LA SEGMENTAZIONE DELLE OCCLUSIVE DELL'ITALIANO MEDIANTE SLAM

Piero Cosi

Istituto di Fonetica e Dialettologia - C.N.R.
Via G.Anghinoni, 10 - 35131 Padova (Italy)
Tel.: 049 8274421
Fax: 049 8274416
E-Mail: cosi@csrf.pd.cnr.it

SOMMARIO

Viene descritta la procedura utilizzata per la segmentazione delle occlusive dell'italiano raccolte in un data-base di prossima distribuzione nell'ambito del progetto infrastrutturale BIBLOS: Proposta di automazione dei cataloghi bibliografici, degli Organi di Ricerca del Consiglio Nazionale delle Ricerche, per lo sviluppo e l'integrazione dei servizi telematici degli Organi afferenti al Comitato Nazionale per le Scienze storiche, filosofiche e filologiche del CNR. Dopo una breve illustrazione del corpora delle occlusive dell'italiano, viene descritto, nei suoi particolari, il sistema di segmentazione semi-automatico, denominato SLAM (dall'acronimo inglese *Segmentation and Labelling Automatic Module*), progettato ed utilizzato, presso il CSRF, allo scopo di velocizzare ed uniformare la complicata ed onerosa operazione di segmentazione di grossi data-base di segnale vocale. SLAM utilizza una rappresentazione spettrale prodotta mediante l'applicazione di un sistema di analisi fortemente ispirato alle funzionalità del sistema uditivo periferico umano, e per questo risulta particolarmente efficace anche nel caso di segnali registrati in ambienti particolarmente rumorosi.

INTRODUZIONE

Numerosi sono ormai i corpora di segnale vocale prodotti da varie organizzazioni e gruppi di ricerca nel campo delle scienze linguistiche e fonetiche in campo nazionale ed internazionale [1], [2], come è possibile osservare, infatti, in un'interessante rassegna sull'ingegneria linguistica recentemente apparse in Internet [3]. Per trasformare questi corpora in materiali scientificamente utilizzabili in vari campi di ricerca applicata e non, finalizzati essenzialmente alle tematiche riguardanti l'analisi, la sintesi ed il riconoscimento automatico della voce, è ovviamente necessario trasformarli in database organizzati e facilmente consultabili. Una delle prime elaborazioni di cui tutti questi corpora necessitano e che spesso vengono tralasciate in fase di progettazione, è senza dubbio la segmentazione e l'etichettatura (dall'inglese *labelling*), a vari livelli (semantico, lessicale, ortografico, fonetico), del segnale

verbale oggetto dei corpora stessi. Non infrequente è, infatti, la presenza nei vari laboratori di ricerca di corpora di segnale verbale la cui utilità scientifica, nonostante la loro complessità, è ridotta quasi a zero a causa della mancanza di questa necessaria elaborazione.

Purtroppo un primo grosso inconveniente introdotto dalla necessità di associare ai corpora un'opportuna etichettatura, e questo soprattutto a livello fonetico, risiede nel fatto che, normalmente, questa viene affidata all'opera manuale di esperti linguisti o fonetisti e, di conseguenza, costituisce un significativo collo di bottiglia a causa dell'enorme spreco di risorse, sia temporali che economiche, che una tale operazione necessariamente richiede.

Nonostante queste considerazioni ovvie e condivisibili a livello scientifico, raramente da parte degli enti preposti alla progettazione ed alla realizzazione dei vari corpora viene riposta quell'attenzione che dovrebbe invece permeare tutte le fasi di sviluppo dei corpora stessi. In altre parole alla fase di definizione dei corpora ed alla loro successiva acquisizione vengono affidate tutte le risorse, scientifiche ed economiche, dimenticandosi dell'organizzazione finale e quindi delle finalità essenziali dei corpora stessi che risiedono ovviamente nella loro effettiva utilizzazione per il progredire della ricerca scientifica nel campo dell'ingegneria linguistica.

Un altro problema di non facile soluzione è rappresentato dal fatto che l'etichettatura manuale è sempre caratterizzata da un elevata anche se controllabile soggettività [4], [5]. Infatti, nonostante l'ausilio di sempre più affidabili strumenti audio visivi, le divergenze nei valori di segmentazione dello stesso materiale vocale, prodotti manualmente da parte di più esperti, non potranno mai essere completamente eliminate. A causa delle diverse capacità percettive, sia visive che uditive, come anche dell'oggettiva difficoltà di definire una inequivocabile strategia comune, è evidente l'implicita incoerenza di un tale approccio manuale. Sulla base di queste considerazioni, l'interesse per la realizzazione di sistemi automatici di segmentazione e *labelling* è ovviamente elevatissimo. Tali sistemi automatici, oltre a minimizzare i tempi di esecuzione, renderebbero implicitamente coerenti i risultati della segmentazione. Infatti, gli eventuali errori di segmentazione risulterebbero facilmente identificabili e categorizzabili a causa della natura algoritmica delle procedure.

METODO

Il sistema semi-automatico di segmentazione descritto in questo lavoro, denominato SLAM (dall'acronimo *inglese Segmentation and Labelling Automatic Module*) [6], studiato e progettato per fornire una risposta pratica a tutti le difficoltà elencate nell'introduzione precedente, fornisce in modo automatico alcune ipotesi di segmentazione allo scopo di ridurre al minimo il compito di esperti fonetisti nell'analizzare grossi corpora di segnale verbale. Nessun istante di segmentazione viene posizionato manualmente, salvo rari casi, e agli esperti viene esclusivamente richiesta un'azione di supervisione sulle ipotesi di segmentazione prodotte automaticamente dal sistema. Gli esperti, infatti, devono scegliere, sulla base della conoscenza ortografica del testo pronunciato, l'allineamento più opportuno fra quelli proposti automaticamente, eventualmente eliminando "marker" sovrabbondanti.

SLAM ottiene la segmentazione del segnale verbale in ingresso essenzialmente in tre fasi: nelle prime due fasi corrispondenti all'elaborazione digitale del segnale verbale ed alla costruzione delle ipotesi di segmentazione, il sistema opera in modo automatico, mentre nella terza fase viene esplicitamente richiesta la collaborazione

interattiva dell'utente che deve scegliere la segmentazione finale sulla base delle ipotesi propostegli. dal sistema. SLAM riceve in ingresso i parametri forniti da un modello del sistema uditivo periferico [7] dimostratosi molto efficace, anche in condizioni rumorose [8], nel codificare le informazioni contenute nel segnale vocale importanti per una corretta segmentazione. Per quanto riguarda l'individuazione dei possibili confini di separazione fra le varie unità, SLAM si basa sulla teoria della segmentazione multi-livello [9], [10], mediante la quale è possibile evidenziare all'interno di un'unica struttura sia i mutamenti rapidi che quelli graduali riscontrabili sul segnale. Per una descrizione dettagliata del sistema si veda [6], [8].

MATERIALI

Il corpus vocale oggetto di questo lavoro si riferisce alle occlusive dell'italiano ed è incluso nei corpora che verranno prossimamente distribuiti all'interno del progetto infrastrutturale del CNR denominato BIBLOS [11], ideato appositamente per lo sviluppo e l'integrazione dei servizi telematici (dati bibliografici, corpora testuali, corpora vocali....) degli Organi di Ricerca afferenti al Comitato Nazionale per le Scienze storiche, filosofiche e filologiche del CNR. In particolare ci si riferisce al corpus vocale denominato MIC-ART 1 contenente il segnale microfonico ed articolatorio (ELITE [12]) relativo alle occlusive dell'italiano (/'VCV/, C=/p,t,k,b,d,g/, V=/a,i,u/) pronunciate 5 volte da 10 soggetti maschi. Per ogni stimolo sono disponibili il segnale microfonico digitalizzato a 16KHz su 16bit PCM, 28 segnali, digitalizzati a 100Hz su 16bit PCM, corrispondenti a movimenti articolatori, catturati in tempo reale dal sistema ELITE [12], di alcuni marker posizionati sulle labbra dei soggetti ed infine la segmentazione e l'etichettatura in formato ASCII. Nelle figure seguenti sono illustrati alcuni casi di segmentazione da cui si può osservare l'efficacia della procedura.

CONCLUSIONI

La procedura descritta in questo lavoro ha contribuito a ridurre sensibilmente i tempi di segmentazione del corpus delle occlusive dell'italiano denominato MIC-ART 1 e si propone come standard futuro per la segmentazione dei corpora vocali che verranno distribuiti dall'Istituto di Fonetica e Dialettologia all'interno del progetto BIBLOS.

BIBLIOGRAFIA

[1] LDC: Linguistic Data Consortium, *Internet www page address: http://www.ldc.upenn.edu/*.

[2] ELRA: European Language Resources Association, *Internet www page address: http://www.icp.grenet.fr/ELRA/home.html*.

[3] Ingegneria Linguistica, *Internet www page address: http://comel.ing.uniroma1.it/~sandro/lengeng.htm*.

[4] P. Cosi, D. Falavigna and M. Omologo, "A Preliminary Statistical Evaluation of Manual and Automatic Segmentation Discrepancies", *Proc. Eurospeech-91*, Genova, 24-26 September, 1991, pp. 693-696.

[5] T. Lander, B. Oshika, J. Carlson, T. Durham, and T. Bailey, "Analysis of Inter-Labeler (dis)agreement in Phonetic Transcriptions of Multiple Languages", *Proc. of the Acoustical Society of America*, Waikiki, Hawaii, December 1996.

[6] P. Cosi, "SLAM v1.0 for Windows: a Simple PC-Based Tool for Segmentation and Labeling", *Proc. of ICSPAT-97*, San Diego, CA, USA, September 14-17, 1997, pp. 1714-1718.
[7] S. Seneff, "A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing", *Journal of Phonetics*, January 1988, pp. 55-76.
[8] P. Cosi, "Ear Modelling for Speech Analysis and Recognition", *ESCA Workshop-92*, Sheffield, 7-9 Apr, 1992.
[9] J.R. Glass, "Finding Acoustic Regularities in Speech: Application to Phonetic Recognition", *Ph. D. thesis, Massachusetts Institute of Technology*, May 1988.
[10] J.R. Glass and V.W. Zue, "Multi-Level Acoustic Segmentation of Continuous Speech", *Proc. Icassp-88*, New York, N.Y., April 11-14, 1988, pp. 429-432.
[11] P. Cosi, "Il progetto BIBLOS: biblioteca umanistica virtuale degli Organi di ricerca del CNR", *in questo volume*.
[12] Magno Caldognetto E., Vagges K., Borghese N.A., and Ferrigno G., "Automatic Analysis of Lips and Jaw Kinematics in VCV Sequences", *Proc. of Eurospeech 1989*, Vol. 2:453-456.

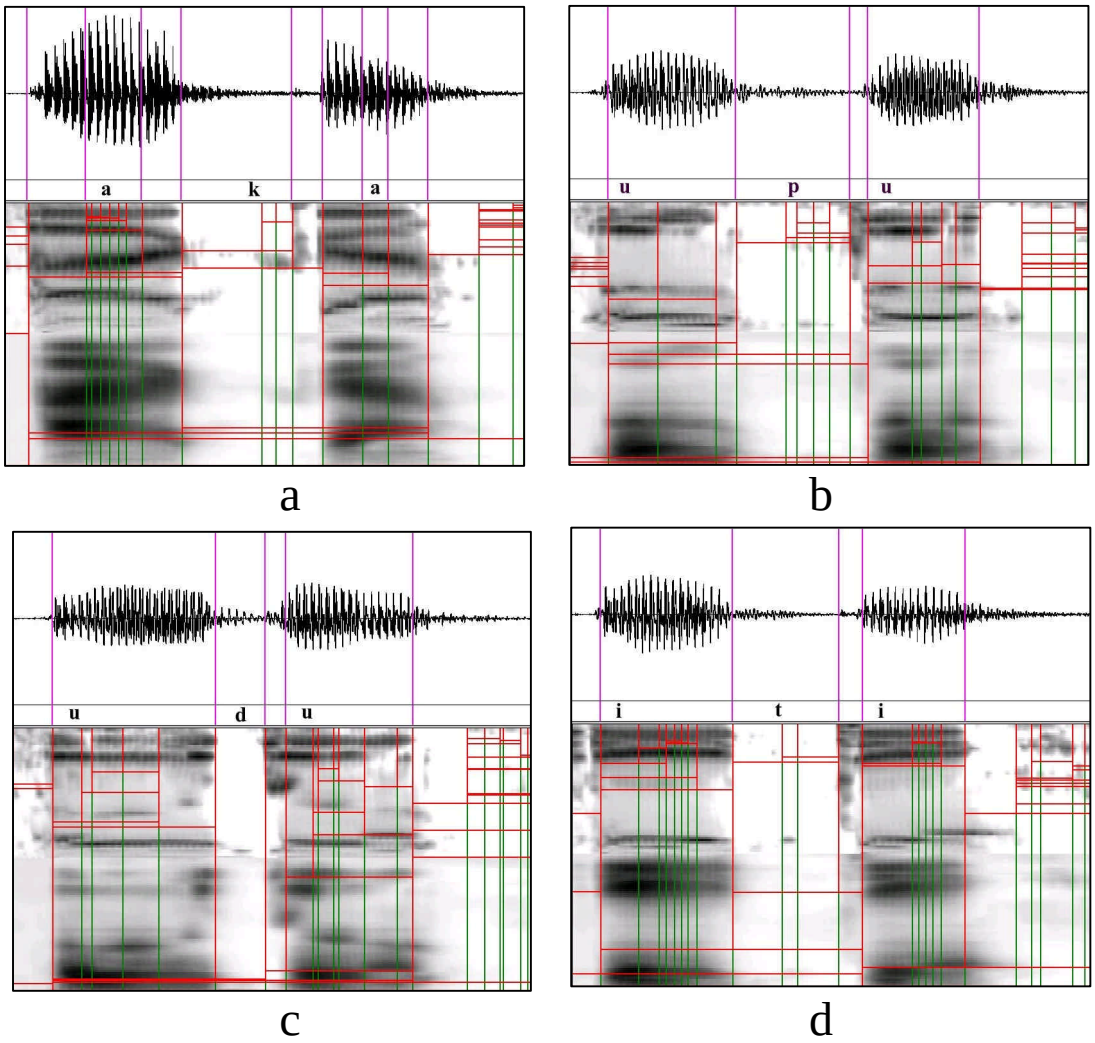

Figura 1. Esempi di segmentazione ottenuti con SLAM relativi agli stimoli /aka/, /upu/, /udu/, /iti/. Risulta evidente come, mediante il reticolo delle ipotesi, possano essere individuate anche zone di particolare interesse quali ad esempio quelle indicate in (a) relative alla zona stabile della vocale /a/.