- racconti;
- lettura di un questionario sviluppato ad hoc per questa ricerca;
- lettura di una lista di parole costruita appositamente per questo lavoro.

Le nuove registrazioni saranno eseguite all'interno della camera silente (4*4 Amplifon) presso il Laboratorio di Fonetica dell'Università della Calabria con un registratore digitale EDIROL 24-bit e un microfono Philips. Il materiale sonoro, in parte etichettato, sarà analizzato attraverso un tool sviluppato appositamente in PRAAT. Questo tool come input utilizza una matrice di valori formantici (F1-F2) estratta dallo stesso programma per la porzione di interesse, cioè l'intero dittongo, e riconosce automaticamente i centroidi dei due differenti segmenti vocalici presenti all'interno del dittongo e in seguito, con porzioni di 20 ms, associa gli spettri successivi o precedenti al primo o al secondo elemento. Tale analisi ci permetterà di classificare automaticamente la natura ascendente o discendente del dittongo e di conseguenza l'elemento tonico e la durata dell'elemento atono. Infine, il risultato ottenuto viene restituito in forma grafica. L'applicazione è stata sviluppata in Matlab ed elabora le misurazioni dei valori di f1 e f2, da inserire nella forma di una matrice contenente per ogni riga la terna dei valori (t,f1,f2) in cui t è l'istante di misurazione e f1 e f2 sono i valori delle formanti all'istante t.

Conclusioni

Lo scopo della ricerca è stabilire una scala di assorbimento, assimilazione o riduzione a monottongo del dittongo analizzato.

LA COARTICOLAZIONE E IL VOT NELLO SVILUPPO FONETICO: STUDIO SPERIMENTALE SU BAMBINI DAI 42 AI 47 MESI D'ETA'

Claudio Zmarich[1], Elisa Bortone[2], Mario Vayra[2], Vincenzo Galatà[1,3]
[1]CNR-ISTC, Padova (I), [2]Università di Bologna (I), [3]CNR-IRAT, Napoli (I)

ABSTRACT

Nelle prime tappe dello sviluppo linguistico, l'aspetto relativo all'acquisizione del corretto controllo motorio pone al bambino una sfida altrettanto complessa e impegnativa dell'acquisizione delle categorie cognitive relative alla produzione fonologica e, più in generale, linguistica (Zmarich, 2010). Lo studio dell'acquisizione del controllo motorio è reso però complicato nel bambino prescolare dalla difficoltà di usare dispositivi per la rilevazione diretta dei movimenti (perché richiedono soggetti collaborativi) e dalla inadeguatezza dell'analisi del percetto uditivo, perché basata sulla trascrizione fonetica e quindi su categorie qualitative. Ecco allora che la metodologia d'elezione diventa l'analisi acustica, che è in grado di quantificare il continuum tempo-frequenziale dei foni, e di ricavare per inferenza informazioni sui movimenti che li hanno prodotti. Il *Voice Onset Time* o VOT, che misura l'intervallo temporale che va dal rilascio dell'occlusione consonantica all'inizio di vibrazione delle corde vocali, è considerato il miglior parametro distintivo per classificare e quantificare la sonorità consonantica, determinata dal rapporto temporale tra l'azione glottale e l'articolazione sopraglottale (Lisker & Abramson 1964). Nello specifico caso delle consonanti occlusive il VOT può essere calcolato sottraendo al valore in (ms) del momento iniziale della vibrazione glottica, il valore in (ms) del momento del rilascio consonantico (*burst*). Smbra che le consonanti sorde e sonore all'inizio dello sviluppo fonetico siano realizzate con *voicing lag* (cioè entrambe vengono realizzate come sorde non aspirate). In seguito incominciano a differenziarsi dal punto di vista acustico, con una distribuzione statisticamente bimodale, ma le differenze non superano la soglia percettiva (stadio nascosto). In uno stadio successivo, sorde e sonore dell'italiano vengono realizzate rispettivamente con valori di *voicing lag* e *voicing lead* (quando le corde vocali vibrano già durante l'occlusione, come nell'italiano) molto alti (cioè sorde e sonore sono esageratamente diverse); alla fine sorde e sonore vengono realizzate rispettivamente con *voicing lag* e *voicing lead* secondo la norma adulta (Macken e Barton, 1980).

Oltre che per le durate, l'analisi acustica risulta particolarmente utile anche per lo studio della coarticolazione (Recasens, 1999). Con il termine coarticolazione ci si riferisce all'influenza (acustica, articolatoria, percettiva) di un fono su un altro, che lo segue (c. perseverativa) o lo precede (c. anticipatoria). Secondo l'ipotesi oggi più accreditata, nello sviluppo fonologico il bambino restringe progressivamente il dominio dell'organizzazione articolatoria dalla sillaba ai singoli gesti C e V, quindi durante lo sviluppo la coarticolazione diminuisce e la distintività fonemica aumenta (Studdert-Kennedy e Goldstein, 2003). Petracco e Zmarich (2006) hanno descritto e quantificato la coarticolazione anticipatoria (di V su C) in sillabe "CV" (C = [p/b],[t/d],[k/g] e V = qualsiasi vocale) prodotte da una bambina, dal babbling dei 10 mesi alle prime parole a 18 mesi, usando l'andamento di F2 nella transizione tra C e V come indice del luogo di occlusione lungo la direzione antero-posteriore del cavo orale (Fant, 1963). Sebbene in nessun mese i gradi di coarticolazione per i tre luoghi articolatori siano uguali a quelli dei soggetti adulti, nondimeno essi seguono profili evolutivi diversi a seconda del luogo consonantico interessato, e le differenze possono essere spiegate dalla forza dei vincoli anatomofisiologici coinvolti nell'interazione tra C e V (cfr. anche Sussman et alii, 1999).

Sui modi e i tempi in cui i bambini apprendenti l'italiano acquisiscono il controllo motorio necessario a produrre valori di tipo adulto si sa ben poco: per quanto riguarda la coarticolazione si è già detto dello studio di Petracco e Zmarich (2006) che però è limitato a un solo soggetto e che si

ferma ai 18 mesi di età. Per quanto riguarda il VOT, c'è solo uno studio di Bortolini et alii (1995) che analizza l'evoluzione di questo parametro in un piccolo gruppo di bambini dai 18 ai 21 mesi. A tutt'oggi non sappiamo cosa succede nei mesi e negli anni successivi, fino all'età di raggiungimento dei valori di tipo adulto. Il presente studio ha preso in esame le produzioni effettuate da 10 bambini (5 maschi e 5 femmine) di età compresa tra i 42 e i 47 mesi, con l'intento di valutare il loro livello di acquisizione per il VOT e la coarticolazione CV. Il campione è stato selezionato dalle registrazioni di quasi 100 bambini effettuate presso due istituti d'infanzia di Padova, in occasione dello studio di Galatà e Zmarich (2011). I soggetti sono stati sottoposti ad un "test di produzione" durante il quale veniva chiesto loro di ripetere una lista di non-parole "CVCV". Per quanto concerne questo studio, il test mirava a stimolare la produzione di consonanti occlusive sorde e sonore, situate soprattutto a inizio di parola, ma anche in posizione intervocalica. Dalle analisi acustiche, condotte in modo semiautomatico con l'aiuto di alcuni script di PRAAT, sono stati ricavati i dati di VOT per tutte le consonanti occlusive iniziali di parola, i valori delle seconde formanti (F2) per le consonanti occlusive in posizione iniziale e intervocalica nel primo e nel secondo ciclo dopo il burst, ed infine i valori di F2 per le vocali che seguivano le occlusive.

Come già in Bortolini et al. (1995), i 10 bambini qui esaminati producono mediamente il contrasto di sonorità differenziando le consonanti sorde dalle sonore per le bilabiali e le dentali, ma hanno qualche difficoltà nella produzione della sonorità per le velari che in qualche caso presentano anche valori positivi (sebbene ci sia una minoranza di bambini che non produce il contrasto di sonorità non solo per le velari ma anche per gli altri luoghi articolatori). Questi casi dimostrano che la difficoltà della produzione della sonorità nelle consonanti occlusive, soprattutto velari, è ancora ben presente tra i 42 e i 47 mesi di età. Per quanto riguarda la coarticolazione, i dati dello studio qui eseguito hanno riscontrato la persistenza, in gradi variabili, degli stessi vincoli di natura anatomo-fisiologica presenti nella bambina studiata da Petracco e Zmarich (2006).

BIBLIOGRAFIA

Bortolini U., Zmarich C., Fior R., Bonifacio R.(1995), Word-initial voicing in the productions of stops in normal and preterm Italian infants, International Journal of Pediatric Otorhinolaryngology, 31, 191-206.

Galatà V. & Zmarich C. (2011) , 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.), Contesto comunicativo e variabilità nella produzione e percezione della lingua , Proceedings of the 7th AISV National Conference, 26-28 January 2011, Università del Salento – Lecce, Bulzoni Editore: Roma, vol. VII, 118-129.

Lisker L., Abramson A. S. (1964), A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements, Word, 20, 192, 384-442.

Macken M.A., Barton D. (1980), The acquisition of the voicing contrast in Spanish: a phonetic and phonological study of word-initial stop consonants, J. Child Lang, 7, 433-458.

Petracco A., Zmarich C. (2006), La quantificazione della coarticolazione nello sviluppo fonetico, in V. Giordani, V. Bruseghini, P. Cosi (a cura di), Atti del III Convegno Nazionale dell'Associazione Italiana di Scienze della Voce (AISV), Trento, 29-30/11- 1/12/2006, EDK Editore srl, Torriana (RN), 135-150.

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

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.

Zmarich C. (2010), Lo Sviluppo Fonetico e Fonologico da 0 a 3 anni, in Bonifacio S., Stefani L. Hvastja, L'intervento precoce nel ritardo di linguaggio: il modello INTERACT per il bambino parlatore tardivo, FrancoAngeli, 17-39.

Le patologie del parlato e il ruolo dello studio strumentale dell'articolazione

Paride Grotta, Barbara Gili Fivela, *Claudio Zmarich

*Università del Salento & CRIL - Lecce, *CNR-ISTC - Padova*

I canali di comunicazione sono molteplici e, quando sono sfruttati contemporaneamente, permettono a parlante e ascoltatore di usare a pieno informazioni di tipo multimodale. La comunicazione, e lo scambio dialogico in particolare, non risente ugualmente della mancanza delle informazioni relative ai vari canali. Ovviamente non è scontato che esista una gerarchia di importanza, visto che il riferimento ai vari canali può variare a seconda delle situazioni (es. informazioni sulla gestualità in conversazioni telefoniche vs di persona) e, addirittura, a seconda delle propensioni individuali (es. differenze individuali nell'integrazione di informazioni visive e acustiche – Fagel, 2005). Tuttavia, l'esito della comunicazione può essere più o meno compromesso a seconda del canale che viene a mancare o lungo il quale il flusso di informazioni risulta alterato.

In questo studio, prenderemo in considerazione alcune patologie che implicano un disturbo della produzione del parlato e che, quindi, possono alterare un aspetto cruciale e di primaria importanza nell'emissione del messaggio verbale e, in particolare, nello scambio dialogico. Il nostro obiettivo è mostrare che lo studio strumentale dell'articolazione dei suoni, ormai usuale in alcuni rami della fonetica e fonologia di laboratorio, può fornire un utile apporto per la descrizione di molte di queste patologie, il loro trattamento e la verifica degli effetti di diverse terapie effettuate per superarle. L'analisi dell'articolazione può quindi avere delle chiare ricadute sul miglioramento della comunicazione e dello scambio dialogico nel caso di patologie del parlato. Si tratta della fase iniziale di un progetto volto ad effettuare un'indagine strumentale di tipo articolatorio che metta in luce le caratteristiche principali delle produzioni orali dei soggetti affetti da disprassia e verifichi se sia possibile ottenere dei sensibili miglioramenti nelle loro produzioni, grazie a sedute di training nelle quali si usino informazioni ricavate strumentalmente sull'articolazione dei suoni linguistici (ad es., ricostruzioni 3D della cavità orale e del movimento della lingua nell'articolazione dei foni).

Come metteremo in evidenza nel nostro contributo, benché la complessità dell'argomento sia tale da rendere difficile una netta ripartizione di disordini e patologie del linguaggio (Darley et al., 1975; Ball et al., 2008; cfr ICD-10, ICF), quelli più interessanti, data la nostra prospettiva, sono legati a:

- incapacità, o all'alterata capacità (es. blesità), di articolare foni e parole, dovute a malattie degli organi dell'apparato fonatorio (es. logoplegia, dislalia)
- incapacità di compiere volontariamente i gesti articolatori (es. aprassia orale – soprattutto quella verbale, glossoplegia) o generali errori e problemi nell'articolazione, dovuti ad alterazioni cerebrali (es. disartria, afasie), o a stati psicofisici che non implicano alterazioni degli organi coinvolti nell'articolazione del parlato (es. balbuzie, cluttering)

Di fatto, tutti questi disturbi del linguaggio hanno una ricaduta più o meno marcata sulla qualità ad efficacia dello scambio dialogico, anche in relazione allo sfruttamento delle informazioni multimodali, visto che spesso sono alterati proprio i movimenti di labbra e lingua, visibili all'ascoltatore.

Nella letteratura relativa alle analisi strumentali del parlato sono presenti molti contributi che dimostrano come la comunità scientifica si sia già orientata allo studio articolatorio delle diverse patologie e della possibilità di migliorare la condizione delle persone affette da disturbi del parlato (Kent, 2000; Kent & Kim, 2003).

In ambito internazionale, nello studio delle patologie del parlato l'uso di indagini strumentali di tipo articolatorio, come l'articulografia elettromagnetica, è abbastanza diffuso (Wong et al., 2010) e i disturbi più indagati e descritti sono certamente la disartria, l'aprassia e la balbuzie (per disartria: Rong et al. 2012, Jaeger et al. 2000, McAuliffe, et al., 2005; Wong et al., 2010a; Wong et al. 2011; per aprassia: Katz, Levitt, Carter, 2003, Katz, Levitt,Carter, 2003; per balbuzie: van Lieshout et al. 1993, van Lieshout et al. 1993, 2004; McClean et al, 2004, McClean, Runyan, 2000). A parte corpora realizzati per studi specifici, sono state costruite anche banche dati preziose per la descrizione della produzione di parlato patologico (es. il database TORGO per la disartria, che include registrazioni video, audio e di dati di articulografia elettromagnetica 3D (AG500) – Rudzicz et al., 2008). In Italia e sull'italiano, invece, sono pochissimi gli studi relativi ad indagini articolatorie di parlato patologico, sicuramente anche per il fatto che gli strumenti più usati in quest'ambito disponibili sul territorio nazionale sono stati, e sono tutt'ora, pochissimi (con ovvie ripercussioni sulla possibilità di poter effettuare studi, in particolare su soggetti italiani). Da questo punto di vista, solo la balbuzie è stata indagata approfonditamente (Zmarich et al, 1994a, 1994b; Zmarich & Magno Caldognetto, 1997; Zmarich et al., 2005; Zmarich, Marchiori, 2006), mentre disartria e aprassia non ci risulta siano state oggetto di indagine articolatorie strumentali per l'italiano.

Paraltro sono anche abbastanza numerosi gli studi nei quali le indagini strumentali, in particolare l'articulografia elettromagnetica, siano volte a verificare l'effetto di terapie per patologie che causino problemi di tipo articolatorio nel parlato (Dromey 2000, Bose et al. 2001). Bose et al. (2001), ad esempio, dimostrano l'utilità del sistema PROMPT (un sistema di "insegnamento" della lingua orale che prevede stimoli uditivi, visivi e tattili - Hayden, 1984), su un soggetto adulto affetto da afasia di Broca e da aprassia. Anche rispetto all'uso delle indagini articolatorio-strumentali per monitorare gli effetti di protocolli riabilitativi, osserviamo la forte carenza di ricerche effettuate in Italia e, in generale, sull'italiano, benché l'utilità di queste indagini sia riconosciuta da tempo anche a livello nazionale, almeno per quanto riguarda la balbuzie (Zmarich, 1999).

Infine, un campo di indagine e di applicazione piuttosto recente e di crescente interesse è quello legato all'uso delle informazioni articolatorio-strumentali per la realizzazione di protocolli di riabilitazione e addestramento che di fatto si basano sul *biofeedback*. In quest'ambito, sono degni di nota i sistemi BALDI (Massaro, 2004) e ARTUR (Eriksson, 2005, Engwall, 2008), sistemi di addestramento tramite computer usati per l'insegnamento della pronuncia (non solo in caso di problemi di parlato/udito, ma anche per le lingue straniere). Di fatto, grazie all'uso di "facce parlanti", ossia facce animate da computer, questi sistemi danno la possibilità all'utente di rendersi conto della reale meccanica del parlato e di auto correggere i propri movimenti. Rispetto all'utilizzo del biofeedback, è particolarmente importante anche il contributo di Katz-McNeil, (2010) che hanno studiato l'effetto di feedback fornito in tempo reale per verificarne l'utilità in pazienti aprassici. Lo studio, effettuato per mezzo dell'articulografo elettromagnetico e grazie a sensori posizionati sulla lingua dei soggetti, descrive come sia possibile fornire informazioni in tempo reale sulla posizione della lingua (v. anche Schulz et al. 2006), mostrando ai soggetti come raggiungere

un target indicato sul monitor del computer, e come questo rappresenti un utile ausilio nel miglioramento dei problemi articolatori dovuti ad aprassia.

Le indagini strumentali di tipo articolatorio sono, quindi, molto promettenti per la descrizione e il trattamento di vari disordini e patologie del parlato, e anche per la verifica degli effetti delle terapie previste per il loro trattamento. È evidente, quindi, l'impatto positivo degli studi articolatori sul miglioramento della capacità e della facilità di comunicazione e di scambio dialogico nel caso di molte patologie del parlato. Peraltro, la ricognizione fatta mostra anche la scarsissima presenza di indagini relative all'italiano, indicando chiaramente future e fertili direzioni di ricerca.

Bibliografia

Adams S. G. and Dykstra A. (2009) *Hypokinetic dysarthria,* Clinical Management of Sensorimotor Speech Disorders, M. R. McNeil, Ed., Thieme, New York, NY, USA, 2nd edition.

Ball M.J., Perkins M.R., Müller N., Howard S. (2008),The Handbook of Clinical Linguistics, Blackwell Pub.

Beskow, J., Engwall, O., Granström, B., Nordqvist, P., & Wik, P. (2008). *Visualization of speech and audio for hearing-impaired persons*. Technology and Disability, 20(2), 97-107.

Bose A. and Square P. A., Schlosser R., Van Lieshout P., (2001) *Effects of PROMPT therapy on speech motor function in a person with aphasia and apraxia of speech.* APHASIOLOGY, *15* (8), 767–785.

Darley F. L., Aronson A. E., and Brown J. R. (1975) *Motor Speech Disorders*, W.B. Saunders Company, Philadelphia, Pa, USA.

Dromey C., (2000)*Articulatory kinematics in patients with Parkinson disease using different speech treatment approaches,*

Journal ofMedical Speech-Language Pathology, vol. 8, no. 3, pp. 155–161.

Jaeger, M., Hertrich, I., Stattrop, U., Schönle, P.-W., Ackermann, H. (2000) *Speech disorders following severe traumatic brain injury: Kinematic analysis of syllable repetitions using electromagnetic articulography*. Folia Phoniatrica et Logopaedica, 52: 187-196.

Engwall, O. (2008). *Can audio-visual instructions help learners improve their articulation? - an ultrasound study of short term changes.* Proceedings of Interspeech 2008 (pp. 2631-2634). Brisbane, Australia.

Engwall, O. (2005). *Introducing visual cues in acoustic-to-articulatory inversion.* Proceedings of Interspeech 2005. Lisbon, Portugal.

Eriksson E., Bälter O., Engwall O., Öster A.-M. (2005) *Design recommendations for a computer-based speech training system based on end-user interviews* (ARTUR), Proceedings of the Tenth International Conference on Speech and Computers, SPECOM 2005, 17-19 October, Patras, Greece 483-486.

Fagel, S., (2005). *Auditory Speech Illusion Evoked by Moving Lips.* Proceedings of the 10th International Conference on Speech and Computer, Patras, 115-118.

Hayden, D. A. (1984), *The PROMPT system of therapy: Theoretical framework and applications for developmental apraxia of speech,* Seminars in Speech and Language, 2,n.2,139-155.

Katz W. F., Levitt J. S., and Carter G. C. (2003). *Biofeedback treatment of buccofacial apraxia using EMA*, Brain and Language 87, 175-176.

Katz, W., Bharadwaj, S., & Carstens, B. (1999). *Electromagnetic articulography treatment for an adult with Broca's aphasia and*

apraxia of speech. Journal of Speech, Language, and Hearing Research, 42, 1355–1366.

Katz W.F., McNeil M. (2010) *Studies of Articulatory Feedback Treatment for Apraxia of Speech Based on Electromagnetic Articulography,* Perspectives on Neurophysiology and Neurogenic Speech and Language Disorders October 2010 20:73-79.

Kent R. D. (2000), Research on speech motor control and its disorders: A review and prospective, J. Communication Disorders, 33, 391-428

Kent R. D. & Kim Y.-J (2003), Toward an acoustic typology of motor speech disorders, Clinical Linguistics & Phonetics, 17, 427-445.

Kjellström, H., & Engwall, O. (2009). *Audiovisual-to-articulatory inversion*. Speech Communication, 51(3), 195-209.

Massaro, D. (2004) *Symbiotic Value of an Embodied Agent in Language Learning*, (BALDI), Proceedings of 37th Annual Hawaii International Conference on System Sciences (CD/ROM), Computer Society Press, 2004, CD Rom, 1-10.

McAuliffe M. J., Ward E. C., and Murdoch B. E., (2005) *Articulatory function in hypokinetic dysarthria: an electropalatographic examination of two cases,* Journal of Medical Speech-Language Pathology, vol. 13, no. 2, pp. 149–168.

McClean MD, Tasko SM, Runyan CM.(2004) *Orofacial movements associated with fluent speech in persons who stutter.*, J Speech Lang Hear Res. 2004 Apr;47(2):294-303.

McClean MD, Runyan CM. (2000) *Variations in the relative speeds of orofacial structures with stuttering severity*. J Speech Lang Hear Res. 2000 Dec;43(6):1524-31.

NeyWong M., Murdoch B. E., and Whelan B.-M; (2011) *Lingual Kinematics in Dysarthric and Nondysarthric Speakers with Parkinson's Disease.* SAGE-Hindawi Access to Research Parkinson's Disease Volume 2011, Article ID 352838, 1-8.

Rong, P. Y., Loucks, T. M., Kim, H. J. & Hasegawa-Johnson, M. (2012). *Assessment of tongue-jaw coordination in spastic dysarthria using simultaneous EMA and EMG recordings.* Clinical Linguistics and Phonetics*,* 26(9), 806-22.

Rudzicz F., Hirst G., Van Lieshout P., Penn G., Shein F., Namasivayam A., Wolff T., (2008) (*Towards a Comparative Database of Dysarthric Articulation* da F. Rudzicz, A. K. Namasivayam e T. Wolff; *TORGO Database of Dysarthric Articulation*, International Seminar on Speech Production, 285-288.

Schulz G.M., Hahn J., Jin G., Kiraly J. e Carstens B. e B. (2006) *Translation Of 3-D Articulatory Signals Acquired By Electromagnetic Articulography To A Visual Display Of Lingual Movements For Biofeedback: Preliminary Results, Presentation during,* Motor speech conference, 2006.

Van Lieshout, P. H. H. M., Alfonso, P. J., Hulstijn, W., Peters, H. F. M. (1993) *Electromagnetic articulography (EMA) in stuttering research*. Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München (FIPKM) 31: 215-224.

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 stutter: A motor skill approach.* In Maassen B., Kent R., Peters H., van Lieshout P. H. H. M. & Hulstijn W. (eds.) *Speech Motor Control in Normal and Disordered Speech.* Oxford: OUP:313-355.

Wong M. N., Murdoch B. E., and Whelan B.-M. (2010), *Kinematic analysis of lingual function in dysarthric speakers with Parkinson's disease: an electromagnetic articulograph study,* International Journal of Speech-Language Pathology, vol. 12, no. 5, pp. 414–425.

Wong M. N., Murdoch B. E., and Whelan B.-M., (2010)*Tongue function in nondysarthric speakers with Parkinson's disease: an electromagnetic articulography investigation,* Journal of 8 Parkinson's Disease Medical Speech-Language Pathology, vol. 18, no. 3, pp. 24–33.

Zmarich C., Magno Caldognetto E., Vagges K., (1994) *Articulatory kinematics of lips and jaw in repeated /pa/ and /ba/ sequences in italia stutterers*, in Proceedings of the First World Congress on Fluency Disorders, Munich, v 1, 43-47.

Zmarich C., Magno Caldognetto E., (1997) "Analysis of lips and jaw multi-peaked velocity curve profiles in the fluent speech of stutterers and nonstutterers", in W.Hulstijn, H.F.M. Peters & P.H.H.M. van Lieshout (Eds.), Speech Production: Motor Control, Brain Research and Fluency Disorders, Elsevier Publisher, Amsterdam, 177-182.

Zmarich C., Danelon L., Lonardi F. (2005), L'indice spazio-temporale (STI): un nuovo strumento per valutare la stabilità articolatoria nel parlato, in P. Cosi (a cura di), Misura dei parametri, Atti del 1° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce (AISV), Padova, 2-4 dicembre, 2004, EDK, Brescia, 377-388.

Zmarich C., Marchiori M., (2006) *Coarticulation and stuttering in fluent syllables under contrastive focus*, 5th International Conference on Speech Motor Control, Nijmegen (NL), June 7 - 10, STEM-, SPRAAK- EN TAALPATHOLOGIE, jaargang 14, Supplement, juni, p. 103.

Zmarich C., Magno Caldognetto E., Vagges K., (1994) *La balbuzie come disturbo della produzione articolatoria*, Acta Phoniatrica Latina, 16, 157-183.

Zmarich C., (1999) L'importanza dell'analisi cinematica: esemplificazioni relative alla balbuzie, in A. Tronconi (a cura di), *Atti del 6° Convegno Nazionale Informatica, Didattica & Disabilità*, Andria (Bari), 101-106.

Percorsi didattici multi-sensoriali per avvicinare i bambini dislessici ai suoni e alle lettere della lingua inglese lingua straniera

Verusca Costenaro; Luciana Favaro
Università Ca' Foscari Venezia

Il presente contributo intende illustrare, sotto forma di poster, i materiali glottodidattici elaborati dal team del Progetto DEAL (Dislessia Evolutiva e Apprendimento delle Lingue) del Centro di Didattica delle Lingue di Ca' Foscari in collaborazione con Oxford University Press (in corso di pubblicazione, ad integrazione del manuale di testo New Treetops 1°). Tali materiali vengono proposti come una risorsa per l'insegnante della scuola primaria per facilitare l'acquisizione della lingua inglese da parte di bambini dislessici o potenziali tali di prima elementare, e in particolar modo per avvicinare i bambini ai suoni e alle lettere della lingua inglese. Per il loro contenuto e la loro impostazione, tuttavia, tali materiali si prestano ad essere impiegati con tutta la classe, ed integrati nel normale svolgimento delle attività in lingua inglese.

Una proposta di lavoro a livello fonemico si traduce in uno strumento cruciale per un bambino dislessico. Secondo la teoria prevalente sull'origine della dislessia, un bambino dislessico presenta infatti un deficit di tipo fonologico (Snowling, 1987), che si manifesta in scarse abilità (meta)fonologiche, intese come capacità di riconoscere e manipolare i suoni del flusso orale (Blachman, 1994). La mancanza di una solida base a livello di abilità (meta)cognitive si traduce, nel momento dell'incontro con il codice scritto, in una difficoltà nella fase di decodifica (conversione delle lettere in suoni). Riuscire a percepire e pronunciare i suoni in maniera corretta diventa dunque di cruciale importanza a vari livelli: in primo luogo, nell'ambito di acquisizione del codice orale della lingua straniera, permette al bambino di instaurare scambi comunicativi chiari ed efficaci – di comprendere e farsi comprendere nell'interazione comunicativa. In secondo luogo, riuscire a sentire e pronunciare i suoni in maniera corretta rappresenta la base per l'attività di decodifica in fase di lettura (e di conseguenza di codifica nella fase di scrittura). La lettura, intesa come decodifica dei simboli grafici, richiede infatti una buona capacità di analisi fonologica, che permette ad esempio di dividere le parole in sillabe, riconoscere parole che rimano fra loro, o identificare suoni simili o diversi all'interno di parole (Adams, 1990). Da qui l'importanza di impostare un percorso glottodidattico per bambini dislessici sul livello fonemico delle abilità (meta)fonologiche. A supporto di tale esigenza metodologica riveste un ruolo cruciale la struttura stessa della lingua inglese. La lingua inglese, infatti, con il suo sistema di suoni diversi o inesistenti rispetto al sistema fonologico italiano, e con il suo sistema di scrittura opaco rispetto a quello trasparente dell'italiano, pone una serie di barriere aggiuntive al bambino che ha difficoltà di analisi fonologica. In tale prospettiva, un lavoro mirato ed esplicito sui suoni dell'inglese e le loro realizzazioni ortografiche potrà rivelarsi un'ottima risorsa per il bambino dislessico, non solo per acquisire maggiore consapevolezza della struttura sonora delle parole, ma anche quando si troverà ad ascoltare o leggere un brano, a parlare o a scrivere (Costenaro, Daloiso, Favaro, 2013).

All'interno del poster in riferimento, si intendono dunque illustrare i percorsi didattici in fase di pubblicazione per la Oxford University Press. Si tratta di otto percorsi sonori che introducono in ogni unità, se si esclude la prima incentrata solo sulla lettera h, una coppia di fonemi della lingua inglese particolarmente difficili da riconoscere e articolare per il bambino dislessico (come ad esempio le coppie /θ/-/f/, /p/-/b/, e /æ/-/ʌ/). Ogni percorso è composto da una scheda-guida per l'insegnante, in cui vengono illustrati gli obiettivi d'apprendimento e le indicazioni per svolgere le attività didattiche, una serie di registrazioni contenute in un CD audio, e 2 schede didattiche fotocopiabili per i bambini. Ogni unità è incentrata su due personaggi-animali che fanno parte dell'ambiente in cui si svolgono le storie del libro di testo, e rappresentano i suoni in riferimento (ad esempio il serpentello *Thumby* porta con sé il suono /θ/, mentre la ranocchietta *Froggy* porta con sé il suono /f/). Ogni percorso didattico è suddiviso in cinque fasi di lavoro, che comprendono la presentazione dei suoni da parte dell'insegnante, il riconoscimento e la produzione dei suoni da parte dei bambini, una fase di associazione dei suoni alla loro rappresentazione grafica più frequente, e una fase finale di sintesi multi-sensoriale di quanto appreso nel percorso (Costenaro, Daloiso, Favaro, 2013). L'aspetto innovativo di tali percorsi didattici risiede non solo nei contenuti – essendo la fonetica spesso trascurata all'interno della lezione di lingua inglese – ma soprattutto nella metodologia adottata per facilitare l'acquisizione da parte di bambini dislessici. I percorsi si fondano infatti su un input multimodale, in cui il suono in riferimento viene associato ad altri canali espressivi – come il canale motorio, tattile, iconico, immaginativo, ecc. In tale modo, il canale deficitario nel bambino dislessico – quello uditivo,

legato alla dimensione fonologica della lingua - viene ad essere sostenuto grazie all'attivazione di ulteriori canali compensatori non deficitari. La presentazione dell'input sonoro attraverso canali sensoriali aggiuntivi costituisce uno strumento prezioso in grado di sostenere e facilitare il processo di acquisizione dell'input stesso (Nijakowska, 2010). Un ulteriore punto di forza e innovazione riguarda la contestualizzazione narrativa di tali percorsi. L'aggancio immaginativo-narrativo permette infatti di inserire un input formale come quello fonologico all'interno di un contesto ludico e piacevole, noto e vicino al mondo del bambino. Tale aggancio affettivo si rivela un fattore metodologico fondamentale per favorire l'instaurarsi di un ambiente di apprendimento sereno e motivante (Daloiso, 2012).

Riferimenti

ADAMS, M.J., 1990, *Beginning to read: Thinking and learning about print*, MIT, Cambridge

BLACHMAN, B.A., 1994, "Early literacy acquisition: The role of phonological awareness.", in Wallach, G.P., Butler, K.G. (a cura di), *Language Learning Disabilities in School-Age Children and Adolescent: Some Principles and Applications*, Macmillan, New York, 253-274.

COSTENARO V., DALOISO M., FAVARO L., 2013, *New Treetops e la dislessia. Risorse didattiche*, Oxford University Press, Oxford.

DALOISO M., 2012, *Lingue straniere e dislessia evolutiva. Teoria e metodologia per una glottodidattica accessibile*, Utet Università, Torino.

NIJAKOWSKA J., 2010, *Dyslexia in the Foreign Language Classroom*, Multilingual Matters, Bristol.

SNOWLING M., 1987, *Dyslexia. A Cognitive Developmental Perspective*, Balckwell, Oxford.

SILLABE FONETICHE APPLICATE AL RICONOSCIMENTO DI EMOZIONI

Antonio Origlia, Francesco Cutugno, Vincenzo Galatà

antonio.origlia@unina.it, cutugno@na.infn.it, vincenzo.galata@pd.istc.cnr.it

Esistono molti problemi nella ricerca sulle emozioni che risultano ancora aperti, come è stato sottolineato recentemente in Schuller et al. (2011). Per quanto riguarda l'estrazione del contenuto emotivo dalle sole proprietà acustiche della voce umana, due punti sembrano dominare il dibattito. Il primo è rappresentato dalla necessità di stabilire una rappresentazione condivisa delle emozioni in termini del modo in cui dovrebbero essere raccolte ed annotate. Il secondo è legato all'identificazione del frammento di analisi più piccolo al quale bisogna far riferimento per le procedure di analisi. Da questo problema dipende direttamente la possibilità di realizzare sistemi che operino in tempo reale piuttosto che aspettando il completamento della produzione.
Per quanto riguarda il primo punto, la ricerca sulle emozioni si è spostata dalla classificazione discreta in categorie, tipicamente il modello a sei classi di Ekman (1992), usato per etichettare corpora come EMO-DB (Burkhardt et al., 2005) ad una rappresentazione più dinamica usando spazi multidimensionali, dove si considerano le componenti delle emozioni piuttosto che i termini emotivi, utilizzata per etichettare corpora come VAM (Grimm et al., 2008), utilizzato per gli esperimenti presentati in questo lavoro.

Per quanto riguarda, invece, il secondo punto, mentre la ricerca sull'estrazione delle proprietà acustiche del parlato per il riconoscimento di emozioni progredisce, il bisogno di studiare metodi di estrazione di features che tengano conto delle necessità di sistemi di analisi che lavorino in tempo reale diventa più importante. In letteratura si è potuto osservare come, tra le altre unità di analisi, la sillaba sia risultata essere importante per la trasmissione di emozioni mentre gli studi classici relativi alla prosodia mostrano che è importante concentrarsi su aree specifiche del parlato per studiare fenomeni intonativi. Gli approcci tecnologici, tuttavia, sono spesso progettati per far uso dell'intera produzione vocale senza tenere presente la variabilità qualitativa del contenuto spettrale. Dato questo contrasto tra la base teorica sulla quale viene condotta la ricerca prosodica, presentiamo un metodo di estrazione di features basato su una interpretazione fonetica del concetto di sillaba. Indichiamo questa unità come sillaba fonetica (talvolta indicata come pseudosillaba) seguendo la terminologia usata in D'Alessandro (1995), che definiva questa unità come *"*[...] *a continuous voiced segment of speech organized around one local loudness peak, and possibly preceeded and/or followed by voiceless segments"*. La definizione che tuttavia si adatta meglio ai segmenti che vengono individuati automaticamente dall'algoritmo impiegato è quella riportata in Roach (2000) che descrive la sillaba fonetica come *"*[...] *consisting of a centre which has little or no obstruction to airflow and which sounds comparatively loud; before and after that centre* [...] *there will be greater obstruction to airflow and/or less loud sound"*. In particolare, ci concentriamo sul contenuto spettrale dei nuclei sillabici, riducendo la quantità di informazione da analizzare, introduciamo un parametro per la descrizione dei movimenti di pitch attraverso il nucleo sillabico in termini di occorrenza di toni dinamici (glissando), features segmentali relative alla velocità di eloquio ed alla durata dei segmenti e introduciamo una pesatura delle features ispirata al concetto di prominenza sillabica, evitando di considerare tutte le unità come ugualmente importanti. Oltre a ciò, viene studiato l'impatto dell'introduzione di features relative a parametri ritmici del parlato. Mentre l'utilità di queste misure è stata recentemente messa in discussione per quanto riguarda la loro capacità di separare le lingue, obiettivo per il quale erano state sviluppate, essere rappresentano un valido descrittore dello stile usato in una determinata produzione vocale e sono pertanto adatte al compito in questione.

Utilizzando come baseline i risultati ottenuti da Wu et al. (2011), valutiamo il nostro approccio su

un modello continuo, tridimensionale, delle emozioni che comprende gli assi relativi a Valenza, Attivazione e Dominanza (Grimm & Kroschel, 2005). Per poter utilizzare lo stesso classificatore dell'approccio di riferimento, le Support Vector Machines (SVM), i parametri estratti da ogni sillaba fonetica vengono riassunti in features contenenti statistiche globali. Tali statistiche, tuttavia, sono calcolate sfruttando la durata normalizzata dei nuclei sillabici come pesi. In questo modo, il parametro tipicamente riconosciuto come particolarmente importante nella percezione della prominenza sillabica, la durata dei nuclei, viene introdotto come elemento di distinzione tra le unità che non contribuiscono, quindi, in maniera uniforme alla definizione delle features globali. Le prestazioni ottenute risultano competitive con lo stato dell'arte pur limitando l'estrazione delle caratteristiche spettrali del parlato ad una porzione inferiore del 40% rispetto a quella di riferimento, che invece impiega l'intera produzione. L'impatto potenziale di questo approccio sulla progettazione di sistemi artificiali affettivi viene inoltre presentato insieme ad una analisi qualitativa delle features utilizzate in termini di correlazione con gli assi del modello tridimensionale ed in termini di intercorrelazione.

Bibliografia

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., Weiss, B., 2005. A database of german emotional speech. In Proc. of Interspeech, pp. 1517--1520.

D'Alessandro, C., Mertens, P., 1995. Automatic pitch contour stylization using a model of tonal perception. Computer Speech and Language 9 (3), pp. 257--288.

Ekman, P., 1992. An argument for basic emotions. Cognition and Emotion, pp. 169--200.

Grimm, M., Kroschel, K., 2005. Emotion estimation in speech using a 3D emotion space concept. In Proc. of IEEE Automatic Speech Recognition & Understanding Workshop, pp. 381--385.

Grimm, M., Kroschel, K., Narayanan, S., 2008. The Vera Am Mittag german audio-visual emotional speech database. In Proc. of ICME, pp. 865--868.

Roach, P., 2000. English Phonetics and Phonology. A Practical Course. CUP.

Schuller, B., Batliner, A., Steidl, S., D., S., 2011. Recognising realistic emotions and aff ect in speech: State of the art and lessons learnt from the first challenge. Speech Communication, pp. 1062-1087.

Wu, S., Falk, T. H., Chan, W., 2011. Automatic speech emotion recognition using modulation spectral features. Speech Communication 53, pp. 768--785.

APPLICAZIONE DI TECNICHE DI CANTO DIFONICO ALL'ANALISI ACUSTICA DELLE FORMANTI NASALI

Antonio Romano [1] & Danilo Pastore[2]
[1,2] Università degli Studi di Torino, [1] Laboratorio di Fonetica Sperimentale "Arturo Genre" – Dip. di Lingue e Lett. Str. e Cult. Mod.
[1]*antonio.romano@unito.it*

RIASSUNTO

In questo contributo ci proponiamo d'illustrare un esperimento da noi condotto per individuare con una certa approssimazione le formanti nasali di una data voce e verificarne le modalità d'interazione con le formanti orali di suoni orali nasalizzati.

L'esperimento poggia sull'assunzione implicita di un modello di produzione del parlato di tipo Sorgente-Filtro (v. Fant 1960) e consiste nel ricorso a un sofisticato controllo dell'articolazione di una nasale uvulare durante la produzione di toni gravi e costanti (il bordone di alcune voci usate nel canto difonico, v. Cosi & Tisato 2003) e nel successivo passaggio ad articolazioni vocaliche con e senza accoppiamento acustico.

In queste circostanze si verifica un insieme di condizioni ideali ai fini dell'individuazione delle risonanze nasali (senza interferenze da parte delle cavità orali e con uno spettro armonico molto denso come sorgente) e delle modalità di sovrapposizione tra la funzione di trasferimento di queste e quella del condotto orale (nelle sue diverse configurazioni).

Nel caso di suoni vocalici, infatti, l'attivazione supplementare (da parte del velo palatino) delle cavità nasali (le cui caratteristiche fisiche sono note in particolare sin da Bjuggren & Fant 1964) conduce a condizioni di accoppiamento acustico col condotto vocale. Questo produce di solito l'apparizione – nello spettro d'energia dei suoni relativi a queste articolazioni – delle formanti relative alla risonanza delle cavità nasali, sovrapposte – in misura più o meno consistente e più o meno variabile – con quelle della cavità orale variamente deformata (v. Romano *et alii* 2005 e bibliografia ivi citata). Tuttavia, anche le caratteristiche di risonanza delle cavità nasali e del condotto rino-faringeo dipendono da quest'accoppiamento (v. Ladefoged & Maddieson 1996; cfr. anche Ferrero *et alii* 1979): il contributo acustico del condotto orale si manifesterebbe con l'aggiunta di un'antirisonanza, cioè uno zero spettrale (individuato come minimo picco negativo tra F_1 e F_2) la cui frequenza aumenterebbe con l'arretramento dell'articolazione orale nell'area uvulare (Fujimura 1962).

Quanto alle consonanti nasali, per le quali il cavo orale si presenta chiuso in qualche punto, si è invece in presenza di risonanze che s'instaurano in una serie di cavità modellizzabili come un tubo aperto a un'estremità nel quale s'inserisce un tubo chiuso di lunghezza variabile in base al punto di occlusione nel condotto orale (Ohala & Ohala 1993). L'ispezione informale degli spettrogrammi di suoni prodotti in queste condizioni conferma la presenza di *pattern* formantici più o meno differenziati per le diverse varianti combinatorie (Recasens 1983, Kurowski & Blumstein 1987, ma già Ferrero *et alii* 1979), ma soprattutto condizioni di transizione fortemente disturbate da modalità di coarticolazione molto *speaker-dependent* (riferimenti in Romano *et alii* 2005).

L'arretramento dell'articolazione orale nell'area uvulare consente di annullare (o, comunque, minimizzare) l'effetto di questa cavità collaterale, stabilendo le condizioni

ottimali per modellizzare il percorso dalla laringe alle narici come un tubo aperto a un'estremità. Sebbene in questi casi si sia in presenza di una cavità tutt'altro che uniforme (e rettilinea), questo tubo può approssimarne bene le condizioni di risonanza (Fujimura 1962). Assumendo per questo tubo una lunghezza media di 23,5 cm (per una voce maschile; v. anche Bjuggren & Fant 1964) possiamo prevedere che nello spettro di un suono di tipo [N] si presentino risonanze con le seguenti frequenze: $F_{N1} \cong 364$ Hz, $F_{N2} \cong 1091$ Hz, $F_{N3} \cong 1819$ Hz, $F_{N4} \cong 2758$ Hz etc.

Ovviamente questi valori variano nel caso delle altre consonanti nasali e si modificano considerevolmente nel passaggio da suoni come questi ai suoni orali contigui (i quali possono essere soggetti a vari gradi di nasalizzazione, con effetti che interessano soprattutto F_3, ma si manifestano, per certe regioni articolatorie, anche per F_2).

In uno studio spettrografico i valori delle frequenze di risonanza nasale sono tanto meglio verificabili quanto più la f_0 dello stimolo è grave e costante e quanto meglio il parlante riesce ad accoppiare o disaccoppiare le cavità nelle diverse fasi della produzione.

Alcune di queste condizioni si possono presentare nel canto difonico (Tisato & Ricci Maccarini 1991, Bloothooft *et alii* 1992, Cosi & Tisato 2003).

Nell'esperimento da noi condotto, due cantanti addestrati alla produzione di voce difonica hanno pronunciato cinque sequenze ininterrotte di circa 10 secondi con articolazione consonantica sostenuta di tipo [N] e con rilascio dorso-uvulare ogni secondo circa al momento dell'impostazione di una delle 5 vocali /i ɛ a ɔ u/ di durata 0,8 s circa (con rese fonetiche risultanti di tipo [nːĩˈnːɛ̃ˈnːãˈnːɔ̃ˈnːũˈ]).

La frequenza fondamentale è stata tenuta costante esattamente a 119 Hz dal primo cantante (Sib_2) e a 73 Hz dal secondo (Re_2). L'intensità delle produzioni di entrambi non è mai scesa sotto i 55 dB e si è presentata particolarmente alta (>70dB) solo nel corso delle realizzazioni vocaliche.

Il primo cantante è l'unico per il quale è stato possibile ottenere anche una ripresa filmata dei movimenti labiali e completare finora tutte le misurazioni relative.

Nelle ultime due ripetizioni della sequenza, il cantante ha articolato le vocali senza accoppiare le cavità (mantenendo la chiusura postdorso-uvulare). Questa condizione è stata usata come verifica dell'assoluta mancanza d'interferenza delle diverse articolazioni orali assunte di volta in volta (e ben visibili nel filmato) sul suono in uscita. Tale suono è quindi da ritenersi prodotto con risonanze delle sole cavità rino-faringee.

Le formanti misurate in queste condizioni (nelle quali si possono presentare altri effetti acustici, alcuni dei quali sono studiati in Ferrero *et alii* 1980) assumono valori che non dipendono dal contesto vocalico ma che confermano solo parzialmente le attese: F_{N1} = 277-358 Hz, F_{N2} = 1404-1502 Hz, F_{N3} = 3070-3333 Hz, F_{N4} = 3855-4018 Hz etc.

Possiamo ipotizzare che le formanti abbiano risentito di un innalzamento causato da un restringimento faringeo basso tipico di queste voci (come illustrato da Cosi & Tisato, 2003), ma questo non si concilia con i bassi valori presentati proprio da F_{N1}. Inoltre, i valori non presentano l'equidistanza tipica delle risonanze delle onde stazionarie di tubi rettilinei uniformi (tra gli altri, Fant 1960). Da un lato questo risultato fa quindi propendere per forti elementi di criticità nei confronti del modello (almeno per queste voci e queste modalità di produzione), ma dall'altro – in virtù delle condizioni di persistenza dei contributi formantici nasali nell'arco delle produzioni orali (nasalizzate) che sono state intervallate – permette di ritenere molto attendibili i valori misurati sul piano della definizione delle caratteristiche di nasalità "assoluta" delle voci analizzate.

BIBLIOGRAFIA

Bjuggren, G. & Fant, G. (1964). The Nasal Cavity Structures. *STL-QPSR, KTH*, 4, 5-7.

Bloothooft, G., Bringmann, E., Van Capellen, M., Van Luipen, J.B. & Thomassen, K.P. (1992). Acoustic and Perception of Overtone Singing. *JASA*, 92/4, Part 1, 1827-1836.

Cosi, P. & Tisato, G. (2003). On the magic of overtone singing. In P. Cosi, E. Magno Caldognetto & A. Zamboni (eds.), *Voce, Canto, Parlato. Studi in onore di Franco Ferrero*, Padova: Unipress, 83-100.

Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton.

Ferrero, F., Genre, A., Boë, L.J. & Contini, M. (1979). *Nozioni di Fonetica Acustica*. Torino: Omega.

Ferrero, F., Croatto, L. & Accordi, M. (1980). Descrizione elettroacustica di alcuni tipi di vocalizzo di Demetrio Stratos. *Rivista Italiana di Acustica*, IV/3, 229-258.

Fujimura, O. (1962). Analysis of nasal consonants. *JASA*, 34, 1865-1975.

Kurowski, K. & Blumstein, S.E. (1987). Acoustic properties for place of articulation in nasal consonants. *JASA*, 81/6, 1917-1927.

Ladefoged, P. & Maddieson, I. (1996). *Sounds of the World's Languages*. Oxford: Blackwell.

Ohala, J.J. & Ohala, M. (1993). The Phonetics of Nasal Phonology: Theorems and Data, in M.K. Huffman & R.A. Krakow (eds.) *Phonetics and Phonology, vol. 5 - Nasals, Nasalization, and the Velum*, San Diego: Academic Press, 225-249.

Recasens, D. (1983). Place Cues for Nasal Consonants with special reference to Catalan. *JASA*, 73, 1346-1353.

Romano, A., Mancini, F. & Zovato, E. (2005). Nasali eterosillabiche in italiano e spagnolo: l'energia di banda come parametro discriminante nella classificazione dei nessi NC. In P. Cosi (ed.), *La misura dei parametri: Aspetti tecnologici ed implicazioni nei modelli linguistici* (Atti del I Conv. Naz. AISV, Padova, 2004), Padova: ISTC/EDK, 101-133.

Tisato, G. & Ricci Maccarini, A. (1991). Analysis and synthesis of Diphonic Singing. *Bulletin d'Audiophonologie*, 7/5-6, 619-648.

Uso del crowdsourcing per trascrizioni di alta qualità del linguaggio parlato: metodologie a confronto - Sprugnoli, R. et alii

Si è recentemente affermato l'uso di piattaforme di crowdsourcing per svolgere vari compiti collegati al trattamento automatico della lingua, tra cui la creazione di corpora di parlato trascritto (si vedano tra gli altri i lavori di Novotney e Callison-Burch (2010), Merge et al. (2010), Parent e Eskenazi (2010), Audhkhasi et al. (2011)) che sono risorse fondamentali per lo sviluppo e la valutazione delle tecnologie ASR. Attraverso queste piattaforme, si ricorre al contributo di un vasto ed indefinito gruppo di persone, non necessariamente esperte di una certa materia, per risolvere un problema. Molti studi hanno mostrato come il crowdsourcing possa ridurre i tempi ed i costi di un lavoro lungo e complesso come quello della trascrizione ma anche che il punto più critico di tale approccio riguarda il garantire l'alta qualità dei dati raccolti applicando degli adeguati meccanismi di controllo.

Il presente contributo vuole descrivere gli esperimenti di crowdsourcing svolti nell'ambito del progetto europeo TOSCA-MP (*Task-oriented search and content annotation for media production*, http://tosca-mp.eu/), che ha come obbiettivo quello di sviluppare tecnologie innovative per la ricerca di informazioni multimediali nell'ambito della produzione di contenuti giornalistici per televisione, radio e Web. Trovare informazioni rilevanti nel parlato è un compito impegnativo nel quale i sistemi di Automatic Speech Recognition (ASR) combinati alle tecniche di Information Retrieval giocano un ruolo chiave. Più specificamente, abbiamo valutato *due diverse metodologie* di crowdsourcing al fine di selezionare il metodo migliore in termini di (i) qualità di trascrizione, (ii) costo e (iii) tempo di raccolta dati per più lingue europee tra cui italiano e tedesco.

Gli esperimenti descritti sono stati realizzati attraverso la piattaforma di crowdsourcing Amazon Mechanical Turk (AMT, www.mturk.com). Poiché l'accesso diretto ad AMT è consentito solo a committenti residenti negli Stati Uniti, per poter utilizzare AMT ci siamo avvalsi dei servizi di intermediazione offerti da CrowdFlower (CF, www.crowdflower.com). Oltre a consentire l'accesso ad AMT, CF mette anche a disposizione un meccanismo di controllo della qualità basato su un data set di riferimento (*Gold Standard*). L'esperimento è stato condotto su italiano e tedesco utilizzando per ciascuna delle due lingue 30 minuti di audio estratti da telegiornali. I due metodi di crowdsourcing messi a confronto differiscono nel meccanismo di controllo della qualità: ciò consente dunque di poter investigare al meglio questo aspetto cruciale nell'acquisizione di dati attraverso crowdsourcing.

Il primo metodo oggetto del presente contributo sfrutta direttamente l'interfaccia ed il meccanismo di controllo della qualità offerti da CrowdFlower. Questa modalità ha richiesto che almeno il 10% dei segmenti audio venissero preventivamente trascritti da due esperti, in modo da produrre un gold standard per il task oggetto del nostro studio. Un esperto ha inoltre prodotto un certo numero di annotazioni volutamente sbagliate. Questi dati sono quindi stati usati per valutare l'affidabilità dei lavoratori a cui è stato chiesto se, ascoltando una clip audio, la trascrizione ad essa associata fosse corretta o meno. Grazie al meccanismo di controllo di CF, i lavoratori che non hanno fornito il giudizio corretto per almeno il 70% delle trascrizioni di riferimento vengono automaticamente esclusi e solo le trascrizioni di lavoratori affidabili sono restituite al committente. Nel nostro esperimento abbiamo richiesto che ogni clip audio venisse trascritta cinque volte (da lavoratori diversi) e alla fine tutte le trascrizioni ottenute per ogni clip sono state unite usando il ROVER (Fiscus 1997).

Il secondo dei metodi di controllo della qualità testati, denominato *metodo iterativo a doppia pipeline* (Liem et al., 2011), è caratterizzato dal fatto che non richiede la presenza di un gold standard prodotto da esperti. Il metodo prevede che le trascrizioni vengano iterativamente migliorate da due gruppi indipendenti di lavoratori fino a che le trascrizioni prodotte da ciascun gruppo non convergano. L'ipotesi alla base di questa metodologia è che, poiché i percorsi di trascrizione sono indipendenti, la convergenza tra i due percorsi garantisca la qualità della trascrizione. In questo modo, grazie ai cicli di revisioni successivi ed iterativi, non è necessario utilizzare il gold standard. Per implementare questo metodo è stata creata un'apposita infrastruttura su database ed un'interfaccia web grafica, accessibile ai lavoratori di CF tramite un link.

I risultati ottenuti dagli esperimenti si sono dimostrati ottimi, in linea con la percentuale di disagreement tra esperti. In particolare, il Word Error Rate (WER) delle trascrizioni dei 30 minuti in lingua tedesca ottenuto con il metodo iterativo a doppia pipeline si è attestato al 4,67% mentre con il metodo basato sul gold standard al 4,14%. È stato perciò registrato un miglioramento di più di 12 punti percentuali rispetto alla corrispondente trascrizione automatica che ha un WER del 17,10%. Per l'italiano, invece, il WER col metodo pipeline è stato del 3,41% e col metodo gold standard del 3,12%: in questo caso, quindi, il miglioramento rispetto alla trascrizione automatica, con un WER del 10,42%, è stato di più di 7 punti percentuali.

In conclusione, con questo lavoro si vuole contribuire al progresso della ricerca sulle tecniche di crowdsourcing nell'ambito dello speech processing a) implementando e valutando il metodo iterativo a doppia pipeline sul compito di trascrizione di file audio usando per la prima volta la piattaforma di AMT; b) valutando la fattibilità del crowdsourcing per la raccolta di trascrizioni di audio in lingue diverse dall'inglese.

Bibliografia

S. Novotney and C. Callison-Burch, Cheap, Fast and Good Enough: Automatic Speech Recognition with Non-Expert Transcription. In Proceedings of HLT-NAACL. 2010, 207-215.

M. Marge, S. Banerjee, and A.I. Rudnicky, Using the Amazon Mechanical Turk for transcription of spoken language. In Proceedings of ICASSP. 2010, 5270-5273.

G. Parent and M. Eskenazi, Toward better crowdsourced transcription: Transcription of a year of the Let's Go Bus Information System data. In Proceedings of SLT. 2010, 312-317.

K. Audhkhasi, P.G. Georgiou, and S.S. Narayanan, Accurate transcription of broadcast news speech using multiple noisy transcribers and unsupervised reliability metrics. In Proceedings of ICASSP. 2011, 4980-4983.

B. Liem, H. Zhang, and Y. Chen, An Iterative Dual Pathway Structure for Speech-to-Text Transcription. In Proceedings of Human Computation. 2011.

J. Fiscus, A post-processing system to yield reduced error rates: recognizer output voting error reduction (ROVER). In Proceedings of IEEE ASRU workshop. 1997.

Singing in German: text-setting rules and language rhythm

Teresa Proto

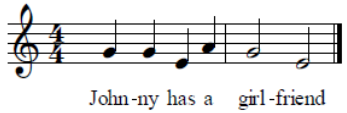
Abstract

Vocal music varies across the world. One of the main sources of difference is of course language: songs are sung in different languages. This is obvious. What is not obvious, however, are the ways in which languages constrain the setting of a text to a tune.

It is generally assumed that the alignment of a text to a tune is not a random process, but one that is governed by a set of rules that may vary from language to language, and possibly from one singing idiom to another within the same language (Dell/Halle 2009). As a matter of fact, when lyrics are set to music, syllables are assigned to musical pitches in such a way as to conform to specific requirements of the language. Some of these requirements are universal, while others are language-specific. Among the universal prerequisites, the most basic one is that each syllable must be matched to at least one musical pitch: no "floating syllables" are allowed. Language-specific requirements depend on the phonetic, phonological and syntactic properties of the language in which the lyrics are composed. Both sets of requirements contribute to the well-formedness of musical settings; when either or both are violated, the resulting settings are rejected as ill-formed (or awkward) by the participants of that singing tradition. This is because the latter have internalized a system of tacit principles and rules that regulate the occurrence in singing of violations to the grammar of the language.

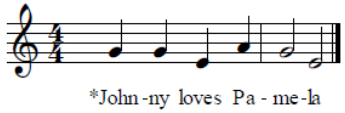
One of the most studied text-setting practices is the English one. The so-called *stress-to-beat matching* rule has been established as a major constraint in this language through a number of works focused on English folksongs (Halle & Lerdahl 1993, Hayes & Kaun 1996, Dell & Halle 2009, Hayes 2009, Liberman 1975, Rodríguez-Vázquez 2010). This rule states that for a language like Present-day English, setting a text to music basically implies assigning prominent syllables in words to strong beats in music. This is shown in the following example of children's chant (Ladd 2008: 57):

(1)



As Liberman points out, ill-formed associations of texts to this tune must be defined in terms of the position of stressed syllables relative to the metrically strong positions in the tune (the notes immediately preceded by bar lines). An example of ill-formed association, due to the misalignment of the stressed syllable *Pam-* with respect to the downbeat, is given below:

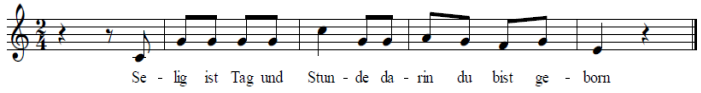
(2)



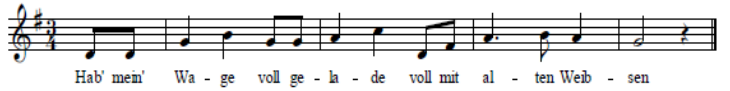
The existence of this constraint in English text-setting has been taken as evidence for strict isochrony in language (Rodríguez-Vázquez 2010). According to the isochrony approach, an equal timing is perceived between stressed syllables in a stress-timed language like English (Pike 1945), reflecting the matching of stressed syllables to strong beats, whereas syllable-timed languages, like Spanish and French, disregard this rule and allow mismatches between prominence in speech and in music.

A preliminary study of the text-to-tune alignment in German, another language traditionally considered as stress-timed, has shown that violations of the stress-to-beat matching are indeed allowed in this language. A survey on a sample of 200 *Volkslieder* has revealed that the stress-to-beat matching principle, although statistically observable in songs, is not an absolute constraint for German. As a matter of fact, this rule can be violated in configurations involving simple as well as compound words, both in line-initial and line-final position, as shown in (3a) and (3b), respectively.

(3a)



(3b)



In (3a) above a mismatch appears at the beginning of the line involving the adjective *selig*. In normal speech this word carries stress on the first syllable; in this setting however, it is the second syllable that is matched to the downbeat. In (3b) the word *Weibsen* at the end of the line carries initial stress, and yet the initial syllable *Weib-* is matched to a weaker position than the following stressless syllable *–sen* (which appears on the downbeat of the final bar).

In the second part of my study I tested the perception of stress-beat mismatches by two German native speakers. The testing revealed that native speakers' perception of this kind of discrepancies relies not only on rhythmical factors, but also on melodic features: besides rhythmical patterns, also differences in pitch interval and duration appear to play a role.

To what extent the stress patterns and intonation contours of the language, on the one hand, and the harmonic structure of the music, on the other hand, affect the perception of mismatches in songs is not known. My research aims at contributing to an answer to this question by providing measurable results for German text-setting.

In particular, it should contribute to the discussion on the very nature of language rhythm and its relation to its cognate in the musical domain, by providing useful insights into how the supra-segmental structures of the language (stress, pitch accent, syllable length) interact with analogous structures in music (downbeats, phenomenal accents, duration).

References

Auer, Peter (2001). 'Silben- und akzentzählende Sprachen'. In Haspelmath, Martin, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible (éd.). *Language Typology and Language Universals. An International Handbook*. Berlin: de Gruyter, 1391-1399.

Dauer, Rebecca. 1987. 'Phonetic and phonological components of language rhythm'. *Proceedings of the XIth International Congress of Phonetic Sciences*. Vol. 5. Tallinn: Académie des Sciences de l'Estonie, 447-450.

Dell, François & John Halle (2009). 'Comparing musical textsetting in French and in English songs'. In Jean-Louis Aroui & Andy Arleo (eds.), *Towards a typology of poetic forms*. Amsterdam: John Benjamins, 63–78.
Halle, John & Fred Lerdahl (1993). 'A Generative Textsetting Model'. *Current Musicology* 55, 3-23.
Hayes, Bruce (2009). 'Textsetting as constraint conflict'. In Jean-Louis Aroui & Andy Arleo (eds.), *Towards a typology of poetic forms*. Amsterdam: John Benjamins, 43–61.
Hayes, Bruce & Abigail Kaun (1996). 'The role of phonological phrasing in sung and chanted verse'. *Linguistic Review* 13, 243–303.
Hannon, E.E., J.S. Snyder, T. Eerola & C.L. Krumhansl. 2004. 'The role of melodic and temporal cues in perceiving musical meter'. *Journal of the Experimental Psychology: Human Perception and Performance* 30, 956-974.
Ladd, Robert D. 2008. *Intonational Phonology*, 2nd edn. Cambridge: CUP.
Lerdahl, Fred & Ray Jackendoff (1983). *A generative theory of tonal music*. Cambridge, Mass.: MIT Press.
Liberman, Mark. 1975. *The intonational system of English*. Cambridge, Mass.: MIT dissertation.
Pike, Kenneth (1945). *The Intonation of American English*. Ann Arbor: University of Michigan Press.
Proto, Teresa & Dell, François (in press). 'The structure of metrical patterns in tunes and in literary verse. Evidence from discrepancies between musical and linguistic rhythm in Italian songs'. *Probus – An International Journal of Latin and Romance Linguistics* (special issue 2012).
Rodríguez-Vázquez, Rosalía. 2010. *The Rhythm of Speech, Verse and Vocal Music: A New Theory*. Bern: Peter Lang.
Szczepaniak, Renata (2007). *Der phonologisch-typologische Wandel des Deutschen von einer Silben- zu einer Wortsprache*. Berlin: de Gruyter.
Wiese, Richard. 2000. *The Phonology of German*. Oxford: Oxford University Press.

Multimodal rhetoric

Verbal, acoustic and body strategies in a Nichi Vendola public speech

Paolo Bravi

The study of rhetoric, seen as the discipline devoted to the *ars bene dicendi* (Quintilianus, Institutio Oratoria), has been historically more concerned with topics related to the invention and construction of the discourse than with the way in which it is delivered to its audience (Plebe & Emanuele, 1988; Mortara Garavelli, 1988). Text has largely overcome pragmatics and description of formal structures of argumentation and phrase has gained more attention than the analysis of performance and of delivery styles (Perelman & Olbrechts-Tyteca, 1966 ed. or. 1958).

However, from the very beginnings of this time-honoured tradition of study there is a clear awareness that the way in which the speaker gives his/her speech is as important as what s/he actually says in terms of words, syntax, discourse strategies. *Hypókrisis, actio, pronuntiatio* are terms that ancient rhetors (Greek and Latin) used to refer to the modulation of voice, to gestures and movements of the speaker (Garver, 1994: Kennedy, 1994).

In this paper, a section of a public speech given by the Italian politician Nichi Vendola has been analysed. Taken from an political meeting held in Milano, 2011, for the electoral campaign in support of Giuliano Pisapia as city's Mayor, the section of the speech is focused on the topic of 'liberty' and is clearly structured as a text with parenetic purpose (see Appendix, fig. 4). The analysis has been carried out on three levels: the verbal, the acoustic and the gestural one. Acoustic and visual data analysis have been performed via the software *Praat* (Boersma & Weenink, 2011) and *Elan* (Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006).

All three levels – verbal, acoustic and gestural – show organizing strategies aiming at developing an atmosphere of enthusiasm, cohesion, determination, tension towards the ideal and the goal. It can be seen that similar patterning expressing stress of meaning and intensification of emotion are present at all levels: on the verbal one, by means of word repetition, anaphoras, enumeration, and other textual devises; on the acoustic one, by means of an appropriate prosodic changes relevant to pitch level, articulation rate, speech fluency); on the gestural one, by means of acceleration and intensification of body movements. Figures 1 to 3 in Appendix, *infra*, show some aspects related to this 'acoustic shape' of the discourse.

Seen as a three-faceted discipline focusing the multimodal structures of speech delivery, the old discipline of rhetoric appears like a very promising field for research based on instrumental means. In this perspective, the *ars bene dicendi* seems to share some of its features with different kinds of communication and interpersonal relations and in particular with musical performances, which are known for their use of strategies similar to that of speech to gain attention and to create emotional involvement in the listeners and among musicians themselves (Meyer, 1956; Imberty, 1986).

Future interdisciplinary work on this area is foreseeable with the aim of identifying common strategies in distinct fields of human expression and communication, and particularly in linguistic and musical performative acts, as well as pinpointing their relevant similarities and differences (Patel, 2008).

Appendix

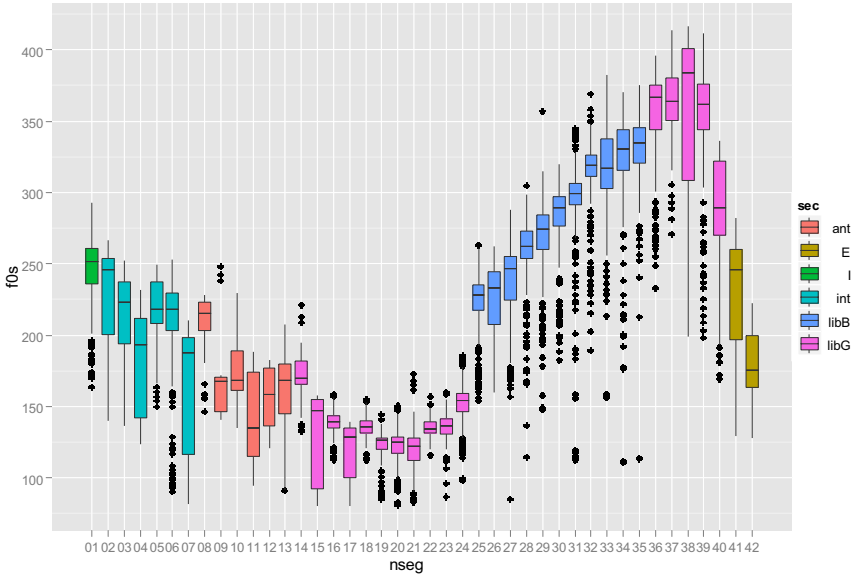


Figure 1. Pitch distributions in the 42 IPS (inter-pause-stretch) comprising the "liberty speech" part of Vendola's rally, divided according to the sections of the speech (see Fig. 4).

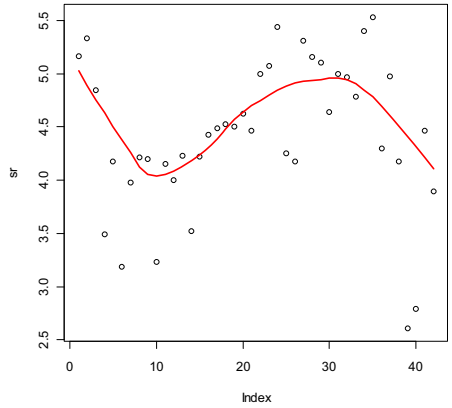


Figure 2. Articulation rate over time (see Goldman-Eisler, 1968; Trouvain, Koreman, Erriquez, & Braun, 2001)

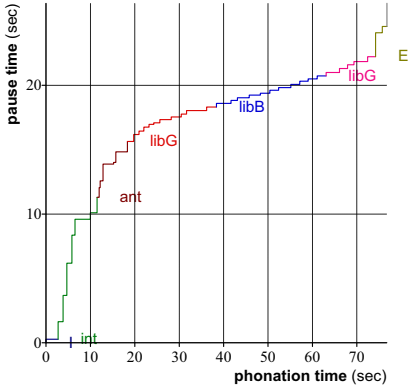


Figure 3. Henderson graph showing evolution over time of speech fluency (Henderson, Goldman-Eisler, & Skarbek, 1966).

N: 1 Dobbiamo riprenderci la parola libertà
N: 2 per una ragione
N: 3 la libertà / N: 4 dei moderni / N: 5 è nata / N: 6 in contrapposizione a un mondo di ingiustizie / N: 7 e di oscurantismo
N: 8 [Era] / N: 9 lo / N: 10 voglio / N: 11 ripetere come un ritornello / N: 12 perché / N: 13 possiamo proprio scolpircelo in testa
N: 14 era libertà dalla miseria
N: 15 era libertà dalla paura
N: 16 era libertà dalla precarietà
N: 17 era libertà dalla superstizione
N: 18 era libertà dall'ignoranza
N: 19 era la libertà che una giovane generazione poteva avere di spalancare le porte per guardare l'orizzonte del futuro
N: 20 la vostra libertà è un circuito commerciale
N: 21 è un supermercato
N: 22 è una sequenza di canali pubblicitari
N: 23 è una libertà artificiale di plastica
N: 24 è una libertà senza contenuto
N: 25 è una libertà predatoria
N: 26 è una libertà della presunzione maschile
N: 27 è una libertà patrimoniale
N: 28 è la libertà dello stupro
N: 29 è la libertà dell'inquinamento
N: 30 è la libertà di privatizzare
N: 31 Noi vogliamo una libertà responsabile
N: 32 è la libertà di abbracciare / N: 33 il genere umano
N: 34 è la libertà di darci calore e forza
N: 35 E' un'altra libertà
N: 36 Dobbiamo riprenderci / N: 37 la libertà

Figure 4. Verbal transcription of Vendola speech. IPS (inter-pause-stretches) are progressively numbered and distinct sections of the speech are displayed through different degrees of text indentation.

Bibliography

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

Garver, E. (1994). *Aristotle's Rhetoric.* Chicago: The University of Chicago Press.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech.* London and New York: Academic Press.

Henderson, A., Goldman-Eisler, F., & Skarbek, A. (1966). Sequential Temporal Patterns in Spontaneous Speech. *Language and Speech , 9* (4), 207-216.

Imberty, M. (1986). *Suoni Emozioni Significati. Per una semantica psicologica della musica.* Bologna: CLUEB.

Kennedy, G. A. (1994). *A New History of Classical Rhetoric.* Princeton: Princeton University Press.

Meyer, L. (1956). *Emotion and Meaning in Music.* Chicago and London: University of Chicago Press.

Mortara Garavelli, B. (1988). *Manuale di retorica.* Milano: Bompiani.

Patel, A. (2008). *Music, Language, and the Brain.* Oxford: Oxford University Press.

Perelman, C., & Olbrechts-Tyteca, L. (1966 ed. or. 1958). *Trattato del'argomentazione. La nuova retorica* (or.: Traité de l'argumentation. La nouvelle rhétorique, Presses Universitaires de France, Paris ed.). Torino: Einaudi.

Plebe, A., & Emanuele, P. (1988). *Manuale di retorica.* Roma-Bari: Laterza.

Trouvain, J., Koreman, J., Erriquez, A., & Braun, B. (2001). Articulation Rate Measures and Their Relation to Phone Classification in Spontaneous and Read German Speech. *Proceedings of ISCA Workshop on Adaptation Methods in Speech Recognition*, (p. 155-158). Sofia-Antipolis.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.*