FEATURE WARPING MULTICANALE NEL DOMINIO CEPSTRALE PER IL RICONOSCIMENTO AUTOMATICO DEL PARLATO

Stefano Squartini, Emanuele Principi, Simone Cifani, Rudy Rotili, Francesco Piazza

A3Lab, Dipartimento di Ingegneria Biomedica, Elettronica, e Telecomunicazioni
Università Politecnica delle Marche
Via Brecce Bianche 31, 60131 Ancona

Le interfacce uomo-macchina (HMI, acronimo inglese) rappresentano un ambito di grande interesse scientifico e commerciale. In particolare, l'utilizzo della voce per veicolare l'informazione in scenari applicativi HMI presenta caratteristiche vantaggiose rispetto all'interazione manuale, da cui la ragione del suo utilizzo. Uno dei problemi principali da affrontare a riguardo è il riconoscimento automatico del parlato (ASR, acronimo inglese), per cui lo stato dell'arte ha raggiunto ai giorni nostri un notevole grado di maturità. Gli attuali sistemi ASR hanno percentuali di riconoscimento vicine al 100% (prestazioni ottenibili però in condizioni particolari, ossia su piattaforme potenti e in ambienti acusticamente ideali).
I sistemi per il riconoscimento vocale generalmente si basano sull'assunzione che il segnale vocale sia una realizzazione di alcuni messaggi codificati come una sequenza di uno o più simboli. Per realizzare l'operazione di riconoscimento della sequenza di simboli il segnale vocale viene preliminarmente convertito in una sequenza di *features*, tipicamente MFCC (Mel-Frequency Cepstral Coefficients). Il compito del riconoscitore è di realizzare un *mapping* fra le sequenze di *features* e le sequenze di simboli desiderate (*words*), attraverso un'opportuna procedura di decodifica, che può essere realizzata seguendo approcci diversi. Come analizzato in [1], l'attuale generazione di riconoscitori vocali si basa su metodologie statistiche, come quello basato su HMM (*Hidden Markov Model*).
Come accennato sopra, 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 tali non-idealità: 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 ristabilire una qualità accettabile facilitando il compito del processing che segue.
Molte proposte sono apparse in letteratura allo scopo ed una classificazione interessante nelle categorie principali di algoritmi *model-based* e *feature-enhancement* è stata proposta in [2]. Nella prima lo scopo consiste nell'adattare i parametri del riconoscitore al fine di massimizzare il grado di adeguatezza del sistema all'ambiente acustico con non-idealità. L'adattamento dei modelli avviene durante la fase di riconoscimento, partendo da un ASR allenato su un database audio pulito. Metodologie interessanti sono state proposte sia nel caso singolo canale [2], basate su un approccio statistico di compensazione e sull'approssimazione di tipo Vector Taylor Series per la stima dei coefficienti, che nel caso multicanale [3], dove i parametri del beamformer sono calcolati sulla base dell'informazione proveniente dal riconoscitore. Questo tipo di approccio presenta un elevato grado di ottimizzazione ma spesso le procedure proposte, oltre a richiedere accesso al motore di riconoscimento, sono pesanti dal punto di vista computazionale e caratterizzate da una parametrizzazione critica.
Per questo in molti casi è preferibile seguire la strategia tipica della categoria degli algoritmi di *feature-enhancement,* dove l'obiettivo è di migliorare la qualità delle *speech features* usate dall'ASR in modo da renderle il più vicino possibile alle condizioni ideali di segnale vocale pulito. Molte tecniche diverse sono state adottate allo scopo: tra di esse si possono citare le

regole di soppressione del rumore basate sia nel dominio della frequenza [4,5] che in quello cepstrale [6]. Da citare anche interessanti estensioni al caso multicanale dove viene sfruttata la capacità di filtraggio spaziale del beamforming [7,8].
Un'altra importante classe di algoritmi di *feature-enhancement* è rappresentata dai metodi di normalizzazione statistica (anche chiamato *feature-warping*), dove le *features* vegono normalizzate per mezzo di opportune trasformazioni con l'obiettivo di rendere la statistica dell'osservazione rumorosa il più vicino possibile a quella del segnale vocale pulito. Il primo tentativo a riguardo è stato fatto con l'algoritmo di Cepstral Mean and Variance Nomalization (CMVN) [1,9], che realizza una trasformazione lineare atta a modificare i primi due momenti della statistica del segnale rumoroso. Poiché la presenza di rumore induce una distorsione non lineare nella rappresentazione del segnale nel dominio delle *features*, altri metodi sono stati proposti orientati a normalizzare momenti di ordine superiore. Di particolare interesse sono i metodi basati sul concetto di Histogram Equalization (HEQ) [1,9,10], che consiste nell'applicare alla CDF (Cumulative Density Function) del segnale rumoroso una CDF di riferimento, tipicamente Gaussiana, come trasformazione non lineare. Il metodo è efficace, come i risultati di riconoscimento possono confermare, anche se non mancano i punti deboli. Uno di questi sta nel fatto che, volendo mantenere i vincoli d'implementabilità in tempo-reale del task di riconoscimento, non si ha una quantità sufficiente d'informazione da utilizzare nella stima della statistica del *noisy-speech*. L'interesse degli autori di questa memoria è di proporre tecniche innovative in grado di alleviare questo problema, permettendo di aumentare le prestazioni di riconoscimento in presenza di non-idealità acustiche, mantenendo la capacità di processamento in tempo-reale. Un possibile approccio, già parzialmente investigato, consiste nello sfruttare la disponibilità d'informazione acustica multicanale per incrementare la quantità di dati utili per il modeling statistico e quindi migliorare le prestazioni dell'approccio HEQ. Le prove sperimentali hanno fornito già risultati più che incoraggianti, ma molte sono ancora le strade da percorrere, sia in termini di ottimizzazione degli algoritmi implementati che di estendibilità a metodi alternativi all'HEQ.

Bibliografia

[1] A. Peinado and J. Segura, *Speech Recognition Over Digital Channels: Robustness and Standards*. West Sussex, U.K.: Wiley, 2006.
[2] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, "A unified framework of hmm adaptation with joint compensation of additive and convolutive distortions," *Computer Speech & Language*, vol. 23, no. 3, pp. 389 – 405, 2009.
[3] M.L. Seltzer, *Microphone Array processing for robust speech recognition*, Ph.D. Thesis, Carnegie Mellon University, Pittsburg, PA, July 2003.
[4] R. Gemello, F. Mana, and R. De Mori, "Automatic speech recognition with a modified Ephraim-Malah rule," *IEEE Sign. Proc. Letters,* vol. 13, no. 1, pp. 56–59, Jan. 2006.
[5] E. Principi, S. Cifani, R. Rotili, S. Squartini, and F. Piazza, "Comparative evaluation of single-channel mmse-based noise reduction schemes for speech recognition," *Journal of Electrical and Computer Engineering,* Hindawi Publishing Corporation, Feb. 2010.
[6] Dong Yu; Li Deng; Droppo, J.; Jian Wu; Yifan Gong; Acero, A., "Robust Speech Recognition Using a Cepstral Minimum-Mean-Square-Error-Motivated Noise Suppressor," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol.16, no.5, pp.1061-1070, July 2008.
[7] Joerg Bitzer, Klaus Uwe Simmer and Karl-Dirk Kammeye, Multi-microphone noise reduction techniques as front-end devices for speech recognition, *Speech Communication*, Volume 34, Issues 1-2, April 2001, pp. 3-12.
[8] E. Principi, S. Cifani, R. Rotili, L. Marinelli, S. Squartini, and F. Piazza, "Robust speech recognition using feature-domain multi-channel bayesian estimators," in *Proc. of ISCAS 2010*, May 2010.
[9] J. Segura, C. Benitez, A. de la Torre, A. Rubio and J. Ramirez, "Cepstral Domain Segmental Nonlinear Feature Transformations for Robust Speech Recognition," *IEEE Sign. Proc. Letters*, vol. 11, no. 5, May 2004.
[10] L. Garcia, J. Segura, A. de la Torre, C. Benitez, and A. Rubio, "Histogram equalization for robust speech recognition," *Speech Recognition*, France Mihelic and Janez Zibert (Ed.), 2008.