IX Giornate di Studio del Gruppo di Fonetica Sperimentale dell'AIA

Aspetti computazionali in fonetica, linguistica e didattica delle lingue: modelli e algoritmi

Università Ca' Foscari – Venezia
Aula Magna Ca' Dolfin
17-19 dicembre 1998

°°°°°°°°°°°°°°°°°°

PRIMA GIORNATA

9.15 Apertura Convegno

9.30
Luciano Nebbia, Silvia Quazza, Pier Luigi Salza

Una tecnica di sintesi vocale specializzata per il dominio lessicale dell'Elenco Abbonati

10.00
Fabrizio Balducci, Loredana Cerrato, Domenico D'Alterio

Valutazione delle prestazioni di un segmentatore per l'italiano

10.30

Emanuela Magno Caldognetto, Claudio Zmarich

Implicazioni linguistiche e psicolinguistiche delle ricerche fonetiche sulle "facce parlanti"

11.00 Pausa

11.30

Anna Zanfei & Cesare Gagliardi & Luca Stefanutti

Algoritmi di assessment per un percorso di apprendimento personalizzato: il self-access per L2.

12.00
Dario Bianchi, Enrico Baricchi, Giovanni Adorni

Automated Learning of Acoustic Indexes from Data in a Text-to-Speech System for Italian

12.30 Pausa pranzo

14.30

Emanuela Cresti, Philippe Martin, Massimo Moneglia

Intonazione delle modalità naturali rappresentative: analisi e sintesi

15.00

Renata Savy

Su alcuni rapporti tra riduzione segmentale e struttura soprasegmentale nel parlato spontaneo.

15.30

Antonella Giannini e Massimo Pettorino

I cambiamenti dell'italiano radiofonico negli ultimi 50 anni

16.15 Pausa

@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@

SECONDA GIORNATA

Tavola Rotonda

Aspetti Quantitativi vs. Aspetti Qualitativi dell'Analisi Fonetico-Acustica

§§

Il riconoscimento e la sintesi del parlato sono divenute tecnologie importanti nella vita quotidiana della società contemporanea. Sono fondate sulla raccolta e analisi di enormi quantità di dati linguistici. Le basi di dati sono quindi utilizzate per costruire modelli predittivi che funzionano in modo probabilistico nelle applicazioni reali.

Attualmente sia il riconoscimento sia la sintesi sono ben lungi dal costituire dei prodotti finiti, ma anche se imperfetti vengono comunque commercializzati. Sino a che punto lo studio della variabilità linguistica interindividuale e intraindividuale può contribuire a perfezionare le tecniche di messa a punto dei modelli, considerato che le lingue naturali contengono indubbie regolarità sistematiche a vari livelli di rappresentazione astratta, in particolare di tipo sovrasegmentale e superiore al livello di parola??

§§

Moderatore: John Trumper

9.00

FUB-L.Cerrato: La modellizzazione della variazione di pronuncia riesce a migliorare le prestazioni dei riconoscitori automatici?

Scuola Normale-B.Gili: Modelli di predizione prosodica

CSELT-G.Micca: Riconoscimento multilingue con modellamento delle transizioni

IRST-M.Omologo: Gli HMMs nel riconoscimento all'IRST

10.00 Discussione

10.30 Pausa

11.00
C.Avesani: Intonazione e modelli linguistici
F.Albano-Leoni: Variabilita' nella realizzazione individuale
E.Caldognetto: La realizzazione delle emozioni
R.Delmonte: Variabilità Prosodica e Modelli Multilinguistici
S.Canazza-A.Vidolin: Il modello di resa delle intenzioni espressive nell'esecuzione musicale/vocale mediante analisi-sintesi del Centro di Sonologia Computazionale di Padova

12.15 Discussione

13.00 Pausa pranzo

14.30

PRESENTAZIONE A INVITO

Piero Cosi
OGI Toolkit.: Il riconoscimento automatico del linguaggio naturale alla portata di tutti.

15.15 Discussione

15.30 Posters

Bacalu-Delmonte, Prosodic Modeling for Syllable Structures from the VESD - Venice English Syllable Database

Bistrot-Delmonte, Il MUSEIKA in giapponese: desonorizzazione, devocalizzazione o elisione vocalica?

Delogu-Aiello-Di Carlo-Nisi-Tummeacciu, Valutazione di corpora generati a partire da scenari testuale e visivi

Zanfei-Gagliardi, Progetto per la sperimentazione di un tutor computerizzato per l'apprendimento della prosodia dell'inglese per italofoni: il modulo prosodico dello SLIM di Venezia

16.30 Pausa

17.00 Assemblea del Gruppo di Fonetica Sperimentale (GFS).

17.30 Assemblea del Centro InterUniversitario di Fonetica (CIUF).

18.00 Partenza per Burano

20.00 Cena Sociale

@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@

ULTIMA GIORNATA

9.30

Antonio Romano & Stefania Roullet

Brevi osservazioni in merito ad alcune differenze tra gli schemi intonativi adottati da uno stesso locutore per comunicare in codici linguistici diversi

10.00

Carlo Schirru

Verso un dimensionamento consonantico-temporale dell'italiano sardo-campidanese.

10.30

Claudio Zmarich

Dinamiche articolatorie nella produzione verbale fluente di normoparlanti e balbuzienti

11.00 Pausa

11.30

Francesco Cutugno

Il tempo della voce.

12.00

Reiko ENDO, Pier Marco BERTINETTO

Caratteristiche prosodiche delle così dette consonanti «rafforzate» dell'italiano

12.30 Chiusura dei lavori

R IASSUNTI

ELENCO AUTORI

La Variabilità nella Realizzazione Individuale

Federico Albano Leoni
CIRASS - Università di Napoli "Federico II"
v. Porta di Massa 1
I-80133 Napoli
tel. +39 81 5420280
fax +39 81 5420370
e-mail: fealbano@unina.it
http://www.unina.it/cirass

La variabilità è un caratteristica intrinseca di tutte le lingue storico-naturali e si manifesta, come noto, sui piani diacronico, diatopico, diastratico, diafasico e diamesico, ai quali va aggiunto quello idiosincratico individuale. Tale caratteristica è particolarmente evidente sul piano fonico.

Se si conviene su questa valutazione, ne consegue che un'analisi fonetica che sia attenta tanto ai modelli fonologici, quanto alle applicazioni pratiche, deve affiancare all'analisi qualitativa anche l'analisi quantitativa dei fenomeni, osservati all'interno di corpora di parlato che riflettano la naturale stratificazione di ciascuna lingua, dei parlanti e dei loro comportamenti comunicativi. L'analisi quantitativa consente di individuare non regole, la cui efficacia è a volte opinabile, ma tendenze associate a probabilità.

L'esigenza di disporre di corpora ampi e stratificati è particolarmente viva nell'ambito degli studi su intonazione e ritmo, che sempre di più si rivelano di una importanza cruciale per la comprensione e la descrizione della comunicazione parlata e per i quali le conoscenze accumulate sono ancora insufficienti. La variabilità del ritmo e dell'intonazione hanno infatti uno statuto diverso da quello della variabilità segmentale: infatti, mentre quest'ultima è in gran parte meccanica, la prima è invece associata a una grande varietà di intenzioni semantico-pragmatiche del parlante.

Prosodic Modeling for Syllable Structures from the VESD - Venice English Syllable Database

Ciprian Bacalu, Rodolfo Delmonte
Laboratory of Computational Linguistics
Section of Linguistics
Ca' Garzoni-Moro, San Marco 3417
Università Ca' Foscari – 30124 VENEZIA
E-mail: bacalu@unive.it
WWW:- http://byron.cgm.unive.it

The VESD has been created in order to be used in the Prosodic Module of SLIM – an acronym for Multimedia Interactive Linguistic Software, developed at the University of Venice. The Prosodic Module is composed of learning activities dealing with phonetic and prosodic problems at word segmental level and at utterance suprasegmental level. The main goal of this module is that of improving student's performance both in the perception and production of prosodic aspects of spoken language activities. The information stored in the syllable database is used both to improve the performance of the automatic segmentation of the speech signal and to give a better feedback to the student, to tell him where or what the mistake is. Preliminary investigation has also been carried out on how the syllable database can be used to build a speech recognition system that uses syllable-like units instead of phoneme-like ones as the building blocks for the recognition process.

Usually read speech databases contain hand-annotated time-aligned phoneme-level and word-level transcriptions of each utterance. Our attempt was to use the information available in order to build a syllable-level transcription of each utterance. Using only phoneme-level information was found to be difficult because the continuous syllable parsing is not as simple at utterance-level as it is at word-level. So both phoneme-level and word-level time-aligned transcriptions have been used. In order to build a database that contains syllable-level information along with word-level and phoneme-level information we used the WSJCAM0 - the Cambridge version of the continuous speech recognition corpus produced from the Wall Street Journal, distributed by the Linguistic Data Consortium (LDC). We worked on a subset of 4165 sentences, with 70,694 words which constitute half of the total number of words in the corpus amounting to 133,080. We ended up with 113,282 syllables and 287,734 phones. The final typology is made up of 44 phones, 4393 syllable types and 11,712 word types. As far as syllables are concerned, we considered only 3409

types.

In order to build syllable structures we tried to develop an automatic procedure. The algorithm for word-level syllable parsing that we used is based on the structure of English syllables and on some phonological rules. The syllable is made of a *nucleus***,** which is a vowel or a vowel-like consonant – usually a sonorant, that can be optionally prefixed and suffixed by a number of consonants, termed the *onset* and *coda* respectively. A LALR(1) grammar has been written based on this syllable structure and on the phoneme-level structure of the *onset*, *nucleus* and *coda*. We found this grammar useful for dividing the syllable into *onset*, *nucleus* and *coda*, but modifying the grammar to parse sequences of syllables resulted in an ambiguous grammar. Some phonological rules have to be applied and more look-ahead has to be done in order to resolve the conflicts during the parsing process. Based on these considerations a modified finite state automata has been built in order to parse words as sequences of syllables. The algorithm has been tested first using the Carnegie Mellon University Pronouncing Dictionary that contains more than 100.000 entries. The errors made by the algorithm were found to be caused mainly by foreign and by compound words. To limit such kind of errors we organized a list of the most frequently used foreign and compound words already divided into syllables and asked the parser to search this list every time a new segmentation is tried at word level.

Using the syllable-parsing algorithm the time-aligned syllable-level transcription of each utterance has been obtained. Then a relational database containing phrases, words, syllables and phonemes has been created. Stress information has also been included into the syllable database. This latter information has resulted from the subdivision of words into function and content words.

Bibliography

Delmonte R., M.Petrea, C.Bacalu (1997), *SLIM Prosodic Module for Learning Activities in a Foreign Language*, Proc.ESCA, Eurospeech97, Rhodes, Vol.2, pp.669-672.

Selkirk, E. (1982), «*The syllable*», *The Structure of Phonological Representations*, volume II, ed. by H. van der Hulst and N. Smith, Dordrecht: Foris, 337-383.

Kahn, D. (1976), *Syllable-based Generalizations in English Phonology*, MIT doctoral dissertation, distributed by IULC.

Hammond Michael (1995), *Syllable parsing in English and French*, University of Arizona, draft: May 25, 1995

Caratteristiche prosodiche delle così dette consonanti «rafforzate» dell'italiano

Reiko ENDO
Pier Marco BERTINETTO
La corrispondenza può essere indirizzata al secondo autore presso:
Scuola Normale Superiore
p.zza dei Cavalieri 7
56126 Pisa
email: bertiNET@sns.it

1. Nella fonetica e fonologia dell'italiano vengono convenzionalmente designate come «rafforzate» le consonanti palatali (/S L N/) nonché tutte le affricate (tS dZ ts dz/), indipendentemente dal loro punto di articolazione.* Con ciò si allude al fatto che, nella pronuncia dello standard (e delle varietà centro-meridionali in generale), le palatali intervocaliche e le affricate possiedono una durata maggiore rispetto alle normali consonanti scempie. Di questo fatto si tiene conto anche nella sillabazione, dove tali consonanti sono considerate alla stregua di geminate, fatta ovviamente salva la posizione iniziale assoluta. Questo trattamento viene mantenuto, per coerenza sistemica, anche nel caso di pronunce chiaramente non rafforzate, come quelle delle affricate palatali del toscano, sottoposte ad un nettissimo processo di fricativizzazione (cf. la pronuncia toscana di *cacio* e *adagio*; ma lo stesso vale, limitatamente alla sorda, per molte varietà centro-meridionali). Per converso, si assume comunemente che nelle varietà settentrionali non si abbia il rafforzamento delle palatali e delle affricate, e che per queste ultime, in particolare, sussista la correlazione di quantità (cf. l'opposizione in *bacio* vs. *laccio* nell'italiano settentrionale).

2. Scopo del lavoro che si intende presentare è la verifica di questo sapere tradizionale, condotta su tre piccoli gruppi di parlanti di diversa provenienza: 5 piemontesi, 6 pisani, 5 napoletani. Oltre alle consonanti rafforzate, è stato preso in considerazione, a titolo di opportuno raffronto, un campione di consonanti scempie e geminate: /n n: l l: z s s: t t: d d:/.Per mantenere il più possibile costanti le condizioni prosodiche generali, tutte le consonanti bersaglio sono state studiate all'interno di verbi, ciascuno dei quali è stato inserito in tre frasi di senso compiuto, concepite in maniera tale da mantenere inalterate le distanze interaccentuali ed il numero di sillabe complessivo (controbilanciando opportunamente la diversa estensione sillabica dei verbi utilizzati). I verbi impiegati erano i seguenti: ba**gn**are, sba**gli**are, fa**sci**are, ringra**z**iare, sgua**zz**are, organi**zz**are, sta**n**are, appa**nn**are,peda**l**are, traba**ll**are, inta**s**are, abba**ss**are, ba**ci**are, ghia**cci**are, ada**gi**are,assa**ggi**are, dila**t**are, alla**tt**are, arre**d**are, raffre**dd**are.

Ogni soggetto ha letto tre volte l'intero corpus. Le misurazioni hanno riguardato la durata delle consonanti bersaglio nonché- per controllare l'eventuale effetto accorciante di queste ultime -la durata delle vocali immediatamente precedenti. Si riportano qui in forma molto succinta i risultati, che saranno debitamente illustrati mediante apposite tabelle nella presentazione orale. Si può comunque anticipare che tali risultati hanno riservato alcune significative sorprese, che scardinano in parte le idee ricevute.

3. Innanzi tutto, i contrasti di durata tra le vocali che precedono le scempie e le geminate non sono risultati altrettanto netti in tutti i casi. Essi lo sono stati per le ostruenti (/d ~ d:/, /t ~ t:/, /s ~ s:/), ma non per le sonoranti (/n ~ n:/, /l ~ l:/). Quanto ai singoli foni rafforzati:

- *Sonoranti palatali* /N L/:

(a) Non si è riscontrata alcuna tendenza ad accorciare la vocale precedente; va peraltro osservato che, nella circostanza considerata,questo parametro non è informativo, poiché nel nostro corpus – come detto - l'opposizione di durata è risultata assente proprio nei casi che avrebbero dovuto fornire il metodo di paragone (ossia, nelle sonoranti dentali).

(b) Circa la durata di /N L/, si è constatata una sostanziale omogeneità tra i nostri locutori. Nonostante la maggior complessità articolatoria, i due foni considerati non raggiungono mai la durata dei corrispondenti foni dentali geminati /n: l:/. Ciò porta a concludere che le sonoranti palatali dell'italiano non manifestano, indipendentemente dalla varietà considerata, proprietà tali da imporle come autentici foni 'rafforzati'.

- *Fricativa palatale* /S/:

(a) Nella pronuncia dei locutori centro-meridionali, /S/ si presenta decisamente rafforzato.

(b) La situazione dei piemontesi è più incerta. La durata della consonante è intermedia rispetto a /s/ e /s:/, mentre la durata della vocale precedente colloca /S/ sullo stesso piano di /s:/ (vedi sotto per ulteriori considerazioni al riguardo).

- *Affricata palatale sorda* /tS/:

(a) Il contrasto di durata tra affricata palatale sorda ortograficamente scempia vs. geminata (cf. *baciare* vs.*ghiacciare*) è ovunque nettissimo: il fono ortograficamente scempio non presenta iconnotati autentici di un fono rafforzato, né coi locutori centro-meridionali (che lo deaffricatizzano), né coi piemontesi (che ne mantengono inalterate le proprietà articolatorie).

(b) Nulla di sistematico si può invece asserire in merito alla durata delle vocali.

- *Affricata palatale sonora* /dZ/:

(a) Coi locutori pisani e piemontesi, la situazione è identica a quella appena descritta per il corrispondente fono sordo, con l'aggiunta che, in questo caso, l'opposizione tra fono ortograficamente scempio vs. geminato è suffragata anche dalla differenza di durata vocalica.

(b) Coi napoletani, abbiamo un quadro più articolato. Rispetto alla durata consonantica, i due foni (ortograficamente scempio e geminato) sono esattamente equivalenti, mentre emerge un netto contrasto di durata vocalica (vedi sotto per ulteriori considerazioni).

- *Affricate dentali* /ts dz/:

(a) Si osserva una differenza di durata tra la sorda e la sonora ortograficamente geminate, da attribuirsi evidentemente alle diverse proprietà articolatorie;

(b) Non si osserva invece alcun contrasto nella durata delle due varianti sorde (ortograficamente scempia vs. geminata), e ciò non solo per iparlanti centro-meridionali, ma anche per i piemontesi.

(c) Benché i dati relativi alle vocali non siano utilizzabili in questo caso, per la non perfetta confrontabilità dei foni vocalici presenti in questa parte nel nostro corpus, ci sembra corretto suggerire che le affricate dentali siano omogeneamente rafforzate in tutte le varietà considerate.

4. Per concludere, ci limiteremo qui a sottolineare i due punti seguenti.

Il sussistere di contrasti di durata vocalica in assenza di analoghi contrasti tra le durate dei foni consonantici bersaglio (un evento osservato in più d'una delle circostanze sopra descritte), ci pone di fronte ad un dilemma. La differenza di durata vocalica indurrebbe ad es. ad attribuire rilevanza fonologica, coi locutori napoletani, all'opposizione tra l'affricata palatale sonora /dZ/ortograficamente scempia e la corrispondente geminata, benché la durata media di questi due foni sia praticamente identica. Ma ciò rappresenterebbe un fatto piuttosto anomalo, visto che in italiano le differenze di durata vocalica (fonologicamente condizionate) possono risultare davvero significative solo quando si sommino a concomitanti differenze di durata consonantica, mentre non possono mai esserlo di per sé sole.

Si deve inoltre, e soprattutto, rilevare che in nessuna delle varietà considerate i foni presunti rafforzati si comportano come un insieme omogeneo. Alcuni di essi manifestano le proprietà di autentiche rafforzate; gli altri no, o comunque non in maniera tale da suggerirne un'interpretazione netta. La contrapposizione tradizionale tra pronunce settentrionali e centro-meridionali appare, alla luce dei risultati ottenuti, come la grossolana semplificazione di una realtà molto più complessa e sfumata, indubbiamente meritevole di ulteriori e più sistematiche indagini.

- Per scongiurare errori dovuti a difficoltà di trasmissione via rete, si adopera qui una trascrizione semplificata ispirata al sistema SAMPA, in cui le maiuscole stanno ad indicare suoni palatali: /S/ = fricativa palatale sorda, /Z/ = fricativa palatale sonora, /L/ = laterale palatale, /N/ = nasale palatale.

Automated Learning of Acoustic Indexes from Data in a Text-to-Speech System for Italian

Dario Bianchi – Enrico Baricchi – Giovanni Adorni
Dipartimento di Ingegneria dell'Informazione Università di Parma

Contact author: Dario Bianchi
Dipartimento di Ingegneria dell'Informazione Università di Parma
Viale delle Scienze - 43100 Parma, Italy
phone: +39 521 905725, fax: +39 521 905723,
email: *bianchi@ce.unipr.it*

We are developing a text-to-speech system for Italian which uses a formant synthesizer to produce vocal output. To obtain a good quality with formant synthesizer, the main problem is the tuning of the parameters (frequency, amplitude, bandwidth of formants, source parameters etc). The number of parameters is very high and it is very important to study the transition zone between a sound and the following one (the coarticulation process).

In this paper we present a technique based on genetic algorithms to optimize the synthesizer parameters by comparing the natural speech signal (obtained by recording the utterance of a human speaker) and a synthetically produced signal. This technique simplifies and speeds up the process of tuning the synthesizer.

A well known synthesizer is that proposed by Klatt where both cascade and parallel architecture were used, the former simulates the poles of the vocal tract transfer function and is used to produce vowels while the latter is used to produce the fricative sounds. The high number of parameters used (39) to control the system can reconstruct any human speech sound with a good output quality. This synthesizer has been extensively studied and used for English language but no specific data are available for the parameters needed to generate Italian phonemes with the Klatt synthesizer.

In order to simplify the process of parameter optimization, we have used a simpler model which uses only 19 control parameters but retains a good output quality. The cascade part of the vocal tract model may be simulated by the parallel resonators configuration if suitable values are chosen for the amplitude controls. So we have removed the cascade resonators. Five resonators, connected in a parallel, are needed to simulate the poles of the vocal tract transfer function. Each resonator has associated a frequency F, a bandwidth B and an amplitude control A. A voicing source gives a periodic signal of amplitude AV at the fundamental frequency F0. A noise source is used for frication and is controlled by the amplitude parameter AF. For voiced fricatives both the periodic and the noise source are

used. To optimize the synthesizer parameters we have used a genetic algorithm which learn the parameters from trails of recorded natural speech. A chromosome represents the set of parameters we want to optimize. In a genetic algorithm we have to define a fitness function which is used to apply a selective pressure to the population. In speech processing the audio signal may be considered almost constant on a time of the order of 5-10 ms. So we have broken the audio recording in frames of 8 ms of duration.

Each recorded frame is compared with a frame of identical duration produced by the synthesizer. This comparison should be in agreement with the recognition based on the human perception. Comparing signals in the time domain does not meet this criterion. Actually, different spectral region have different weight in the phoneme discrimination process (some spectral region are characteristic the phoneme and others of the speaker). Also the fundamental frequencies of the signal may change and the phases are arbitrary. The comparison is performed better on the spectral properties i.e. in the frequency domain.

To evaluate the fitness for a single frame the following steps are considered:

a) a low pass filter is applied to the recorded signal (the target) to eliminate high frequency (up to 3 kHz for vowels and to 5 kHz for the other sounds)
b) using the parameters coded in the chromosome the signal is synthesized for the same time duration and with the same sampling rate of the target
c) a windowing is applied to the target and synthetic signal (with square or gaussian shape)
d) a 128 points Fast Fourier Transform is applied to both data
e) the fitness function is obtained comparing the Fourier spectra of the target and of the synthetic signal, using an euclidean distance between the two Fourier component vectors.

So for each frame the optimization process gives the synthesizer parameters corresponding to a single sound. The process is repeated for all the frames of the recorded wave.

For vowels it is necessary to optimize the frequency and the bandwidth of the first three formants. The 4^{th} and 5^{th} formants give the talker characteristic and don't affect the intelligibility of synthesis. They are related with some spectral details rather then perceptive features and so were not optimized. Also the amplitudes of the first 3 formants can be held constant. Fricative unvoiced source can be generated using the noise source and the high frequency resonators. The fricative amplitude AF and frequency, bandwidth and amplitude of third, fourth end fifth formants were optimized. The first and second resonators are not used. For voiced fricative there are two active source of sound, one located at the glottis (voicing) and one at the constriction of the vocal tract (friction). So the voicing and frication amplitudes (AV and AF) should be included in the number of parameters to optimize. The same parameters are used for sonorants

For nasals, a nasal pole and a nasal zero with fixed parameters were added to the system. Only the voicing source is used. Plosives are very short in time and distributes energy on all

the five formants. High frequency amplitude A4 and A5 are also used.

The easier case for the GA is that of vowels, in which only 6 parameters are used and a convergence can be usually obtained with a population of 70 individuals and 3000 generations. By the other hand voiced fricatives and plosives require the highest number of parameters (13) and a higher number of generations is required to obtain a good result (typically 15000). A convergence to a good fit was reached for all the frames in the recordings test set. The GA can be runned independently for each frame starting with a new random population every time. Nevertheless we can take advantage from the fact that in the speech signal the formants frequencies, amplitudes and bandwidth vary with continuity as a function of time. So an elithistic method was used. From one frame optimization to the next a small fraction of the final population was maintained. Typically, using a population of 70 chromosomes, the 10 best individuals obtained at the end of the optimization are transferred into the initial population used for the optimization of the next frame (while the remaining 60 individuals were initialized randomly). This method result in a shortest time required to obtain convergence and in a better overall fitness.

This method is very useful in the study of the transition zone between a phoneme a the next one. We can follow the changes of the formant parameters as a function of time and reconstruct what happens during the coarticulation process. With this method we was able to follow different transitions for example involving plosives. In order to evaluate the results obtained from the optimization we have used a perception test. Using the parameters found in the GA optimization a number of complete isolated word was generated with the synthesizer. Each word was presented 3 times (with a pause of 10 seconds) to each subject which has to write what he/she has understood. The results in audio file format are available at:

http://www.fis.unipr.it/baricchi/tesi/tesi_progress.html.

Il MUSEIKA in giapponese: desonorizzazione, devocalizzazione o elisione vocalica?

Antonella Bristot & Rodolfo Delmonte
Sezione di Linguistica
Dipartimento Studi Asia Orientale
Università Ca' Foscari – Ca' Garzoni-Moro
San Marco 3417 – 30124 VENEZIA
Tel.:0412578464/52/19 – Fax:0415287683
www: http://byron.cgm.unive.it

Il *museika*, è un fenomeno che interessa principalmente le vocali alte giapponesi [i] e [ɯ], ed è riscontrato essenzialmente nel linguaggio informale della parlata di Tokyo. Dal punto di vista fonologico, quando queste vocali si trovano tra due consonanti ostruenti [-sonoro], o a fine morfema precedute da consonante [-sonoro], esse sono sottoposte a "devoicing", rappresentate foneticamente come [i̥] e [ɯ̥]. In questo lavoro proporremo una diversificazione graduata in tre livelli del "devoicing" a seconda che vi siano oppure no tracce formantiche nello spettro, che definiremo "desonorizzazione" e "devocalizzazione", oltre ovviamente ad individuare i casi di vera elisione vocalica.

L'esperimento che abbiamo compiuto prevede uno studio comparativo che mette a confronto termini che costituiscono prestiti lessicali provenienti dalla lingua inglese, denominati "gairaigo" in giapponese e termini preesistenti nel giapponese standard, allo scopo di verificare se l'applicazione della regola fonologica di museika venga estesa automaticamente a questi nuovi elementi lessicali. Vale la pena notare che l'applicazione della regola modifica comunque la pronuncia originale delle parole gairaigo che per poter essere introdotte nel giapponese devono rispettare in tutto e per tutto la fonologia del giapponese. Le parole sono state pronunciate da parlanti di sesso maschile e femminile all'interno di una frase quadro. L'analisi spettrografica delle parole interessate dal *museika* ci ha permesso di distinguere chiaramente tra diversi tipi o gradi di "devoicing" dei vocoidi giapponesi. Dalle analisi compiute, la natura acustica e articolatoria della consonante che precede le vocali alte è fondamentale per la realizzazione del *museika*, ma vi sono altri fattori importanti che interagiscono con questo fenomeno.

Vari studi compiuti sull'argomento hanno rivelato che è possibile distinguere tra due tipi di *museika*. Quando la vocale alta è preceduta da una consonante occlusiva, sullo spettrogramma si riscontra la presenza sia di una barra vocale indicante la lieve vibrazione

della glottide sia la presenza delle formanti. È opinione generale che, in questo contesto, il *museika* può essere definito un caso di "devoicing". Al contrario, se la vocale è preceduta da una fricativa, di norma, sullo spettrogramma si visualizza un rumore di tipo fricativo che non è caratterizzato né dalla presenza della barra vocale né delle formanti. In questo secondo caso, molti parlano di elisione vocalica.

Tuttavia l'elisione vocalica implica non solo la sparizione del segmento, ma anche la creazione, a conseguenza di ciò, di nessi consonantici non permessi dalla fonotattica giapponese. Poiché la lingua giapponese ha una struttura sillabica semplice in cui la sillaba meno marcata è del tipo (C)V(C), gli unici nessi permessi sono quelli creati da una sequenza di geminate, mentre sequenze di consonanti diverse sono evitate. Ne è una riprova il largo ricorso all'epentesi vocalica nei *gairaigo*, dove i nessi consonantici del termine d'origine vengono separati tramite l'inserimento di una vocale, inserimento finalizzato all'adattamento della struttura del termine straniero al sistema sillabico giapponese.

Inoltre, chi definisce il "devoicing" nel contesto di una fricativa sorda un caso di elisione, non sembra prendere in considerazione lo sforzo compiuto dal parlante giapponese nel rilasciare lentamente la consonante che precede la vocale. L'estensione articolatoria della consonante nella zona occupata dalla vocale non solo indica che del vocoide rimane una traccia, ma permette anche di preservare la struttura sillabica del termine.

Partendo dalla considerazione che il "devoicing" dovrebbe implicare non solo l'assenza di vibrazione delle corde vocali, ma anche l'assenza delle formanti che caratterizzano questi suoni, proporremo un'analisi diversificata delle vocali sottoposte a *museika* basata sul riscontro della presenza – desonorizzazione - o assenza – devocalizzazione - di tracce formantiche nello spettro. Sosterremo, inoltre, che anche in assenza delle formanti il *museika* può e deve venire trattato come un caso di "devoicing" e non di elisione vocalica.

Analisi spettrografiche dei termini che presentano al loro interno vocali [-sonoro] hanno dimostrato che la presenza nello spettro di tracce formantiche non è condizione unica e obbligatoria per la realizzazione del *museika*. Si cercherà quindi di determinare quali altri fattori possono essere presi in considerazione e in base a quali argomentazioni si possa parlare di diversi tipi o gradi di "devoicing"

Problematiche e Risultati delle Ricerche Acustiche sull'Espressione Vocale delle Emozioni

Emanuela Magno Caldognetto

Istituto di Fonetica e Dialettologia – C.N.R.
Via G. Anghinoni, 10 - 35121 Padova (ITALY)
e-mail: Emanuela Magno@csrf.pd.cnr.it
www: http://www.csrf.pd.cnr.it

L'individuazione degli indici acustici che veicolano la trasmissione delle informazioni sullo stato affettivo del parlante e' diventata una tematica rilevante anche nel campo della fonetica perche' emozioni e attitudini sono riconosciute ormai come una delle componenti del contesto dell' interazione comunicativa faccia -a-faccia e come tali inserite nei piu' attuali modelli pragmatici della conversazione elaborati , nelle elaborazioni computazionali di dialogo proposti dall' Intelligenza Artificiale e nelle diverse applicazioni delle Tecnologie del parlato (sintesi e riconoscimento del parlato).

Tale individuazione risulta complessa a causa della cooccorrenza e interazione degli indici acustici specifici delle informazioni paralinguistiche con quelli che trasmettono le informazioni extralinguistiche e linguistiche che determinano una serie di problemi metodologici e teorici.

Tra i primi rientrano ,per esempio:

1) La tipologia del materiale da analizzare, che prevede la scelta tra emozioni reali e simulate, cioe' tra enunciati spontanei e testi predeterminati o strutture fonologiche senza significato lessicale prodotti, con l'aiuto di scenari, da attori o da parlanti non addestrati.

2) La scelta dei parametri acustici presi in considerazione per definire le variazioni imposte dalla produzione delle emozioni rispetto alle produzioni non emotive, cioe' tanto caratteristiche spettrali (formanti, concentrazioni di rumore) delle unita' segmentali, quanto parametri cosiddetti soprasegmentali: F0, durata, intensita' . Questi ultimi possono essere valutati sia globalmente, cioe' relativamente a tutto un enunciato (il cosiddetto livello macroprosodico), sia localmente (livello microprosodico), cioe' con analisi rivolte ad individuare le specifiche variazioni dei singoli parametri in relazione o alle vocali toniche o alle sillabe iniziali o finali di sintagmi o dell'enunciato.

Tra questi parametri F0 e' il piu' dettagliatamente studiato tanto quantitativamente, in termini di valori sia assoluti (massimi, minimi, medi, rang, span) che relativi (per i quali si richiedono operazioni di normalizzazione e il ricorso ad unita' di misura diverse dagli Hertz, p.es. le ottave) e di qualita' non modali della voce (presenza di laringalizzazioni, voce rauca, voce

sussurrata, atti fonatori speciali quali risate, singhiozzi, inspirazioni), quanto qualitativamente, in termini di andamenti intonativi globali e locali.
In relazione a tali andamenti fonetisti e fonologi si devono porre una serie di interessanti problemi teorici: se gli andamenti intonativi determinati dalle emozioni siano discreti o scalari, se si sovrappongano addizionalmente all'intonazione definita dalla struttura sintattica degli enunciati non emotivi o se invece si debbano ipotizzare delle intonazioni pianificate globalmente per enunciati con nuove caratteristiche pragmatiche, per le quali quindi si dovrebbe poter definire un inventario di unita' con specifiche forme e funzioni.
Nel corso dell'intervento saranno presentati esempi sperimentali atti ad illustrare le problematiche esposte e i risultati ottenuti dalle ricerche in corso presso l'IFeD che riguardano prevalentemente la caratterizzazione macroprosodica multidimensionale delle emozioni.

Implicazioni Linguistiche e Psicolinguistiche delle Ricerche Fonetiche sulle "FACCE PARLANTI"

Emanuela Magno Caldognetto, Claudio Zmarich

Istituto di Fonetica e Dialettologia – C.N.R.
Via G. Anghinoni, 10 - 35121 Padova (ITALY)
e-mail: Emanuela Magno@csrf.pd.cnr.it
www: http://www.csrf.pd.cnr.it

La prima parte delle comunicazione sarà dedicata all'esposizione riassuntiva delle ricerche condotte presso l'IFD sui movimenti labiali e mandibolari in strutture VCV (C=21 consonanti italiane; V= i, a, u). Queste ricerche hanno lo scopo di quantificare le caratteristiche del segnale ottico che trasmette l'informazione linguistica. In questa relazione non si illustreranno le possibili utilizzazioni di tali dati nelle tecnologie multimodali del parlato, ma la loro rilevanza linguistica e psicolinguistica.

La seconda parte della comunicazione discuterà alcune problematiche linguistiche:

- a livello di produzione, la maggior adeguatezza delle nuove fonologie articolatorie alla spiegazione dei rapporti tra fonetica e fonologia, cioè tra movimenti articolatori e rappresentazione linguistica [1];

13. quantificazione dell'informazione fonologica trasmessa dai movimenti articolatori visibili (**visemi**);

14. necessità degli studi contrastivi, perché l'italiano presenta caratteristiche peculiari sia per inventario fonetico che per diversa utilizzazione dello spazio articolatorio (cfr. per es. [2] per l'inglese e [3] per il giapponese).

15.dati sperimentali alla base del fenomeno del sinergismo nell'interazione tra le informazioni fonologiche trasmesse dalla modalità uditiva e quella visiva.

La terza parte della comunicazione illustrerà alcuni problemi psicolinguistici e cognitivi:

1. rappresentazione plurisensoriale delle unità fonologiche;
2. la quantificazione dell'effetto **sinergico**, cioe' dell'aumento dell'intelligibilita' nella percezione contemporanea dei due segnali visivo e uditivo, che ha luogo normalmente nell'interazione comunicativa faccia a faccia [4, 5].
3. la riflessione sui **modelli di integrazione** dei due tipi di informazione, uditiva e visiva, che prevedono differenti possibilità:

· identificazione separata delle due fonti di informazione;
· identificazione basata su ricodificazione a modalità dominante (o uditiva, o articolatoria);
· identificazione basata su ricodificazione amodale fondata sulla dinamica articolatoria.

4. la formulazione di teorie quali la **teoria motoria di percezione della parola** [6] e la **teoria diretta di percezione dell'evento** [7], che sono alternative rispetto alle tradizionali teorie acustico-uditive.

Per concludere, la quarta parte discuterà le implicazioni del concetto di visema nelle problematiche dell'accesso al lessico. Infatti, se la conversazione faccia a faccia avviene in presenza di condizioni ambientali avverse o le capacità uditive dell'ascoltatore sono ridotte, la confluenza dei 21 fonemi consonantici dell'italiano in un numero ridotto di visemi determina una riduzione della trasmissione dell'informazione fonologica, con la creazione di parole **omofene**, cioè quelle parole che sono diverse uditivamente, ma che invece sono simili in termini di lettura labiale. L'omofenia ha effetti negativi sull'accesso al lessico perché molte parole omofene e non omofone corrispondono ad una stessa stringa di visemi.

BIBLIOGRAFIA

1. Browman C.P. & Goldstein L., Dynamics and Articulatory Phonology, in R.F. Port & T. van Gelder (Eds.), *Mind as Motion*, MIT Press, Cambridge (Mass), 1995, 175-193.
2. Walden B.E., Prosek R.A., Montgomery A.A., Scherr C.K. & Jones C.J., Effects of Training on the Visual Recognition of Consonants, *Journal of Speech and Hearing Research* 20, 1977,130-145.
3. Hiki S. & Fukuda Y., Negative effect of omophones on speechreading in japanese", in C. Benoit & R. Campbell (Eds.), Proceedings of AVSP'97, Rhodes, 26-27 Sept. 1997, 9-12.
4. Erber N.P., Auditory-Visual Perception of Speech, *Journal of Speech and Hearing Disorders*, 40, 1975, 481-492.
5. Mohamadi T. & Benoit C., Apport de la vision du locuteur à l'intelligibilite' de la parole bruitée en français, *Bulletin de la Communication Parlée* 2, 1992, 31-41.
6. Liberman A.M. & Mattingly I.G., The Motor Theory of Speech Perception Revised, *Cognition*, 21, 1985, 1-36.
7. Fowler C. A., An Event Approach to the Study of Speech Perception from a Direct-Realist Perspective, *Journal of Phonetics*, 14, 1986.

Il modello di resa delle intenzioni espressive nell'esecuzione musicale/vocale mediante analisi-sintesi del Centro di Sonologia Computazionale di Padova

Sergio Canazza
Centro di Sonologia Computazionale
Dip. di Elettronica e Informatica
Universita' di Padova

Si presentera' una classe di tecniche di analisi in grado di dare una descrizione su piu' livelli della performance, sia sul piano simbolico che su quello acustico. Al fine di produrre una completa descrizione della voce cantante, differenti livelli di analisi del segnale e della performance sono stati integrati in una stessa metodologia. Queste analisi sono utilizzate al fine di giungere alla costruzione di un modello in grado di sintetizzare (mediante MIDI o tecniche di post processing) performance musicali o vocali con diverse intenzioni espressive. Saranno presentate diversi esempi di analisi e risintesi della performance musicale/vocale.

Valutazione delle prestazioni di un segmentatore per l'italiano.

Fabrizio Balducci, Loredana Cerrato, Domenico D'Alterio
Contact author:Dott.Loredana Cerrato
Speech Communication Group
Fondazione Ugo Bordoni
via B. Castiglione, 59
00142 Roma
tel. 06 54803355
fax 06 54804405
e-mail: loredana@fub.it
www.fub.it

Abstract
La segmentazione e l'etichettamento del parlato forniscono un'interfaccia tra parametri acustici fisicamente misurabili e categorie fonologiche astratte. Tra le applicazioni di un segmentatore vi è quella di etichettare segnali, anche di cattiva qualità, al fine di studiare particolari fonemi ad esempio nell'ambito applicativo del riconoscimento del parlatore.
Per valutare le prestazioni del sistema di riconoscimento RES, realizzato dal gruppo di

comunicazioni vocali della Fondazione Bordoni, si presentano i risultati di un test di valutazione eseguito sul database APASCI opportunamente modificato per applicazioni telefoniche.

Il sistema RES (Recognition Experimental System), con il quale sono stati condotti gli esperimenti, è un sistema per il riconoscimento statistico del parlato continuo, che è stato opportunamente modificato per la realizzazione di un algoritmo di segmentazione.

I risultati ottenuti sono stati valutati confrontando i limiti individuati automaticamente con quelli della segmentazione manuale di riferimento fornita con il database. Il 41.5% dei limiti di fonema individuati presentano uno scostamento inferiore ai 5ms rispetto alla segmentazione di riferimento, mentre l'88% è al di sotto dei 20ms.

OGI Toolkit.
Il riconoscimento automatico del linguaggio naturale alla portata di tutti.

Piero Cosi

Istituto di Fonetica e Dialettologia – C.N.R.
Via G. Anghinoni, 10 - 35121 Padova (ITALY)
e-mail: cosi@csrf.pd.cnr.it
www: http://www.csrf.pd.cnr.it

SOMMARIO

I sistemi di riconoscimento automatico del linguaggio naturale rendono possibile all'uomo di interagire con il computer mediante la voce, il metodo di comunicazione umana più naturale e comune. Questi sistemi sono studiati e realizzati mediante le conoscenze acquisite nel corso degli ultimi anni nel campo del riconoscimento automatico, dell'elaborazione del linguaggio naturale e delle tecnologie per l'interfaccia uomo-macchina. Essenzialmente si basano sul riconoscimento delle parole pronunciate, sull'interpretazione della loro sequenza al fine di ottenerne un opportuno significato e sull'attuazione di un'adeguata risposta. Le potenziali applicazioni sono numerosissime e, sebbene questi sistemi siano sostanzialmente agli albori, è oltremodo facile intuire la loro enorme potenzialità nel poter rivoluzionare il modo in cui le persone nel futuro si rapporteranno con le macchine. Interagendo in modo naturale, senza cioè dover sottostare ad una specifica fase di addestramento, un sempre gran numero di persone, non necessariamente specializzate, sarà introdotto all'uso di queste tecnologie. Negli ultimi anni le tecnologie relative alla realizzazione di questi sistemi di riconoscimento automatico del linguaggio naturale hanno subito una fortissima accelerazione e numerosi e notevoli sono stati i passi avanti compiuti nel campo della ricerca. Come risultato, si possono a tutt'oggi osservare un gran numero di sistemi funzionanti in compiti specifici quali la pianificazione di viaggi, l'esplorazione urbana etc.. Ormai, non si può più parlare di esclusivi prototipi di ricerca appannaggio di elitari laboratori scientifici, ma di vere e proprie

applicazioni operanti in tempo-reale, su parlato continuo, indipendentemente dal parlante che non deve più sottostare a lunghe sedute di addestramento, e supportati da vocabolari di 1000 e più parole. Questi sistemi devono essere assai più robusti degli iniziali prototipi di ricerca in quanto devono essere utilizzati in condizioni naturali quindi in presenza di rumore, sia di canale sia di ambiente, in condizioni di utilizzo che devono essere ugualmente soddisfacenti indipendentemente dal variare della velocità di eloquio, dell'accento o del sesso dell'utilizzatore. Devono esibire inoltre un comportamento 'intelligente', devono cioè essere in grado di saper reagire anche in condizioni di parziale riconoscimento, che può avvenire in seguito all'occorrenza di pronunce scorrette da parte dell'utente o a causa di altri fenomeni indesiderati. Dovranno esibire inoltre la capacità di integrarsi efficacemente con altre modalità di comunicazione, cercando di 'capire' in anticipo le intenzioni dell'utente attraverso le sue espressioni facciali, il movimento delle labbra, degli occhi etc. e sfruttando tutte le molteplici potenzialità multimediali offerte dalla tecnologia per elaborare le proprie azioni in risposta ai quesiti dell'utente rendendo l'interazione oltremodo naturale ed immediata.

Per poter sfruttare al massimo questa nuova tecnologia un sempre maggior numero di laboratori deve poter disporre di strutture informative adeguate ed è impensabile che le conoscenze necessarie allo sviluppo di una tale tecnologia siano parcellizzate e non comunemente utilizzabili. E' con questo obiettivo che all' '*Oregon Graduate Institute'* (*OGI*) di Portland ed in particolare presso il '*Center for Spoken Language Understanding*' (*CLSU*) sono stati sviluppati dei 'tools' denominati *OGI-Toolkit* [1] che hanno proprio lo scopo di fornire ad un sempre più elevato numero di ricercatori, come anche di non addetti ai lavori, degli strumenti necessari per creare e sviluppare personalmente in modo semplice ed interattivo nuovi sistemi di riconoscimento del linguaggio naturale sempre più orientati alle applicazioni. In questo lavoro vengono descritte le principali funzionalità degli OGI-Toolkit che costituiscono un insieme integrato di tecnologie software specializzate rappresentanti lo 'stato dell'arte' negli strumenti per la ricerca, lo sviluppo e l'apprendimento dei sistemi di riconoscimento del linguaggio naturale. Sono presentati inoltre i risultati ottenuti nel riconoscimento automatico di stringhe di numeri pronunciati in modo connesso in modalità indipendente dal parlante, su canale telefonico, mediante la realizzazione di tre differenti architetture completamente realizzate mediante l'applicazione degli OGI-Toolkit. In particolare, saranno descritti in dettaglio, rispettivamente, un sistema basato sulle Reti Neurali Artificiali, un sistema basato sulle Catene di Markov Nascoste ed uno su una struttura ibrida comprendente entrambe le due strutture precedentemente citate. I risultati ottenuti sono fra i migliori apparsi in letteratura su un corpus analogo.

BIBLIOGRAFIA

[1] P.Cosi, J.P. Hosom, J. Shalkwyk, S. Sutton, and R. A. Cole, 'Connected Digit Recognition Experiments with the OGI Toolkit's Neural Network and HMM-Based Recognizers', to be published in Proceedings of The 4th IEEE workshop on Interactive Voice Technology for Telecommunications Applications and ESCA Tutorial and Research Workshop on Applications of Speech Technology in Telecommunications, Torino (Italy), 29 - 30 September 1998.

Intonazione delle modalità naturali rappresentative: analisi e sintesi

Emanuela Cresti, Philippe Martin, Massimo Moneglia
(Università di Firenze e Università di Toronto)
Comunicazioni a Massimo Moneglia
e-mail moneglia@cesit1.unifi.it

E' noto che esiste un rapporto molto stretto tra l'intonazione di frase e l'espressione della modalità: dichiarativa, interrogativa, iussiva (Denes, 1960; Fonagy, 1987; Fava, 1995; Bertinetto-Magno Caldognetto, 1993).

Nonostante vari lavori (Magno-Caldognetto et alii, 1978 Sorianello, 1995; Caputo, 1996; Maturi, 1998; Schirru, 1981; Canepari, 1985; Cresti, 1994) non esiste ancora per l'italiano una descrizione complessiva delle tipologie intonative corrispondenti all'espressione della modalità di frase, ma soprattutto risulta difficile a livello teorico, isolare i caratteri pertinenti di tale indagine.

Un'importante linea di ricerca segue dall'idea che nel parlato spontaneo le entità linguistiche da considerare non siano tanto frasi caratterizzate da una diversa modalità ma enunciati che realizzano atti linguistici (Austin 1962) con diversa forza illocutoria. Essi sono sistematicamente letti da gruppi di unità tonali (pattern intonativo 't Hart et alii 1990), e si strutturano a partire dall'unità tonale di comment, che è caratterizzata dall'espressione dell'illocuzione e che è necessaria e può essere sufficiente a costituire il pattern (Cresti 1987 e seguenti).

Le variazioni di F0 che permettono il riconoscimento percettivo dell'unità di comment sono oggetto di indagine. All'interno di tale indagine assume un'importanza particolare lo studio dell'intonazione di enunciati semplici, composti da una sola unità di comment, e l'attribuzione di valore alle loro variazioni prosodiche su base percettiva.

La comunicazione presenta i risultati di una ricerca che considera: a) un comment semplice, come un predicato nominale "è Filippo", che dovrebbero essere valutata come frase di tipo dichiarativo; b) la sua realizzazione in diversi contesti "elicitanti" di risposta, deissi, identificazione, conclusione, azioni che sono tradizionalmente tutte considerate all'interno della illocuzione rappresentativa (Searle 1978); c) le sistematiche variazioni d'intonazione dell'espressioni realizzate nei diversi contesti.

Le produzioni, realizzate in laboratorio da locutori maschili e femminili, in contesti elicitanti controllati sulla base di una definizione esplicita delle caratteristiche informative di ciascun contesto, mostrano una notevole costanza formale dei profili tonali sia al variare della struttura accentuale del predicato nominale in questione (ossitona, parossitona, proparossitona) e della micromelodia di parola ("è Massimo"; "è Marilù), che al variare dei locutori.

I profili tonali dei comment prodotti in ciascun contesto elicitante saranno descritti in dettaglio secondo la teoria di Ph. Martin (1979, 1987).

L'esistenza di profili tonali con valore differenziale all'interno di una più generale tipologia rappresentativa, non è immediatamente inferibile sulla base dell'analisi dei profili, ma può essere sostenuta a livello sperimentale a seguito di una validazione percettiva condotta su due gruppi distinti di dieci parlanti competenti confrontati con enunciati naturali e di sintesi. La validazione ha mostrato che:

1) date le restrizioni informative che definiscono i contesti, i soggetti mostrano una chiara preferenza e giudicano naturali i profili elicitati in ciascun contesto, mentre sono rifiutate

massicciamente, al di là di ogni attesa, le varianti proprie degli altri contesti;
2) la sintesi dei profili tonali sottoposta agli stessi soggetti ottiene risultati paragonabili, mostrando la pertinenza melodica delle variazioni.
Ne deriva l'evidenziazione di profili tonali di valore differenziale, connessi all'espressione di modalità rappresentative diverse, e il valore naturale delle restrizioni informative che le definiscono (Moneglia-Cresti, 1997).
L'analisi e la sintesi dei diversi enunciati è stata eseguita con il sistema Win Pitch di Ph. Martin.

Il tempo della voce

Francesco Cutugno

CIRASS - Centro Interdipartimentale di Ricerca per l'Analisi e la Sintesi

dei Segnali.

tel +39 81 5420281 fax +39 81 5420370

Università degli Studi di Napoli Federico II

http://www.unina.it/cirass

e-mail cutugno@unina.it

Il ritmo delle lingue, inteso come l'insieme delle regolarità temporali presenti all'interno del codice fonetico, è oggetto di studio di fonologi, fonetisti e psicolinguisti. Le teorie di fonologia prosodica, danno una descrizione delle caratteristiche ritmiche della lingua basata sulla definizione di strutture gerarchiche ad albero binario. Le unità minime di analisi sono il piede e/o la sillaba, e l'alternanza regolare fra posizioni forti e deboli viene considerata come caratteristica universale (Principio dell'Alternanza Ritmica - PRA).

Sul versante fonetico, coerentemente con le descrizioni fonologiche, tali regolarità si realizzano nelle caratteristiche di sviluppo temporale del codice acustico. Ne conseguono disegni sperimentali che cercano di inquadrare foneticamente le lingue in tipologie metriche rigide (cfr. la distinzione fra isocronismo sillabico e isocronismo accentuale) che difficilmente giungono a conclusioni scevre da incertezza.

Il punto di vista del ricevente è quello più trascurato dalla ricerca linguistica, ma è caro a quella psicolinguistica. Ad ogni modo l'influenza delle teorie fonologiche e fonetiche è forte anche fra gli specialisti di questo ambito: buona parte della ricerca sul *processing* del segnale acustico indaga sulla forma dei presunti meccanismi che consentono al ricevente di "allinearsi" temporalmente al flusso *regolare* delle informazioni ricevute e su come il ritmo, pensato come fenomeno ciclico, possa essere impiegato per segmentare la catena fonica continua.

In questo tipo di studi è celata una visione della comunicazione audioverbale come processo *sincrono*, cioè un processo di scambio di informazioni nel quale il codice è dotato di caratteristiche temporali regolari che consentono una contemporaneità relativa fra il processo di produzione e quello di ricezione.

Scopo della presente comunicazione è indagare su questa presunta sincronicità. Ciò verrà fatto mediante un confronto delle problematiche di ambito fonetico e quelle di ambito percettivo-psicolinguistico articolato in due fasi:

1) una rassegna di studi fonetici relativi alla misura delle durate di unità segmentali e soprasegmentali in parlato letto, o comunque prodotto per scopi specifici di ricerca e in parlato connesso: mentre nel primo dei due stili possono (ma non sempre) riscontrarsi proprietà temporali regolari, nel secondo è praticamente impossibile individuare caratteristiche ritmico-temporali regolari;

2) una analisi dei principali modelli psicolinguistici sul *processing* del segnale incentrate sull'uso da parte dell'ascoltatore delle proprietà temporali del segnale vocale per la segmentazione della catena fonica continua, l'individuazione delle unità minime di analisi, l'accesso al lessico e il riconoscimento delle parole.

Seguirà una discussione sulla validità dell'ipotesi di sincronia a cui si è fatto cenno in precedenza. Verrà anche formulata una proposta alternativa delle modalità di comunicazione audio-verbale in termini di processo cognitivo di tipo *asincrono*.

In questo tipo di processi il codice trasmesso non possiede necessariamente regolarità temporali prevedibili, i partecipanti alla comunicazione eseguono i loro rispettivi compiti con modalità e scansioni temporali indipendenti e la ricostruzione delle informazioni codificate da parte del ricevente, avviene grazie all'accumulo delle informazioni in un *buffer* di memoria a breve termine. Il *processing* deve in questo caso prevedere un meccanismo di analisi capace di muoversi retroattivamente nel buffer o di aspettare l'arrivo di ulteriori informazioni in tempi successivi non prevedibili.

Prosodic Variability: from Syllables to Syntax through Phonology

Rodolfo Delmonte
Sezione di Linguistica
Dipartimento Studi Asia Orientale
Università Ca' Foscari – Ca' Garzoni-Moro
San Marco 3417 – 30124 VENEZIA
Tel.:0412578464/52/19 – Fax:0415287683
e-mail: delmont@unive.it
www: http://byron.cgm.unive.it

Variability in timing and phrasing can be nicely accounted for in multilingual contexts by taking syllables as the phonological interface from higher linguistic levels - syntactic, lexical, semantic, to lower phonetic-acoustic levels. We shall comment on the hypothesis put forward by Campbell, Isard, Breen among others, that syllables be used as the most appropriate linguistic units both in timing models for text-to-speech synthesis systems and in recognition systems: in both cases HMMs or NNs would be trained on higher level linguistic factors affecting syllable timing rather than on phone-like segments, whose duration would be inferred/induced from the interplay of syllable type and its constraining factors, with intrinsic phone mean duration values. As to phrasing, we shall comment on the need to extend the number of Boundaries to better approximate spontaneous speech, by introducing a set of syntactic prosodic boundaries intended to cover an extended number of linguistic phenomena typical of dialogue from work being done in Verbmobil on the German language reported by Batliner et al. We shall also comment on factors related to syllable-based duration model in English by the above mentioned autors. Finally we shall use this kind of prosodic labelling to comment on our duration data for Italian.

Bibliography

Batliner A., R.Kompe, A.Kiessling, M.Mast, H.Niemann, E.Noeth (1998), *M - Syntax + Prosody: A syntactic-prosodic labelling scheme for large spontaneous speech databases*, in Speech Communication, 25, 4, 193-222.

van Santen J., C.Shih, B.Moebius, E.Tzoukermann, M.Tanenblatt (1997), *Multi-Lingual Duration Modeling*, in Eurospeech'97, Rhodos, Vol.3, 2651-2654.

Breen A.P. (1995), *A Simple Method of Predicting the Duration of Syllables*, Eurospeech'95, 595-598.

van Son R.H., J. van Santen (1997), *Strong Interaction between Factors Influencing Consonant Duration*, in Eurospeech '97, Rhodos, Vol.1, 319-322.

Campbell W., S.Isard (1991), *Segment durations in a syllable frame*, in Journal of Phonetics 19, 37-47.

Campbell W. (1993), *Predicting Segmental Durations for Accomodation within a Syllable-Level Timing Framework*, Eurospeech '93, 1081-1085.

Valutazione di Corpora Generati a Partire da Scenari Testuali e Visivi

C. Delogu, D. Aiello, A. Di Carlo, M. Nisi, S. Tummeacciu
Cristina Delogu
Speech Communication Group
Multimedia Communication Department
Fondazione Ugo Bordoni
v. B. Castiglione 59 00142 Roma Italy
tel: + 39 06 54803354; fax: + 39 06 54804405
e-mail: cristina@fub.it
http://www.fub.it

I corpora per applicazioni come l'interrogazione di banche dati via voce o la traduzione automatica voce-voce, devono contenere parlato spontaneo ottenuto in modo naturale da parlatori a cui vengono presentate delle descrizioni del dominio di applicazione, ovvero degli scenari. Gli scenari devono stimolare i parlatori nella generazione di frasi con un'ampia varietà di parole e frasi. Generalmente, nell'acquisizione di corpora vocali vengono usati scenari testuali (ST). In questo lavoro viene presentato un confronto tra ST e scenari visivi (SV), in cui la situazione viene visualizzata sotto forma di vignetta. Allo scopo di valutare possibili differenze tra scenari testuali e visivi, abbiamo condotto un esperimento con 100 soggetti suddivisi in due gruppi: Gruppo ST e Gruppo SV. In particolare, volevamo testare le seguenti ipotesi: (i) le frasi prodotte dal gruppo SV sono più complesse di quelle prodotte dal gruppo ST; e (ii) le frasi prodotte dal gruppo SV sono più difficili da modellare di quelle prodotte dal gruppo ST. La verifica delle ipotesi è stata fatta attraverso l'analisi della "word intersection" e l'analisi della perplessità. I risultati di tali analisi suggeriscono che i soggetti del gruppo SV hanno utilizzato più parole per indicare uno stesso concetto; mentre i soggetti del gruppo ST hanno utilizzato sempre le stesse parole presenti nello scenario testuale. Le analisi condotte sui due corpora hanno confermato quindi la nostra ipotesi che gli scenari testuali influenzano la scelta del lessico usato per esprimere i concetti dello scenario molto più degli scenari visivi. Inoltre, le frasi prodotte a partire dagli scenari visivi mostrano una maggiore differenziazione lessicale. Questi risultati possono essere spiegati considerando il ruolo del linguaggio nella cognizione e nella psicologia sociale sottostante agli esperimenti con soggetti umani. Il linguaggio infatti aiuta a segmentare e a categorizzare la realtà. Avere a disposizione una situazione già linguisticamente descritta risparmia lo sforzo di trovare la segmentazione e la categorizzazione appropriate alla situazione ma al tempo stesso non incoraggia a trovare segmentazioni e categorizzazioni alternative. Inoltre, un esperimento è un contesto sociale specifico che induce deferenza verso lo sperimentatore da parte del soggetto sperimentale. Per questi motivi, i soggetti esposti a situazioni descritte

linguisticamente tendono a usare lo stesso linguaggio usato nel materiale sperimentale che hanno ricevuto. Mentre i soggetti esposti a situazioni visive devono trovare un loro modo per descrivere linguisticamente il materiale visivo loro proposto. Per concludere, i corpora vocali generati con scenari testuali e visivi a partire dalle stesse descrizioni concettuali sono sostanzialmente diversi e possono essere combinati vantaggiosamente in un unico corpus più ricco per addestrare sistemi di comprensione automatica di languaggio parlato.

Il sistema "RES" per il riconoscimento del parlato

Mauro Falcone
Speech Communication Group
Multimedia Communication Department
Fondazione Ugo Bordoni
v. B. Castiglione 59 00142 Roma Italy
tel: + 39 06 54803354; fax: + 39 06 54804405
e-mail: falcone@fub.it
http://www.fub.it

Si presenta il sistema RES sviluppato in Fondazione Ugo Bordoni, per il riconoscimento automatico del parlato. Il sistema fa riferimento ad un software di riconoscimento del parlato basato sulla tecnologia dei Modelli Markoviani Nascosti, sviluppato in C++ e che verrà distribuito in versione aperta, ovvero con codice sorgente in un volume di prossima pubblicazione (*C. Becchetti, L. Prina Ricotti "Speech Recognition: Theory and C++ Implementation", John Wiley & Sons*).

Verranno descritti i moduli base del software e le principali filosofie di implementazione; inoltre sarà esposto lo schema del riconoscitore ed alcuni esperimenti eseguiti su database tra i più noti quali TIMIT, ed ATIS.

Le librerie "INTEL" per in Signal Processing e lo Speech Processing

Nell'ambito della elaborazione del segnale vocale ed in particolare del riconoscimento del parlato diviene sempre più importante la possibilità di effettuare elaborazioni veloci del segnalo parlato su sistemi commerciali standard quali i personal computer su piattaforma Intel. A tal proposito verranno brevemente descritte le librerie distribuite "gratuitamente" dalla Intel per la elaborazione ed il riconoscimento del parlato.

Algoritmi di assessment per un percorso di apprendimento personalizzato: il self-access per L2.

Cesare Gagliardi & Anna Zanfei
Centro Linguistico di Ateneo
Università degli Studi di Verona
Via S.Francesco, 29
37129 - VERONA
Tel.:0458009844 Fax: 0458009372
E-mail: azanfei@chiostro.univr.it

Per determinare un percorso personalizzato di ICALL (intelligent computer assisted language learning) è necessario costruire un modello di riferimento e permettere ad un sistema tutoriale intelligente di ricostruire di volta in volta il percorso di apprendimento sulla base di procedimenti algoritmici. Il CLA di Verona è il luogo in cui si attua un progetto per l'aula multimediale in cui verrà implementato un sistema di assessment adattivo basato su un modello ricavato da una procedura di interrogazione dei collaboratori ed esperti linguistici che operano nel centro. Al modello di base verranno applicati gli algoritmi di assessment per rendere i test personalizzati ed economici in una parola: adattivi. La procedura di assessment viene in primo luogo utilizzata per testare il modello costruito con la raccolta dei dati ricavati dall'interrogazione degli esperti e l'applicazione di regole matematiche specifiche. L'assessment adattivo deve poi essere testato secondo procedure sperimentali. Il progetto è assai ampio e in questa fase presentiamo solo il funzionamento degli algoritmi utilizzabili per la "query" degli esperti e l'assessment degli studenti.

Bibliografia:

Koppen M.(1993), "Extracting Human Expertise for Contructing Knowledge Spaces: an algorithm", Journal of Mathematical Psychology 37, 1-20.

J.C.Falmagne, J.P.Doignon (1988) "A Markovian Procedure For Assessing The States Of A System", Journal of Mathematical Psychology:32, 232-258.

«I cambiamenti dell'italiano radiofonico negli ultimi 50 anni» Aspetti segmentali (prima comunicazione) e aspetti ritmico-prosodici (seconda comunicazione)

Massimo Pettorino, Adriana Giannini
Università di Napoli
E-mail:Massimo Pettorino mapettor@unina.it

Il parlato dei mass media, radiofonico e televisivo, viene spesso preso come punto di riferimento quando si va alla ricerca del cosiddetto italiano «standard». Tuttavia, come è ovvio, anche l'italiano standard cambia nel tempo in funzione di fattori di vario tipo. Ma quali sono le variazioni in atto e quali le tendenze di sviluppo per il futuro? Alcuni studi sono stati gia' condotti in tal senso, comparando l'italiano prodotto negli anni '50 con quello di oggi. Tuttavia in questi lavori i testi confrontati erano necessariamente diversi. Per superare questa difficoltà, abbiamo operato un artificio. Alcuni brani tratti da radiogiornali degli anni '50, sono stati trascritti e, in collaborazione con la sede Rai di Napoli, sono stati inseriti tra le notizie di un giornale radio attuale. Lo speaker ha così letto il testo secondo i canoni che quotidianamente segue. Questo materiale è stato poi analizzato strumentalmente. I risultati costituiscono l'oggetto delle due comunicazioni qui proposte, una riguarda il piano segmentale, l'altra quello ritmico-prosodico.

Nel primo lavoro vengono rilevate le formanti delle vocali e, per ciascun parlante, viene comparato il sistema tonico con quello atono. E' importante notare che, grazie all'artificio descritto sopra, e' possibile confrontare elementi che occorrono in situazioni contestuali del tutto simili, in quanto il testo, come si e' detto, e' lo stesso e prodotto in una medesima situazione ambientale, ma pronunciato a distanza di circa mezzo secolo. I risultati di tali analisi mostrano le variazioni occorse in questo intervallo di tempo e indicano la linea di tendenza per il futuro. Il secondo lavoro proposto si sofferma sugli aspetti ritmico-prosodici. Dei due enunciati messi a confronto vengono calcolati gli indici di fluenza, velocita' di eloquio e di articolazione, la distribuzione e durata delle pause e infine l'andamento intonativo. Un esame che, se pur condotto su un campione ridotto, rivela aspetti interessanti dei cambiamenti avvenuti nella nostra lingua negli ultimi 50 anni.

Modelli di predizione prosodica

Barbara Gili
Scuola Normale Superiore
p.zza dei Cavalieri 7
56126 Pisa
E-mail: gili@alphalinguistica.sns.it
http://alphalinguistica.sns.it

La struttura prosodica arricchisce gli enunciati di informazioni di tipo pragmatico, esplicitando, ad esempio, l'intenzione comunicativa del parlante, ma fornisce anche indicazioni utili sulla struttura delle frasi a cui viene associata. Poiché si è essenzialmente liberi di scegliere quali variazioni prosodiche realizzare, esiste un alto grado di variabilità nelle relazioni tra struttura prosodica e struttura di frase. Tuttavia, è possibile individuare alcune 'tipiche' corrispondenze tra prosodia e testo, e pensare, quindi, che le informazioni relative ad una struttura possano gettare luce sull'altra, e viceversa. Tale corrispondenza consente alle tecnologie vocali sia di sfruttare le variazioni prosodiche per ricavare indicazioni sulla struttura delle frasi (procedimento utile nei sistemi di riconoscimento vocale), sia di ricavare alcune informazioni prosodiche a partire dall'analisi strutturale degli enunciati (processo caratteristico dei sistemi di sintesi vocale).

Nel corso dell'intervento, considereremo prevalentemente il secondo aspetto, fornendo una panoramica delle tecniche utilizzate per generare automaticamente la prosodia a partire dal testo. Ci concentreremo, quindi, sui sistemi di sintesi vocale piuttosto che su quelli di riconoscimento, in cui l'elaborazione prosodica ha avuto fino ad oggi un ruolo secondario. In particolare, prenderemo in considerazione esempi concreti di sistemi in cui la struttura prosodica viene ricostruita per mezzo di regole, altri in cui si utilizzano algoritmi di apprendimento automatico, e, infine, sistemi in cui il problema viene, in un certo senso, aggirato concatenando unità che siano già caratterizzate prosodicamente.

Riconoscimento vocale per vocabolari multilingua

Giorgio Micca
CSELT
Via G. Reiss Romoli 274
10148 Torino, Italia
Tel: +39 011 228 6241
Fax: +39 011 228 6207
E-mail: giorgio.micca@cselt.it

I riconoscitori vocali per vocabolari flessibili richiedono un modellamento con unita` piu` piccole della parola, a differenza dei riconoscitori specializzati per particolari applicazioni dove l'insieme delle parole chiave e` costituito da poche decine di elementi. In quest'ultimo caso si adotta un modello a parole intere, che ha il pregio di ottimizzare le prestazioni in termini di accuratezza di riconoscimento. Normalmente si definiscono insiemi di alcune centinaia di unita` acustico-fonetiche che fanno riferimento alla struttura fonetica di una data lingua. Tali unita` possono modellare anche gli allofoni, e in genere differenziano un modello fonetico in funzione del contesto in cui tale modello si trova. Si ottengono cosi` i trifoni, dove il fonema e` considerato nel contesto dei fonemi sinistro e destro adiacenti, i polifoni, dove l'orizzonte di dipendenza si estende anche a distanze di ordine due o superiore, i modelli sillabici, dove la coarticolazione e` modellata implicitamente all'interno della sillaba, ed infine i modelli di tipo transizione-stazionarieta`, dove si rappresentano in modo differenziato ed esplicito le componenti stazionarie dei fonemi e le traiettorie di transizione da un fonema a quello immediatamente successivo. Tanto piu` e` dettagliato il modello, tanto maggiore e` la sua precisione di rappresentazione all'interno dello spazio acustico della lingua, ma tanto piu` e` difficile addestrare in modo robusto il modello stesso, perche` la cardinalita` dell'insieme delle unita` aumenta. Ad esempio, per poter garantire una copertura molto elevata di tutti I fenomeni fonotattici della lingua italiana, sarebbe necessario definire un insieme di trifoni di circa sette-ottomila elementi; se si considera che per avere un modello sufficientemente robusto dal punto di vista statistico e` necessario disporre di almeno un centinaio di occorrenze di ogni elemento all'interno della base dati di addestramento, e se si tiene conto che, anche in una base dati lessicalmente progettata in modo tale da risultare foneticamente bilanciata, la distribuzione di frequenza delle unita` presentera` egualmente delle forti asimmetrie dovute alla struttura stessa della lingua, si capisce come un modello trifonico ad alta copertura fonotattica richiederebbe una base dati di dimensioni difficilmente

gestibili e troppo costosa. Un modello sillabico avrebbe difficolta` analoghe, ed un modello polifonico ancora maggiori. Normalmente, ci si limita a rappresentare I fenomeni a piu` alta incidenza statistica, cosicche` una data parola verra` rappresentata in definitiva da un insieme misto di unita` di vario grado di contestualita` (ad esempio, trifoni, bifoni destri e sinistri e semplici fonemi). In questo modo, si raggiunge un compromesso tra *precisione* del modello e sua *addestrabilita`*. I modelli di tipo transizione-stazionarieta` hanno il vantaggio di generare una minore complessita` dell'insieme di unita` - alcune centinaia – e sono stati recentemente utilizzati nel riconoscimento vocale con buoni risultati. Una ulteriore dimensione viene introdotta quando si considera la multilingualita`, la` dove si intendere sviluppare un modello di riconoscimento tendenzialmente indipendente dalla lingua, o per lo meno utilizzabile con un vocabolario applicativo costituito da parole appartenenti ad un insieme definito di lingue diverse. In questo caso la cardinalita` dell'insieme delle unita` da rappresentare aumenta linearmente con il numero delle lingue interessate.

Recentemente ho affrontato questo tema di ricerca, cercando di dare una risposta a due quesiti:

1) come definire un modello acustico-fonetico multilingua per riconoscitori vocali a vocabolario flessibile adatti ad applicazioni che richiedono la presenza di parole appartenenti a lingue diverse;
2) come sfruttare un modello multilingua di questo tipo per "interpolare" un riconoscitore in una nuova lingua, diversa da quelle del modello originario, tenendo conto delle similarita` fonetiche dei suoni della nuova lingua rispetto ai suoni di una delle lingue del modello multilingua. Quest'ultimo punto ha una variante applicativa interessante, che e` quella di poter sviluppare un riconoscitore statisticamente robusto anche per una lingua per la quale la quantita` di materiale vocale disponibile per l'addestramento e` scarsa, anche qui sfruttando la similarita` dei suoni della lingua considerata con quelli delle N-1 lingue appartenenti al modello.

L'approccio seguito, ancora in corso di sviluppo, ha due componenti:

a) introduzione di metriche "a posteriori" per la misura del grado di similarita` dei modelli acustico-fonetici dei suoni di due o piu` lingue, ottenedo una struttura gerarchica che rappresenta gli aggregamenti in classi in funzione delle metriche adottate;
b) Introduzione dell'unita` acustico-fonetica di base, che` quella di tipo classe di transizione-stazionarieta`, dove pero` vengono effettuati due tipi di accorpamenti: 1) tra elementi di

stazionarieta` dove la metrica introdotta al punto a) individua simililarita` superiori ad una soglia prefissata, e 2) tra elementi di classi di transizioni, la` dove la stessa metrica individua elementi cosi` vicini nello spazio acustico da dover essere unificati.

Se la granularita` delle classi fonetiche e` elevata, come ad esempio avviene in un modello dove si distinguano essenzialmente suoni plosivi, fricativi e liquidi-nasali, piu` le classiche tre classi vocaliche (frontali, centrali e posteriori), il modello risulta sufficientemente generale da non dover richiedere – probabilmente – nessuna unificazione di modelli di transizioni tra le suddette classi, almeno per le principali lingue europee considerate. In questo modo, e` possibile ottenere un modello acustico-fonetico allo stesso tempo semplice – poche centinaia di unita` - e preciso, o almeno ottimale rispetto alla cardinalita` ammissibile dell'insieme. Il modello risultera` tanto piu` generale quanto piu` lingue saranno state considerate in fase di progetto, e anche tanto piu` capace di adattarsi ad una lingua nuova o "povera", nel senso dato a tale termine in precedenza. Vengono presentati risultati ottenuti nel caso bilingue Italiano e Spagnolo, e si presenta il progetto relativo alla estensione del modello all'Inglese ed al Tedesco, con una possibile sperimentazione alla "interpolazione" di un riconoscitore per il Rumeno.

Bibliografia

1) P. Bonaventura, F. Gallocchio, G. Micca, "Multilingual Speech Recognition for Flexible Vocabularies", EuroSpeech '97, Rodi, Gracia, 22-25 settembre 1997, pp. 355-358.
2) P. Bonaventura, F. Gallocchio, J. Mari, G. Micca, "Speech Recognition Methods for Non-Native Pronunciation Variations", Workshop on "Modeling Pronunciation Variation for Automatic Speech Recognition", Rolduc, Olanda, 4-6 maggio 1998, pp. 17-22

Gli HMMs nel riconoscimento all'IRST

Maurizio Omologo
ITC-IRST
Povo - Trento
E-mail: omologo@irst.itc.it
http://poseidon.itc.it:6116/~omologo

Negli ultimi anni i piu' efficaci sistemi per il riconoscimento automatico della voce (ASR) sono basati sui Modelli di Markov nascosti (HMM) e il loro impiego per la modellizzazione acustica del parlato continuo prevale in questo campo. Benche' il paradigma HMM rappresentera' per lungo tempo la tecnologia dominante, presenta comunque degli aspetti

critici: la debole modellizzazione della durata, l'assunzione di indipendenza delle osservazioni data una sequenza di stati, le limitazioni di parametri acustici basati sull'analisi a finestre.
L'efficacia nel trattare l'intrinseca variabilita' del parlato e le buone prestazioni sono spiegabili con la capacita' di gestire sorgenti non-stazionarie ma la teoria non prevede un esplicito meccanismo per modellare le variazioni dei segnali e le corrispondenti relazioni temporali dato un contenuto fonetico fissato. Viene assunta la stazionarita' condizionata dallo stato per i dati osservati e solo la catena di Markov nascosta si "adatta" alla non-stazionarieta' della produzione vocale. Questa ipotesi di stazionarieta' degli stati appare ragionevole se uno stato rappresenta un segmento breve di alcuni suoni (es. fricative) mentre per segmenti piu' lunghi tale assunzione si mostra inadeguata; le regioni di transizione tra fonemi rivelano la dominante natura non stazionaria della voce. Il rilassamento dell'ipotesi di indipendenza condizionata delle osservazioni e' oggetto di numerosi studi. Un semplice meccanismo in uso per considerare questa dipendenzatemporale e' l'aumento dello spazio delle osservazioni acustiche con le derivate dei parametri. E' anche possibile impiegare parametri segmentali piuttosto che ricavati dall'analisi a finestre ma il modello statistico corrispondente risulta notevolmente piu' complesso.

Brevi osservazioni in merito ad alcune differenze tra gli schemi intonativi adottati da uno stesso locutore per comunicare in codici linguistici diversi

Antonio Romano[1] *& Stefania Roullet*[1 & 2]

[1] Centre de Dialectologie de Grenoble
Université Stendhal BP 25
Domaine Universitaire
38040 Grenoble (France)

[2] B.R.E.L. - Bureau Régional pour l'Ethnologie et la Linguistique
via Grand Eyvia, 59
11100 Aosta (Italia)

Autore da contattare:
Antonio Romano, e-mail: romano@u-grenoble3.fr
Tel.: +33 4 76 82 68 64 Fax: +33 4 76 82 43 56

Con questo intervento vorremmo stimolare la discussione in merito ad alcuni interessanti fenomeni di "interferenza" tra gli elementi prosodici di uno o più codici linguistici, a cui uno

stesso locutore può fare riferimento, pur disponendone con diversi "gradi di sicurezza".

Da più parti si sostiene che le caratteristiche prosodiche, in particolar modo intonative, siano le ultime ad essere abbandonate e le prime ad essere acquisite. Secondo D. Hirst & A. Di Cristo (1998, p. 2), ad esempio, « The prosodic characteristics of a language are not only probably the first phonetic features acquired by a child [...] but also the last to be lost either through aphasia [...] or during the acquistion of another language or dialect ». In ogni caso, sembra che i mutamenti che coinvolgono le strutture ritmiche e intonative siano meno rapidi di quelli che riguardano il lessico e la morfosintassi (cfr. Bertinetto & Magno-Caldognetto, 1993, p. 141).

Rivolgendoci in particolar modo alle varietà linguistiche che costituiscono il principale oggetto della nostra attenzione (e cioè l'italiano e le altre varietà romanze che con esso convivono) siamo immediatamente messi a confronto con una realtà in cui tale processo di "acquisizione/oblìo" mostra un'intensa dinamica, spesso caratterizzata da interazioni tra sistemi diversi piuttosto che da pressioni unilaterali.

Già G.B. Pellegrini (1960, p. 16) sosteneva che « la pronuncia dell'italiano regionale svela quasi sempre il sottofondo dialettale che fa capolino con maggiore o minore evidenza secondo l'attenzione e la cultura del parlante ». Inoltre, come fa notare T. Telmon (1990, p. 14), « Ciò che è assai più interessante è che anche oggi, in una situazione di drastica riduzione numerica dei dialettofoni, i fatti intonativi, prosodici e fonologici continuano ad essere le spie acutissime della regionalità di qualsiasi parlante italiano ».

Concentrando la nostra attenzione sui fenomeni prosodici, ci accorgiamo che, come descritto, ad esempio, da M. Voghera (1992, p. 88), « [...] è proprio a livello intonativo che si sedimentano differenze tra parlanti di diversa provenienza geografica e/o culturale ». Dello stesso avviso è anche L. Canepari (1979, p. 276), che sostiene: « spesso coloro che hanno eliminato le caratteristiche articolatorie (più) marcatamente regionali della loro pronuncia conservano le strutture intonative della loro parlata originaria: ché sono le più difficili da modificare ».

Malgrado queste interessanti considerazioni di ordine generale - che un qualsiasi attento ascoltatore può confermare sulla base della sua quotidiana esperienza -, ci sembra che nessuna analisi descrittiva di approccio diretto, sia mai stata rivolta all'analisi dei cambiamenti osservabili nel sistema intonativo di una determinata area linguistica italiana, nel passaggio da un registro all'altro (per es. formale vs. informale), o allo studio di una determinata situazione di dinamiche linguistiche in cui siano coinvolti "due" codici distinti[1].

[1] Una prima descrizione quantitativa di questo tipo di fenomeni è stata tentata, in maniera alquanto semplicistica e per un ristretto numero di locutori di varietà salentine, da A. Romano (1997, p. 175), il quale, descrivendo alcune caratteristiche prosodiche dialettali riscontrate anche nell'italiano regionale parlato dagli stessi locutori, precisa: « Quantitative data obtained from our speech signal analysers showed close correlation between prosodic parameters of utterances in dialect and those in standard Italian. Obviously this conclusion cannot be generalised because various

Al tentativo di dare una spiegazione di tale fenomeno potrebbe essere d'aiuto il quadro fornito da A.A. Sobrero & I. Tempesta (1996, p. 110), secondo i quali : « La pressione standardizzante della scuola ha sempre agito sulla fonetica e sul lessico : per più di un secolo la didattica linguistica dell'Italia unita ha fornito, regione per regione, elenchi di suoni e di parole che non dovevano essere usati perché troppo vicini al dialetto. [...] Sul ritmo e sull'intonazione i libri non dicevano nulla [...]. Questi livelli di lingua si sono così tramandati senza particolari censure e si sono preservati al punto che oggi l'usare una determinata cadenza è il segnale più forte e sicuro - spesso l'unico - dell'appartenenza a una determinata area o areola linguistica ».

Resta però ancora da stabilire come e quando i caratteri di un dato sistema linguistico di cui il locutore dispone *ab ovo*[2] si affermano nell'uso di un codice linguistico diverso ("quanto?") o quantomeno avvertito come "altro" dallo stesso locutore, indipendentemente dal fatto che egli sia in grado di esercitare un controllo sulle mutue interferenze, e a prescindere dall'aver appreso i due codici linguistici in egual misura e contemporaneamente o in tempi e modi diversi.

Come risulta evidente, la questione è spinosa, ma è soprattutto resa inavvicinabile dall'estrema variabilità delle esperienze individuali del locutore. Nella presente occasione, volendo quindi evitare di far rientrare nelle nostre valutazioni le delicate problematiche dell'acquisizione del linguaggio e dell'apprendimento di una seconda lingua, ci limitiamo a riportare i nostri dati discutendo le nozioni di resistività e di labilità delle caratteristiche prosodiche nei casi delle lingue a nostra disposizione in questo primo momento.

Alla luce dei dati di cui disponiamo, ci proponiamo di verificare specialmente il grado di persistenza (o di variazione) degli schemi intonativi in relazione al mutamento di codice linguistico.

Per quel che riguarda l'analisi svolta in Valle d'Aosta, si è scelto di prendere in considerazione due paesi che, pur appartenendo alla stessa area linguistica, si differenziassero il più possibile quanto al grado di "esposizione all'italiano".

Da un primo sommario esame dei dati raccolti, sembrano esistere differenze significative tra gli schemi intonativi dei due codici linguistici e nei due punti considerati.

Grazie a questi primi dati, integrati dall'esperienza linguistica personale di uno degli autori, sembra quindi possibile poter affermare che schemi intonativi caratteristici delle varietà francoprovenzali e dell'italiano regionale si estendano su aree linguistiche non coincidenti e che, in generale, quelli dell'italiano regionale siano comuni ad aree più estese al cui interno

---

reactions are possible following different contextual condition and various degrees of articulation spanning from hypo- to hyper-speech. [...] In the natural and spontaneous conditions we inspected, intonational patterns as well as duration values of the Italian messages match those of the dialectal ones ». (Notare che in riferimento alla varietà di italiano sarebbe meglio ricorrere qui all'aggettivo "regionale" che in quella sede avrebbe richiesto ulteriori precisazioni).

[2] Ontogeneticamente, ma sarebbe altrettanto interessante approfondire gli aspetti filogenetici del fenomeno.

sono invece riscontrabili, nelle diverse varietà francoprovenzali, soluzioni prosodiche (ma anche fonetiche, morfosintattiche ecc.) differenti.

Per il salentino, invece, una prima descrizione piuttosto approssimativa induce a distinguere due sottosistemi intonativi diffusi rispettivamente in un'area meridionale estrema e in una centro-settentrionale (cfr. anche Romano, 1997); il locutore originario di una di queste due aree utilizza il sistema intonativo della sua area quando si esprime nel suo dialetto e nella maggior parte delle situazioni in cui è chiamato ad esprimersi in italiano in ambiti familiari o subregionali. L'"italiano" di riferimento al di fuori della sua regione può improvvisamente divenire, in maniera del tutto accidentale, un "italiano" presunto dell'interlocutore o un "italiano" di riferimento di provenienza imprevedibile che ricorda quello da lui usato nella lettura (cfr. Savino & Refice, 1997) o quello appreso in determinati contesti extraregionali. Al di là dell'esistenza di fasce geografiche di transizione[3], sembrerebbe anche possibile che delle preferenze individuali facciano emergere uno dei due sistemi in aree in cui il sistema dominante è messo in ombra da complicate situazioni di contatto o dal progressivo diffondersi di varietà di prestigio.

Un sistema potrebbe essere più arcaico e l'altro "innovativo". Come spiegare però la graduale rinuncia da parte di un'intera comunità a un sistema intonativo che si pretende profondamente radicato negli usi linguistici di ogni singolo individuo?

Bibliografia

Bertinetto P.M. & Magno Caldognetto E. (1993). "Ritmo e intonazione". In A.A. Sobrero (a cura di), *Introduzione all'italiano contemporaneo. Le strutture.* Bari, Laterza, 141-192.

Canepari L. (1979). *Introduzione alla fonetica*. Torino, Einaudi.

Hirst D. & Di Cristo A. (in corso di stampa). "A survey of intonation systems". In D.J. Hirst & A. Di Cristo (a cura di), *Intonation Systems: a Survey of Twenty Languages*, Cambridge Univ. Press.

Pellegrini G.B. (1960). Tra lingua e dialetto in Italia. *Studi mediolatini e volgari*, VIII, 137-153 (v. anche Pellegrini G.B., 1975, *Saggi di linguistica italiana*, Torino, Boringhieri).

Romano A. (1997). Persistence of prosodic features between dialectal and standard Italian utterances in six sub-varieties of a region of southern Italy (Salento): first assessment of the results of a recognition test and an instrumental analysis, *Proc. of Eurospeech '97*, 175-178.

Savino M. & Refice M. (1997). "L'intonazione dell'italiano di Bari nel parlato letto e in quello spontaneo". In F. Cutugno (a cura di), *Fonetica e fonologia degli stili dell'italiano parlato.* Atti delle VII Giornate di Studio del G.F.S., Esagrafica, Roma, 1997, 79-88.

Sobrero A.A. & Tempesta I. (1996). La Puglia una e bina. *Italiano e Oltre*, XI, 2, 107-114.

Telmon T. (1990). *Guida allo studio degli italiani regionali*. Dell'Orso, Alessandria.

Voghera M. (1992). *Sintassi e intonazione nell'italiano parlato*. Il Mulino, Bologna.

[3] La suddivisione nelle due aree, nettissima sul versante occidentale della penisola, con una frontiera segnata tra comuni distanti solo 4 km, non lo è altrettanto sul lato orientale, dove i due sottosistemi, pur caratterizzati da sfumature che ne riducono le opposizioni, sono attestati in ugual misura durante le prime inchieste e potrebbero appartenere a registri diversi dello stesso locutore.

Una tecnica di sintesi vocale specializzata per il dominio lessicale dell'Elenco Abbonati

Luciano Nebbia, Silvia Quazza, Pier Luigi Salza
CSELT, Centro Studi E Laboratori Telecomunicazioni
Via G. Reiss Romoli, 274 - 10148 Torino
Autore di riferimento: Pier Luigi Salza (pierluigi.salza@cselt.it)

Riassunto

Si descrive la realizzazione di una versione specializzata di Eloquens®, il sintetizzatore vocale da testo scritto sviluppato in Cselt, progettata con l'obiettivo di conseguire un sostanziale miglioramento dell'intelligibilità e della naturalezza della risposta vocale nel servizio "1412", il servizio informazioni elenco abbonati Richiesta Numerica, che fornisce automaticamente nominativo e indirizzo dell'abbonato a partire dal numero telefonico. Lo sviluppo di una voce più naturale dovrebbe rendere il servizio automatico più accettabile all'utenza, che, abituata ai servizi erogati con voce preregistrata, ha accresciuto le proprie aspettative nei confronti della qualità acustica delle risposte vocali. L'obiettivo è stato conseguito sfruttando le caratteristiche peculiari del dominio applicativo: il sistema è stato specializzato sul dominio lessicale, per quanto riguarda sia la pronuncia delle parole sia il modo di comporle. Vediamo brevemente come.

Un sintetizzatore vocale da testo per uso generale, quale è Eloquens® standard, deve essere predisposto a trattare qualunque testo di una data lingua. Così la maggior parte della capacità del sistema di elaborare il testo è dedicata ad individuare la struttura sintattica della frase e ad assegnarle la corretta prosodia, mentre le unità acustiche sono progettate per coprire tutti i possibili contesti fonetici. Le unità che da più tempo vengono usate in sistemi di questo tipo sono *difoni* con struttura fonetica uniforme e non contestuali. Il vantaggio principale dei difoni di questo tipo è dato dalla copertura offerta su testi qualsiasi, perchè con un unico rappresentante per ciascuna delle combinazioni di suoni possibili in una data lingua (in genere sull'ordine del migliaio) è possibile generare qualsiasi messaggio in quella lingua. Inoltre, la voce prodotta con questo tipo di tecnica risulta generalmente di alta intelliggibilità. Per contro, l'uso dei difoni comporta due notevoli svantaggi:

- la voce prodotta, pur risultando altamente intelliggibile, presenta gravi carenze dal punto di vista acustico, in particolare per quanto riguarda la naturalezza del timbro, solitamente caratterizzato da uno sgradevole suono "artificiale" o "robotico", le cui

possibili cause sono: l'eccessiva uniformità delle unità acustiche registrate originariamente; il procedimento utilizzato per la modifica prosodica, il quale provoca delle rilevanti distorsioni nei casi in cui la "distanza" fra i valori prosodici "target" e quelli intrinseci del segnale presente nella base dati originale è rilevante;

- cambiando lingua occorre riprogettare *ex novo* la base dati dei difoni, in conformità alle regole fonotattiche della nuova lingua.

Tuttavia, importanti applicazioni della sintesi da testo non traggono beneficio da una impostazione così generale. Ad esempio, nel servizio Richiesta Numerica la struttura del messaggio è fissa, essendo sostanzialmente composta da una lista di pochi elementi (il nome, il cognome e l'indirizzo), e i contorni prosodici sono semplici e ripetitivi. Data la finalità del servizio e la struttura semplice del messaggio, è stato valutato che la lettura parola per parola sia una modalità accettabile (se non addirittura preferibile) di fornire l'informazione. Inoltre, il dominio lessicale è limitato alle sole parole contenute nell'elenco abbonati. E' stata anzi individuata una lista limitata di parole che da sole coprono un'alta percentuale prefissata delle parole da sintetizzare. Queste caratteristiche dell'applicazione hanno suggerito un approccio alla tecnica di sintesi nel quale il concetto di difono viene profondamente rivisto. Tale approccio punta alla riduzione delle discontinuità ai confini di unità e delle distorsioni in fase di concatenazione, utilizzando unità acustiche *non uniformi*, cioè di lunghezza fonetica variabile e comunque più ampia del difono, e *contestuali*, cioè presenti con più occorrenze prosodicamente differenti.

Se dovessero essere utilizzate per sintetizzare testi qualsiasi e a struttura sintattica variabile, il numero di tali unità acustiche potrebbe risultare enormemente elevato; tuttavia il loro numero si ridimensiona notevolmente se la copertura si deve invece limitare alle parole di un determinato dominio. Inoltre, la lettura a parole isolate semplifica ulteriormente i problemi: da un lato, infatti, riduce fortemente il numero delle unità necessarie, essendo escluse quelle ai confini di parola; d'altro lato, il modello prosodico per una data unità acustica risulta altamente predicibile sulla base della sua posizione all'interno della parola, ciò che consente di contestualizzare efficacemente le unità mediante una semplice classificazione posizionale e accentuale, in grado già di per sè di riflettere la struttura sillabica e prosodica della parola isolata. E' diventato così possibile, dalla registrazione di parole lette isolatamente con prosodia il più possibile uniforme, scegliere le unità contestuali che possono essere semplicemente concatenate, senza ulteriori aggiustamenti prosodici.

La scelta delle unità avviene escludendo innnanzi tutto di segmentare le vocali. In secondo

luogo, si sono definiti due livelli gerarchici di segmentazione, basati su criteri acustici e percettivi, detti primo livello (o livello ottimo) e secondo livello (o livello sub-ottimo), che si differenziano in base alla "resistenza" alle discontinuità coarticolatorie offerta dalle consonanti segmentate. Si è stabilita quindi una gerarchia delle unità da coprire ai due livelli in base alla loro occorrenza in parole più o meno frequenti. Il dato della frequenza è stato ricavato dall'analisi statistica di un ampio corpus di cognomi, nomi, indirizzi e parole comuni, inclusi anche parole e cognomi stranieri. Per unità al di fuori di questa copertura (che in ogni caso ammontano a meno del 4% del totale), si ricorre ai difoni, che rappresentano la soluzione di recupero del sistema. La realizzazione di questo approccio si è avvalsa della struttura flessibile del sintetizzatore Eloquens®, che ha reso possibile integrare dunque 3 tipi di unità acustiche diverse (primo livello, secondo livello e difoni) per un totale di circa 21.200 segmenti acustici fisicamente conservati nel dizionario (primo livello più difoni), dai quali si ricavano ancora ulteriori 20.000 sottosegmenti di secondo livello, contenuti nei primi. Sottoposta a valutazione soggettiva, la versione specializzata di Eloquens® mostra, rispetto a quella standard, un incremento della comprensione soggettiva del 7.6 % nel campo Nome, un incremento di 0.7 punti su 5 della media MOS, un miglioramento dell'intelligibilità reale del 4.2% riferito alla parola singola nel campo Nome.

SU ALCUNI RAPPORTI TRA RIDUZIONE SEGMENTALE E STRUTTURA SOPRASEGMENTALE NEL PARLATO SPONTANEO.

Renata Savy
CIRASS - Centro Interdipartimentale di Ricerca per l'Analisi e la Sintesi dei Segnali.
tel +39 81 5420281 fax +39 81 5420370
Università degli Studi di Napoli Federico II
via Porta di Massa 1 80133 Napoli
E-mail: rensavy@unina.it

Il lavoro che si intende presentare è parte di una più ampia ricerca su fenomeni di riduzione e/o cancellazione di materiale fonico segmentale nell'italiano parlato spontaneo.

L'indagine si basa su un *corpus* di parlato conversazionale selezionato tra le registrazioni del LIP (cfr. De Mauro et al. 1993) e orientato su due diverse varietà diatopiche: si tratta di

quattro conversazioni spontanee bidirezionali (faccia a faccia), due registrate a Napoli e due a Milano.

Su questo campione sono state svolte precedentemente alcune indagini incentrate sulle realizzazioni foniche dei suffissi morfologici flessivi dei costituenti nominali e verbali. L'analisi spettroacustica del materiale ha evidenziato che in misura significativa (oltre il 50%) i suffissi flessivi sono soggetti a riduzione; la riduzione si manifesta come a) riduzione timbrica delle vocali finali di parola, b) sostituzione timbrica di vocali finali di parola, c) cancellazione di vocali finali di parola, d) cancellazione di sillabe finali di parola.

Ulteriori analisi sono state effettuate sul corpus allo scopo di indagare i rapporti che intercorrono tra riduzioni sul versante segmentale e organizzazione prosodica delle stringhe. Il lavoro comprende:

1) la segmentazione dei brani di conversazione in unità tonali (**TU**) e costituenti prosodici minori (sintagmi intermedi [**SI**]);

2) la classificazione dei contorni melodici e l'individuazione dei *pitch accents* all'interno delle unità prosodiche;

3) un'analisi distribuzionale delle riduzioni di ordine morfologico-segmentale all'interno delle unità intonazionali individuate.

Il risultato delle analisi mostra che esiste una correlazione tra la frequenza di riduzione e la posizione di un determinato elemento (o gruppo di elementi) dentro la stringa prosodica.

In particolare sembra che si possano identificare luoghi della TU particolarmente soggetti ad accogliere fenomeni di riduzione delle strutture sillabiche e della timbrica vocalica e luoghi in qualche modo 'protetti' rispetto a tale processo. Questa suddivisione è correlata da un lato ad un puro fattore posizionale, dall'altro alll'intero schema accentuale dell'unità.

In questa comunicazione verranno analizzate, esemplificate e discusse alcune condizioni prosodiche che favoriscono o inibiscono la riduzione segmentale; le prime presentano come caratteristica comune la presenza di una prominenza accentuale immediatamente prima della sede del fenomeno di riduzione.

In altre parole, si intende dimostrare che la riduzione sul piano segmentale è più probabile e più frequente in posizione seguente un *pitch accent*, mentre risulta in qualche modo bloccata nelle posizioni precedenti un *pitch accent*.

Verrà proposta, infine, un'interpretazione del rapporto tra distribuzione delle riduzioni segmentali e schema accentuale nei termini di un principio fisiologico sottostante, alla base della produzione di ciascuna unità intonazionale: in stretta connessione con il meccanismo della respirazione, ogni unità viene prodotta attraverso una fase di *impostazione* seguita da una fase di *declinazione modulata*.

Tutto ciò che avviene nella fase di impostazione fino alla segnalazione di una prominenza è preservato dai fenomeni di riduzione che si verificano, invece, con il rilassamento della tensione necessaria alla produzione, dopo la prominenza (e sul finire della stringa).

L'intero processo sembra pertanto governato da una 'legge del minimo sforzo' articolatorio, violata solo in concomitanza con le parti che vengono poste in rilievo.

Verso un dimensionamento consonantico-temporale dell'italiano sardo-campidanese.

Dr. Carlo Schirru

Dipartimento di Linguistica

Universita' di Padova Via B. Pellegrino 1

Tel.+ 39 49-827 49 11

Fax:39 49-827 4919

35137 PADOVA (ITALY)

E-mail: schirrc@maldura.unipd.it

Abstract

A seguito di uno studio pilota sul vocalismo dell'area campidanese della Sardegna [4], si intende qui portare un primo parziale contributo allo studio sperimentale del consonantismo corrispondente. Riferita all'aspetto temporale, l'analisi intende focalizzare alcune fra le caratteristiche peculiari espresse dalla variante in oggetto, relativamente a tre località site rispettivamente al centro, al

[4] **Schirru C. (1994**), *Aspetti vocalico-temporali dell'italiano in Sardegna. Primi dati sperimentali*, Atti delle 4e Giornate di Studio del G. F. S. (A.I.A.), Torino, 11-12 Novembre 1993, XXI: 131-140.

nord e al sud di una fascia mediana dello stesso Campidano. Costituito da una serie di registrazioni-intervista - tesa a massimizzare il grado di naturalezza e omogeneità - il corpus contiene le dichiarazioni delle proprie generalità da parte di locutori diversificati in funzione dell'età, del sesso e del grado di scolarizzazione. Il tutto, in funzione di vari campi applicativi quali il linguistico, l'industriale, il forense, il medico.

Progetto per la sperimentazione di un tutor computerizzato per l'apprendimento della prosodia dell'inglese per italofoni: il modulo prosodico dello SLIM di Venezia.

Anna Zanfei & Cesare Gagliardi
Centro Linguistico di Ateneo
Università degli Studi di Verona
Via S.Francesco, 29
37129 - VERONA
Tel.:0458009844 Fax: 0458009372
E-mail: azanfei@chiostro.univr.it

La metodologia della ricerca in linguistica applicata si avvale oggi di metodi qualitativi e quantitativi. Cionostante la letteratura in merito a metodi e strumenti di misurazione accreditati è molto più solida dal lato quantitativo. La sperimentazione non può quindi non partire da questi strumenti. In questo intervento vengono perciò presentare le procedure e i metodi utilizzabili, con la costruzione di test adeguati, per la sperimentazione del sistema di riconoscimento del il modulo prosodico dello SLIM di Venezia che mira a sviluppare la prosodia dell'inglese come L2 per studenti adulti. La sperimentazione sarà attuata secondo questo progetto presso il Centro Linguistico di Ateneo dell'università di Verona. Il metodo scelto è uno dei più classici della metodologia quantitativa mentre il pre-test e il post-test sono assolutamente originali. I risultati di questa indagine verranno pubblicati in seguito. Del modulo prosodico preso in considerazione non esistono antecendenti nella formazione assistita da computer: in Italia oggi questi strumenti sono i primi ad avere un'applicazione didattica ed in particolare nella formazione linguistica autonoma di italofoni sono da considersi un prototipo di riferimento. Infine la sperimentazione rivelerà se l'utilizzo di un

sistema di tutoraggio automatico per la prosodia così come è stato concepito e realizzato all'interno dello SLIM abbia un effettivo ed efficace riscontro nell'iter dell'italofono che apprende l'inglese come seconda lingua. La percezione e la produzione della prosodia dell'inglese parlato nello SLIM è affrontata tramite attività di apprendimento a livello interno della parola e attività a livello di frase fonologica. Il riconoscimento degli enunciati dello studente che imita una voce master fornisce un feedback che permette all'utente di visualizzare istantaneamente dove e in quale misura la lunghezza dei foni emessi dalla voce master differisce dai propri. La visualizzazione della propria performance dovrebbe essere di aiuto per un'immediata valutazione della stessa che permette all'utente di organizzare mentalmente l'intensità e la lunghezza con cui emettere i suoni che compongono una parola e una strategia nel caso di una frase di riferimento. I risultati potrebbero indicare in particolare per gli italofoni quali aiuti si rivelano più efficaci e se sviluppare ulteriormente il modulo prosodico. La pratica operativa rimane comunque un ambiente creativo anche se basato sulla falsariga di una metodologia già data e per questo motivo la sperimentazione sul campo ha sempre un valore significativo sia per le problematiche poste nell'attuazione del progetto che per i risultati attesi o disattesi che siano. La parte dell'indagine qualitativa invece sarà svolta in un secondo momento non perché considerata di minore importanza ma piuttosto per una ragione di tempi di elaborazione e di costruzione del progetto relativo.

Dinamiche Articolatorie nella Produzione Verbale Fluente di Normoparlanti e Balbuzienti

Claudio Zmarich

Istituto di Fonetica e Dialettologia – C.N.R.
Via G. Anghinoni, 10 - 35121 Padova (ITALY)
e-mail: zmarich@csrf.pd.cnr.it
www: http://www.csrf.pd.cnr.it

Per descrivere, valutare, diagnosticare e riabilitare varie patologie articolatorie è fondamentale disporre di dati di analisi qualitativi e quantitativi, affidabili ed esaustivi, dei movimenti degli organi articolatori, in termini di spostamento, durata, velocità e accelerazione. Questi dati permettono di controllare e integrare i tradizionali metodi di descrizione, e cioè la trascrizione fonetica su base uditivo-percettiva e l'analisi elettroacustica, le cui limitazioni risiedono rispettivamente nella soggettività della valutazione del percetto uditivo e nell’assenza di biunivocità tra dato acustico e dato articolatorio.

Le limitazioni della trascrizione fonetica e dell'analisi elettroacustica risultano ancor

più evidenti in quei casi in cui un parlato "fluente" dal punto di vista percettivo viene invece prodotto con movimenti articolatori parzialmente o del tutto anomali. E' questo un caso non infrequente con i soggetti balbuzienti: alcuni di loro non presentano le ripetizioni di parte di parola e i prolungamenti di suono che sono considerati tradizionalmente i sintomi della balbuzie, ma avvertono spesso sensazioni di "sforzo" muscolare e tensione cognitiva che difficilmente sono colti dall'osservatore. Se con disordine della fluenza o balbuzie definiamo un quadro sintomatologico in cui viene a mancare o è ridotta la capacità di produrre enunciati privi di discontinuità in modo rapido e senza sforzo, allora anche queste persone dovrebbero essere considerate balbuzienti. Inoltre lo studio della produzione verbale percettivamente fluente dei balbuzienti ha un notevole significato teorico, non solo per i clinici ma anche per i fonetisti, perché lo sforzo di chiarire la natura della loro fluenza porterebbe ad approfondire la riflessione sul concetto di "fluenza" dei normoparlanti. Lo studio della balbuzie fluente è poi importante perchè permette di isolare i sintomi legati direttamente alla menomazione (sintomi primari) dai sintomi che invece indicherebbero un comportamento di reazione o compensazione della menomazione (sintomi secondari).

Il problema pricipale insito in tale tipo di ricerca e' però rappresentato dalla grande capacità compensativa esibita dalle strutture articolatorie in condizioni normali, che rende difficile distinguere la normalità dalla patologia. Ciò è ben riassunto da Folkins (1991), che a proposito della produzione verbale normale invita a tenere presenti le compensazioni definite "flessibili" e quelle definite "plastiche". Le prime consentono il ricorso a diverse alternative strutturate all'interno di un dato sistema in equilibrio per gestire contesti fonetici di tipo diverso, e fenomeni quali l'accento, l'intonazione, la velocità di esecuzione ecc. Le seconde vengono utilizzate in presenza di perturbazioni dell'equilibrio del sistema esistente per raggiungere un nuovo stato di equilibrio (ad es., in situazioni quali parlare mentre si mangia, con la sigaretta tra le labbra, mentre si sta correndo ecc.). A questo proposito Folkins invita a ricordare che "*processi inusuali in una parte o livello del sistema possono produrre compensazioni sia plastiche che flessibili in altre parti o livelli. Solamente quando si superano i limiti di plasticità e flessibilità dell'intero sistema i processi inusuali compromettono l'output comportamentale. Non possiamo definire la fisiologia atipica come patologica finché non abbiamo esaminato i limiti di flessibilità e plasticità per comprendere come i livelli fisiologici interagiscono per produrre l'output comportamentale indesiderato*".

In questa linea di ricerca diventa importantissima l'identificazione delle strategie del controllo motorio. La parola strategia si riferisce ad una regolazione preferenziale di parametri attraverso cui è possibile manipolare caratteristiche dei singoli movimenti, come la velocità di spostamento ecc. Tali strategie di controllo possono essere adottate, all'interno di una più vasta organizzazione neuro-muscolare e attraverso un apprendimento di tipo subconsciente, utilizzando sinergie funzionali che riducono i gradi di libertà del sistema motorio per la produzione del parlato, poichè vincolano reciprocamente i movimenti articolatori individuali in strutture coordinative o gesti. In questo modo ogni individuo può acquisire un pattern di coordinazione motoria stabile di fronte a variazioni nell'ampiezza, nella durata o nella velocità dei movimenti. Questo modello di organizzazione motoria, conosciuta come *Task Dynamic model* (Saltzman e Munhall, 1989), e, nella sua versione più fonologica, come "fonologia articolatoria" (Browman & Goldstein, 1986, 1996) si oppone all'impostazione più classica, in cui gli articolatori sono controllati in modo individuale e indipendente.

In questi anni più recenti, gli studi sulla balbuzie di tipo eziologico si sono progressivamente concentrati sul livello fisiologico della menomazione. Quasi tutti gli studi che hanno trattao l'argomento hanno trovato differenze significative tra il comportamento motorio dei b/i e quello dei normoparlanti. Di questi sforzi sono testimonianza le tre conferenze internazionali organizzate dall'Università di Nijmegen (Olanda) nel 1985, nel 1990 e nel 1996. Quest'approccio fornisce una chiave privilegiata di comprensione unitaria di quel fenomeno multidimensionale che è la balbuzie, che appare condizionata da variabili di natura socioculturale, psicologica, genetica, fisiologica etc. Questa disperante varietà può essere utilmente semplificata con la considerazione che, per ricoprire un certo ruolo causale nella balbuzie, ciascuna di queste variabili deve alla fine influenzare direttamente o indirettamente i processi del controllo motorio del sistema pneumo-fono-articolatorio.

In questo studio saranno presentati alcuni indici di natura cinematica e dinamica che permetterebbero di differenziare in modo non banale la produzione verbale fluente dei balbuzienti da quella dei normoparlanti. Una particolare enfasi verrà portata sui risultati di un esperimento condotto con la strumentazione ELITE da Zmarich e Magno Caldognetto (1995), che partendo dalla considerazione che la presenza di profili di velocità (del gesto articolatorio di apertura o chiusura bilabiale) con un unico picco può essere un indice di normalità per la coordinazione di sistemi articolatori complessi, hanno scoperto che i balbuzienti esibivano una percentuale molto maggiore di curve di velocità irregolari (perche'

dotate di picchi multipli) rispetto ai non balbuzienti. Questi risultati sono stati spiegati ricorrendo all'ipotesi che i balbuzienti fanno un uso più intenso e continuo di un meccanismo di feedback articolatorio inusuale, quello propiocettivo-cinestetico (poichè le curve di velocità con più picchi riflettono una sequenza di submovimenti usati per effettuare adeguamenti spaziali e temporali durante il movimento principale), probabilmente a causa di un deficit nella fase pianificatrice della parametrizzazione ottimale della forza muscolare.

Bibliografia

Folkins J.W.: "Stuttering from a speech motor perspective". In H.F.M., Peters, W, Hulstijn, & C.W., Starkweather, (Eds.), *Speech motor control and stuttering*, Excerpta Medica-Elsevier, Amsterdam, 1991, 561- 579.

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