RICONOSCIMENTO DI EMOZIONI NEL PARLATO PER MEZZO DI PARAMETRI PROSODICI

Roberto Gretter, Dino Seppi
ITC-irst, Trento, Italia
gretter@itc.it, *seppi@itc.it*

SOMMARIO

Il presente lavoro si propone di descrivere la realizzazione di un sistema automatico per il riconoscimento di emozioni nel parlato. La peculiarità del progetto consiste nell'aver utilizzato *database* di parlato spontaneo, in particolare registrazioni di interazioni vocali uomo-macchina. L'intero progetto può essere suddiviso in tre differenti fasi: la raccolta e l'etichettatura dei dati, l'estrazione di parametri prosodici dalle registrazioni audio e la classificazione emozionale di ciascuna frase.

Sono stati considerati due *database*. Il primo, in italiano, consiste in registrazioni di utenti di *call-center* automatici. Questi dati sono stati selezionati ed etichettati da annotatori professionisti in base alle emozioni espresse da ciascun parlatore. Il secondo, in tedesco, consiste invece in registrazioni raccolte durante la messa punto di un sistema di dialogo automatico. Questa seconda raccolta contiene un numero più significativo di frasi non emotivamente neutre ed è stata utilizzata per un raffronto con i dati italiani. La scelta di fare uso di parlato spontaneo comporta inevitabilmente alcuni problemi: la presenza preponderante di frasi emotivamente neutre che rende i *database* molto sbilanciati, la disomogenea distribuzione delle emozioni rilevate e i bassi livelli di consenso registrati tra gli annotatori.

Ulteriori problematiche derivano dall'estrazione automatica di parametri significativi: a tutt'oggi, infatti, le tecniche utilizzate in letteratura non consentono di fare affidamento su un insieme limitato di *features* robuste. L'approccio adottato prevede quindi il calcolo di un gran numero di parametri, anche molto correlati tra loro, per il successivo addestramento di un classificatore automatico. Tali parametri, comunemente utilizzati in letteratura, derivano da funzioni del segnale audio come energia, frequenza fondamentale, durata ed eventuale presenza di pause tra parole e sono stati calcolati a livello di parola. La segmentazione del segnale vocale è stata ottenuta per mezzo di un segmentatore automatico.

Date le potenziali ambiguità delle annotazioni e le lacune nelle informazioni codificate dai parametri acustici il compito svolto dalla classificazione diviene difficile se non critico. Per questo motivo abbiamo testato due diversi tipi di classificatori: le reti neurali e gli alberi binari di classificazione; per entrambi forniamo i risultati nelle configurazioni che si sono rivelate più robuste. Nonostante entrambi i metodi proposti non si comportino particolarmente bene se applicati a dati sparsi e molto sbilanciati, siamo riusciti a ottenere, per entrambi i *database* utilizzati, risultati equiparabili e prestazioni più che soddisfacenti.

In conclusione restano ancora irrisolti numerosi problemi di carattere tecnico e concettuale, tra cui la significatività dei parametri prosodico-acustici utilizzati e l'affidabilità dell'etichettatura manuale dei dati. Quindi sforzi futuri per il miglioramento dei risultati andranno direzionati soprattutto verso un'attenta e più approfondita analisi dei parametri, che andranno modificati, selezionati e affiancati da altri, non necessariamente solo di natura prosodica.