IL MODULATION SPECTROGRAM NEL RICONOSCIMENTO AUTOMATICO DELLE SILLABE

Gianpaolo Coro

Dipartimento di Scienze Fisiche – gruppo NLP – Università di Napoli "Federico II"
coro@na.infn.it

RIASSUNTO

La scelta delle "**features**", parametri numerici che contengono informazione discriminante su unità linguistiche e che permettono di operare un riconoscimento automatico del Parlato, è una questione fondamentale da affrontare nella progettazione di un ASR (Automatic Speech Recognizer). La tecnica di analisi del segnale chiamata **Modulation Spectrogram** (**MS**) ha avuto interessanti impieghi nella costruzione di sillabatori e di riconoscitori automatici che hanno affrontato il problema del riconoscimento di unità segmentali più ampie del fono (sillabe e mezze sillabe). In questa sede mostrerò dei vantaggi della tecnica MS soprattutto nell'ambito del parlato spontaneo, evidenziandone i vantaggi e gli svantaggi rispetto allo spettrogramma "classico" e la sua resistenza a variazioni ambientali. I risultati mireranno ad indurre un ragionamento sulla necessità della costruzione di riconoscitori automatici "robusti", che siano più indipendenti dalla situazione ambientale (parlato chiaro o rumoroso) nella quale avviene il riconoscimento e che si avvicinino ulteriormente all' emulazione del sistema di elaborazione acustica umano.

La tecnica si basa su un risultato di un'analisi psicolinguistica che evidenzia il fatto che le lente variazioni del parlato hanno, mediamente, una frequenza di 4Hz, approssimativamente coincidente con la cadenza delle sillabe. Questo permette di scandire il segnale mediante finestre di 250 ms analizzando la distribuzione in frequenza dei segmenti in analisi. Si cerca quindi di sfruttare questa regolarità per individuare segmenti sillabici (e quindi costruire a partire da ciò un sillabatore automatico), oppure per ipotizzare che i picchi della trasformata di Fourier a 4 Hz possano distinguere una sillaba dall'altra.
Il MS, nella versione da me presentata, permette anche discussioni sul fatto che il numero di features che esso riesce a fornire per descrivere un segmento sillabico è sicuramente inferiore a quello che le moderne e "pluricollaudate" tecniche di analisi del segnale, quali l'MFCC o il RASTA-PLP, offrono per i foni: nel primo caso si parla di almeno 30 features per segmento, più altrettante per l'informazione contestuale, nel secondo ci sono tipicamente 18 parametri per segmento e l' informazione contestuale non è così fondamentale, inoltre la sua resistenza a fattori come rumore o riverbero la rendono una tecnica più potente rispetto a quelle tradizionali.
La tecnica che presenterò sarà applicata ad un insieme di campioni fonici estratti dai dati messi a disposizione ai partecipanti dagli organizzatori del convegno AISV 2004.