IL RICONOSCIMENTO BIMODALE DEL PARLATO: UN ESEMPIO DI IMPLEMENTAZIONE PER L'ITALIANO

**Nello Galiano, *Mario Refice , **Michelina Savino*

*Dipartimento di Elettrotecnica ed Elettronica, Politecnico di Bari
**Dipartimento di Psicologia, Università di Bari

RIASSUNTO

Il riconoscimento bimodale del parlato rappresenta una sorta di estensione del tradizionale riconoscimento automatico del parlato, nel senso che vengono utilizzate sia le informazioni acustiche sia quelle visive, integrandole opportunamente, per migliorare l'accuratezza del riconoscimento soprattutto in ambienti rumorosi. Questa strategia di integrazione di informazioni acustiche e visive nel riconoscimento del parlato è, d'altra parte, tipica degli esseri umani, come dimostrato sperimentalmente dal celebre "effetto McGurk" (McGurk & MacDonald, 1976).

Da un punto di vista più tecnico la giustificazione a priori viene anche dalla considerazione che il canale visivo può essere considerato ortogonale a quello acustico, capace quindi di fornire informazioni di natura diversa, possibilmente integranti quelle fonetiche. Il rumore acustico inoltre non influenza i dati visivi e questa constatazione rende plausibile il ricorso al secondo canale informativo, soprattutto quando il primo non si presenta nelle migliori condizioni. Da un punto di vista pratico si deve inoltre sottolineare che il costo dei sistemi di acquisizione video, grazie al progresso tecnologico, tende a scendere ed a rendere quindi concretamente possibile l'impiego di queste apparecchiature anche come apparati complementari.

Come in altri tipi di applicazioni, l'utilizzo di sistemi di riconoscimento visivo si fonda sul notevole bagaglio di conoscenze derivanti dal riconoscimento delle immagini in quanto tali.

Pertanto si possono distinguere in generale i metodi basati sull'aspetto da quelli basati sulla forma: nei primi si assume che tutti i pixel dell'immagine, all'interno di una determinata regione di interesse, forniscano informazioni sul fonema pronunciato, mentre nei secondi l'informazione visiva si assume sia contenuta prevalentemente nel contorno del viso e delle labbra del parlante. In quest'ultimo caso si ricorre quindi all'estrazione di *features* di tipo geometrico (larghezza e altezza delle labbra, perimetro e area delle stesse, protusione, ecc.) misurate tramite opportuni parametri. Sviluppo di modelli basati su quest' ultimo tipo di approccio e relative sperimentazioni di riconoscimento bimodale sono stati condotti per l'italiano dal gruppo di ricerca del CNR di Padova (cfr per esempio Cosi & Magno Caldognetto, 1996).

In questo lavoro viene presentata una prima sperimentazione, fatta dal gruppo di ricerca di Bari, nell'utilizzo di tecniche per il riconoscimento bimodale di parlato continuo, dove il metodo utilizzato per il riconoscimento visivo si basa invece sull' aspetto, mentre per il canale acustico i parametri utilizzati sono costituiti dai classici coefficienti MFCC (*Mel Frequency Cepstrum Coefficients*). Le informazioni provenienti dai due canali vengono opportunamente integrate mediante l'utilizzo di modelli statistici "accoppiati" (CHMM, *Coupled Hidden Markov Models);* il sistema pertanto è al momento dipendente dal parlatore. I primi risultati relativi alla *performance* di tale sistema indicano che l'integrazione delle informazioni relative al canale visivo produce un miglioramento del tasso di riconoscimento del parlato del 45% superiore a quello ottenuto basandosi unicamente sulla sorgente di informazione acustica.

Riferimenti

Cosi P., Magno Caldognetto E., *Lips and Jaw Movements for Vowels and Consonants: Spatio-Temporal Characteristics and Bimodal Recognition Applications*, in D.G. Starke & M. E. Henneke (eds.), Speechreading by Humans and Machine: Models, Systems and Applications, NATO ASI, vol.150, Springer-Verlag, 1996, pp.291-313.

McGurk H., McDonald J., *Hearing lips and seeing voices*, Nature 264 (5588), 746-748.