Front-end robusto per il Riconoscimento Automatico del Parlato in presenza di parlatori multipli e di riverbero

Stefano Squartini, Rudy Rotili, Emanuele Principi, Francesco Piazza

Dipartimento di Ingegneria Biomedica, Elettronica e delle Telecomunicazioni
Università Politecnica delle Marche
Via Brecce Biance 1, 60131, Ancona, Italy

Le interfacce uomo-macchina (HMI, Human-Machine Interface) rappresentano un ambito di grande interesse scientifico, con numerose ricadute commerciali. In particolare, l'utilizzo della voce per veicolare l'informazione in scenari applicativi di questo tipo presenta caratteristiche vantaggiose rispetto all'interazione manuale. Uno dei problemi principali da affrontare a riguardo è il riconoscimento automatico del parlato (ASR). I sistemi ASR si basano sull'assunzione che il segnale vocale sia rappresentabile come una sequenza di uno o più simboli. Per realizzare l'operazione di riconoscimento la forma d'onda continua del parlato viene preliminarmente convertita in una sequenza di vettori parametrici discreti (feature vectors). Il compito del riconoscitore è di realizzare un mapping fra le sequenze di feature-vectors e le sequenze di simboli desiderate (words), attraverso un'opportuna procedura di decodifica, che può essere realizzata seguendo approcci diversi.
Le prestazioni del riconoscitore dipendono fortemente dalle condizioni acustiche in cui si trova a operare [1]. Un notevole interesse è stato mostrato negli ultimi anni allo studio e sviluppo di soluzioni robuste a non-idealità acustiche: i segnali che arrivano alla fase di riconoscimento sono in generale deteriorati per effetto di fenomeni quali rumore di fondo, parlatori simultanei e riverbero e necessitano quindi di un trattamento per ristabilirne una qualità accettabile che faciliti il compito del processing successivo.
Nelle memorie passate la ricerca si è orientata a considerare il problema del rumore di fondo [2], ed interessanti sviluppi si sono avuti anche recentemente nel caso multicanale, ovvero in presenza di più microfoni [3,4]. In questa memoria viene invece descritta una architettura circuitale in grado di processare sorgenti vocali multiple in un ambiente acustico riverberato al fine di garantire al sistema ASR susseguente di operare al meglio delle sue potenzialità. Il front-end è stato implementato su piattaforma NU-Tech [5] ed è in grado di operare in tempo-reale: da questo punto di vista presenta una certa appetibilità per applicazioni reali.
Come mostrato in Figura 1.a, l'architettura proposta [5,6] consta di tre stadi principali che assolvono i seguenti compiti rispettivamente: separazione delle sorgenti, dereverbero e stima "blind" delle risposte impulsive (BCI). L'assunzione di base consiste nell'avere un numero di microfoni maggiore alle sorgenti esistenti nell'ambiente. Il sistema MIMO (*multiple input multiple output*) che ne consegue viene trasformato da blocco di separazione in un certo numero di sistemi SIMO (*single input multiple output*) che forniscono in uscita segnali riverberati ma privi di interferenze. Tali segnali vengono quindi dati in pasto al blocco di dereverbero, che tramite una procedura adattativa [7] è in grado di ridurre la distorsione dovuta al riverbero e mettere a disposizione dei sistemi ASR segnali vocali simili a quelli emessi dalle sorgenti. Al fine di far lavorare adeguatamente i due stadi menzionati, è necessario stimare le risposte impulsive MIMO corrispondenti ai canali acustici tra le sorgenti ed i microfoni: lo stadio BCI serve appunto a questo.
La stima delle risposte (assunte debolmente non-stazionarie) non funziona adeguatamente se ci sono più sorgenti contemporaneamente attive. Al fine di ridurre questa limitazione, è stato introdotto un ulteriore blocco operativo, quello indicato in Figura come "Speaker Diarization" [6,8] che è in grado di capire chi parla e quando, e come tale è in grado di pilotare l'attività non solo della BCI (quindi stimare le risposte quando un solo parlatore è attivo) ma anche dell'ASR (che eseguirà i suoi compiti solo in corrispondenza del parlatore a cui è associato).
Sono state eseguite delle simulazioni che hanno premesso di verificare il buon funzionamento degli

algoritmi implementati nell'architettura globale proposta. A tale scopo è stato realizzato un database opportunamente legato allo scenario acustico d'interesse, a partire dal "WSJ" un corpus LVCSR (*large-vocabulary continuous speech recognition*) largamente usato in applicazioni ASR. Sono state considerate due diverse condizioni di riverbero (T60 uguale a 120 e 240ms) e valutate le prestazioni in termini di *word-accuracy* sui file completi del database e sulla parte di essi in corrispondenza dei quali gli algoritmi adattativi del Front-end possono considerarsi arrivati a convergenza. Come si può osservare dalla Figura 1.b un notevole miglioramento è ottenibile rispetto al caso "unprocessed", ovvero quando il Front-end non è attivo.
Sviluppi futuri sono orientati all'ottimizzazione degli algoritmi sviluppati sia in termini di prestazioni che di implementabilità su piattaforme embedded, come recentemente fatto da alcuni autori di questa memoria per l'algoritmo di Speaker Diarization [8].

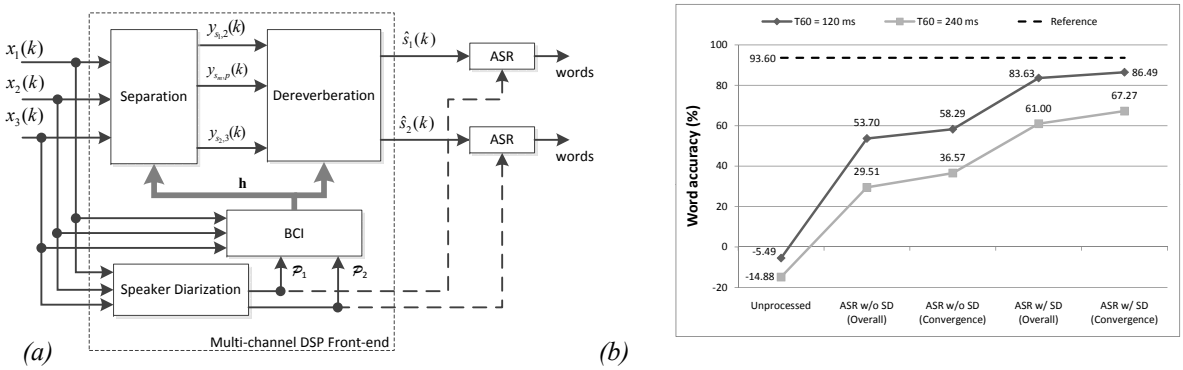

Figura 1. *(a)* Schema a blocchi del Front-end. *(b)* Risultati delle simulazioni

Bibliografia

[1] A. Peinado and J. Segura, Speech Recognition Over Digital Channels. John Wiley & Sons, Ltd, 2006.

[2] Rotili, E. Principi, S. Cifani, F. Piazza, S. Squartini "Multichannel Feature Enhancement for Robust Speech Recognition", *Speech Technologies / Book 1*, Ivo Ipsic Ed., ISBN 978-953-307-152-7, to be published, 2011.

[3] S. Squartini, E. Principi, R. Rotili and F. Piazza, "Environmental Robust Speech and Speaker Recognition through Multi-channel Histogram Equalization", *Neurocomputing*, 2011, accepted for publication.

[4] S. Squartini, E. Ciavattini, A. Lattanzi, D. Zallocco, F. Bettarelli, and F. Piazza, "NU-Tech: implementing DSP algorithms in a plug-in based software platform for real time audio applications," *Proceedings of 118th Convention of the Audio Engineering Society*, 2005.

[5] R. Rotili, C. D. Simone, A. Perelli, S. Cifani, and S. Squartini, "Joint multichannel blind speech separation and dereverberation: A real-time algorithmic implementation," in *Proc. of 2010 International Conference on Intelligent Computing,* 2010, pp. 85-93.

[6] Rudy Rotili, Emanuele Principi, Stefano Squartini, Francesco Piazza, "Real-time Joint Blind Speech Separation and Dereverberation in Presence of Overlapping Speakers", *Advances in Neural Networks - ISNN2011*, eds. D. Liu, H. Zhang,M. Polycarpou, C. Alippi, H.He, Springer LNCS, 2011.

[7] R. Rotili, S. Cifani, E. Principi, S. Squartini, and F. Piazza, "A robust iterative inverse filtering approach for speech dereverberation in presence of disturbances," in *Proc. of APCCAS 2008*, pp. 434-437.

[8] V. Colagiacomo, E. Principi, S. Cifani, and S. Squartini, "Real-time speaker diarization on TI OMAP3530," in *Proc. of EDERC2010*, 2010.