STABILITA' DEI PARAMETRI NELLO SPEAKER RECOGNITION: LA VARIABILITA' INTRA E INTER PARLATORE

(N.B.: presente solo come abstract)

Luciano Romito & Rosita Lio
Laboratorio di Fonetica
Università della Calabria
luciano.romito@unical.it; lio.rosita@libero.it

SOMMARIO

Un sistema di riconoscimento del parlatore ha lo scopo primario di identificare una persona attraverso la sua voce. Deve, innanzi tutto, ricercare quelle informazioni quanto più *oggettive* possibili presenti nella voce umana ed analizzare la produzione di un parlatore senza interessarsi della sfera semantica, della produzione linguistica[1], o della costruzione sintattica e morfologica.

La voce è molto più di una semplice sequenza di suoni, è intrinsecamente complessa e gran parte della sua complessità è legata ai rapporti tra le singole variabili che operano al suo interno come ad esempio il senso, il significato, le intenzioni, le emozioni, lo stato di salute, lo stato sociale, il livello di autostima, il livello di scolarizzazione ecc. Si veda, a tal proposito, quanto riportato in J. Laver, Principles of Phonetics (1994:2) <<The voice is the very emblem of the speaker, indelibly woven into the fabric of speech. In this sense, each of our utterances of spoken language carries not only its own message, but through accent, tone of voice and habitual voice quality it is at the same time an audible declaration of our membership of particular social and regional groups, of our individual physical and psychological identity, and of our momentary mood.>>. Tutto ciò, ovviamente, assume una maggiore importanza dal punto di vista forense (almeno potenzialmente), ma è, allo stesso tempo, molto difficile da analizzare e stimare. Le variazioni del parlato spontaneo, dipendono essenzialmente da un livello *Paradigmatico*, responsabile della sintassi, della morfologia, della semantica, della costituzione della frase e delle parole, della programmazione dell'intonazione, degli accenti primari e secondari, dell'isocronia e quindi dell'uso del tempo, e da un livello *Sintagmatico,* responsabile delle variabili diafasiche, della velocità di eloquio, delle regole fonologiche, delle variabili diatopiche, della centralizzazione (delle vocali toniche e delle vocali atone), della cancellazione, della riduzione, della neutralizzazione, della declinazione, delle variabili diastratiche, ecc.

Una comparazione di voci è un'analisi estremamente complessa. Nella maggior parte dei casi il modo corretto per valutare i campioni di parlato e di conseguenza valutare il peso delle variabili fonetiche(*-forensi*) è quello di stimare la probabilità e osservare la variabilità interparlatore e intraparlatore. Questo metodo è intrinsecamente probabilistico e non può condurre *mai* ad una assoluta identificazione o esclusione del sospetto.

Acusticamente esistono molti parametri che possono essere utilizzati per comparare due voci e la loro scelta è determinata da una approfondita analisi linguistica. Ovviamente, non esistono parametri ideali ma solo alcune caratteristiche da soddisfare:

[1] Questo soprattutto perché oggi sempre più l'esperto si trova a dover trattare segnali sonori intercettati di breve durata. Ovviamente, qualunque analisi linguistica che tende al riconoscimento del parlatore fallisce in partenza.

a) mostrare una alta variabilità interparlatare e una bassa variabilità intraparltore;
b) essere resistente al camuffamento;
c) avere una alta frequenza di occorrenza;
d) essere robusto durante la trasmissione;
e) essere relativamente facile da estrarre e misurare.

Sui metodi utilizzati per lo speaker recognition in Italia e nel mondo, la letteratura è veramente abbondante (si veda Romito–Galatà (2006) per l'Italia e P. Rose (2002) per il resto).

In generale, esistono tre grandi famiglie di metodi di SR: uditivo-percettivi, parametrici e completamente automatici.

Tra gli uditivo-percettivi, i metodi riconosciuti sono:

1. metodo uditivo attraverso ascoltatori inesperti;
2. o attraverso un campione ristretto di esperti fonetisti (trained phonetician):
    a. comparazioni di single vs multiple choice;
    b. comparazioni di familar vs unfamilar voices;
3. metodo del *Panel Approach*, comparazione di coppie di frasi e risposta in percentuale. Tale metodo prevede sia risposte di tipo qualitativo sia identificazione di parti molto tecniche ed acustiche;
4. *Direct processing*, dove un esperto ascolta un intero brano e identifica la voce;
5. *Aural-Perceptual Approach* dove all'esperto vengono chieste informazioni precise come la valutazione del pitch (level, variability, patterns);
6. *Aural-Spectrografic identification* dove l'esperto confronta e compara contemporaneamente sia i sonogrammi che audio.

Si tratta di metodi basati sulla capacità che la specie umana possiede di analizzare e riconoscere le voci come appartenenti alla stessa persona o a persone differenti. Nonostante questa riconosciuta competenza, però, si tratta comunque di metodi soggettivi e i parametri utilizzati non sempre vengono specificati.

Ulteriori limiti nascono dal fatto che non tutti possiedono la stessa abilità (Ladefoged and Ladefoged 1980:45; Hollien 1995:15, Foulkes and Barron 2000:182), che alcune voci sono più facilmente identificabili (Popçun et al. 1989, Rose and Duncan 1995:12,16), altre sono più simili tra loro, e infine, l'esperto fonetista che ascolta, e quindi con competenza giudica, non è automaticamente un esperto *riconoscitore* o un esperto perito[2].

I metodi automatici, invece, (si veda Hollien 2002, Furui 1989, Rose 2002) si basano esclusivamente su parametri oggettivi scorrelati da ogni impostazione articolatoria. L'uso, per esempio, del terzo e quarto coefficiente *cepstrale* non presenta alcuna correlazione articolatoria (anche se gli studi di Clermont e Itahashi (1999) tentano di dimostrare che la

[2] Non esiste quindi una professionalità, universalmente riconosciuta, basata su tale competenza, si veda Nolan 1983:17 dove si riporta che nel Meeting del 1980 del British Association of Academic Phoneticiens fu approvata la seguente mozione <<*Phoneticians should not consider themselves expert in speaker identification until they have demonstrated themselves to be so*>>.

qualità vocalica potrebbe essere interpretata come variazione del II e III coefficiente cepstrale)[3].

Secondo gli studi di Ladefoged (2001:78-95), gli uomini e i computer riconoscono le voci attraverso procedimenti completamente differenti, quindi è anche ovvio che i due metodi, uditivo percettivo e automatico, abbiano parametri che siano completamente differenti.

Il giudice in quanto uomo, notoriamente preferisce parametri che abbiano una correlazione con le impostazioni articolatorie. Grazie all'Acoustic Theory of Speech Production, il comportamento di alcuni parametri acustici è articolatoriamente interpretabile. Il metodo parametrico nasce e si sviluppa proprio grazie a questa correlazione.

Tale metodo per essere definito oggettivo e quindi godere di rilevanza nell'ambito delle procedure atte all'identificazione del parlatore, deve basarsi su parametri acustici, strettamente dipendenti dalla voce del singolo parlatore e quindi, fortemente caratterizzante[4], che godono di precise caratteristiche, svincolate, per quanto possibile da informazioni linguistiche e soprattutto stabili[5].

Questo lavoro si prefigge di testare la stabilità dei parametri utilizzati in ambito di SR attraverso lo studio della *variabilità intraparlatore* e della *variabilità interparlatore* A tale fine viene utilizzato come luogo d'indagine il corpus PRIMULA[6].

PRIMULA è un corpus ristretto di voci calabresi ideato e creato presso il Laboratorio di Fonetica dell'Università della Calabria per la valutazione delle metodologie e dei sistemi di riconoscimento del parlatore con particolare attenzione all'ambito forense. Allo scopo di simulare una situazione reale al fine di avere, a prodotto finito, situazioni simili o quantomeno assai vicine a quelle che si presentano di norma nella maggior parte dei casi forensi, sono state effettuate delle registrazioni con attrezzature normalmente utilizzate per le intercettazioni. È stato così possibile registrare lo stesso materiale prodotto sia attraverso la microspia installata su un'autovettura sia attraverso un cellulare collegato con un telefono fisso presso il Laboratorio di Fonetica dove la registrazione veniva acquisita su un registratore DAT. Il materiale registrato e così derivato ha portato quindi ad avere una intercettazione ambientale (in automobile) e una registrazione telefonica (tra utenza cellulare e utenza di rete fissa).

All'interno del corpus sono presenti la voce di 5 interlocutori maschili di simile statura, peso e classe di età. I tipi di registrazione sono Ortofonico (in camera silente), Ambientale e

---

[3] Anche uno degli autori del presente lavoro collabora in una ricerca internazionale Italia-Usa, su un metodo completamente automatico basato sulle funzioni dissipative (progetto Interlink tra l'Università della Calabria la Mason University di Washington e la Chapman University di Los Angeles).

[4] Romito L. (2000), Manuale di fonetica articolatoria, acustica e forense, Centro Editoriale e Librario, Unical.

[5] Tra l'altro la scienza in merito ha opinioni molto controverse: per Badwin (1979), Baldwin and French (1990:9) il dato uditivo è sufficiente, mentre invece non è assolutamente necessario per Furui (1989). Kunzel (1987), (1995:76-81); French (1994:173-4) ritengono invece che bisogna combinare le due tecniche acustico e uditivo.

[6] Romito L., Galatà V. (2006), Speaker Recongnition: stato dell'arte in Italia. Valutazione dei corpora, dei metodi e delle professionalità coinvolte, in Atti III° Convegno AISV (Associazione Italiana Scienze della Voce) "Scienze Vocali e del Linguaggio Metodologie di Valutazione e Risorse Linguistiche" Trento, 29-30 Novembre - 1 Dicembre 2006.

Telefonico. Per ogni tipo si hanno registrazioni di lettura di tre frasi foneticamente bilanciate ripetute da 10 a 50 volte, lettura di 10 frasi singole e diverse sessioni di parlato spontaneo sia in dialetto calabrese che italiano regionale. Per studiare e verificare l'influenza del canale è stata effettuata come già detto la stessa identica registrazione sia in modalità ambientale (intercettazione) che attraverso il telefono cellulare. Per studiare e valutare l'influenza del rumore e l'intensità del locutore in presenza di rumore abbiamo la stessa registrazione in strada, ad una fermata di autobus, in un aula universitaria molto rumorosa e in automobile con finestrino aperto. Tutte le registrazioni sono state acquisite, nonostante la presenza dei diversi canali di registrazione, in formato *.wav con una frequenza di campionamento di 44100 Hz, 24-bit in modalità monoaurale.

Partendo, dunque, da queste impostazioni metodologiche, questo lavoro si prefigge di investigare sui seguenti punti:

1. studio dei parametri formatici (F1, F2, F3) e della Frequenza Fondamentale (F0) nelle vocali sia toniche che atone;
2. studio della velocità d'eloquio in particolar modo per quanto riguarda l'articulation rate[7] ;
3. variabilità intraparlatore, verrà studiata attraverso le ripetizioni delle singole frasi, in contesti differenti (aula, strada fermata autobus, telefono) e attraverso modalità diversa (voce Alta, Normale e Bassa). Verrà anche studiata la differenza tra lettura, ripetizione e parlato spontaneo;
4. variabilità del canale, verranno studiate le variazioni dei valori formantici indotte dal canale di trasmissione (unica sessione di registrazione su canali differenti in camera silente - microspia – telefono fisso ecc);
5. variabilità interparlatore, ovviamente lo studio comparato delle analisi effettuate sul singolo parlatore condurrà allo studio della variabilità interparlatore;
6. verrà anche analizzata l'influenza della variabile diafasica sulla velocità di elocuzione sia per lo studio della variabilità intraparlatore che per quella interparlatore;
7. verrà analizzato l'effetto del rumore esterno sull'innalzamento dell'intensità del parlatore e il conseguente effetto sul valore della frequenza fondamentale.

[7] Kunzel (1997) e Zavattaro (Tesi di Dottorato).