USO DELLA PROSODIA PER LA DISAMBIGUAZIONE DI SEQUENZE DI NUMERI

Roberto Gretter, Silvia Rocchi e Dino Seppi

ITC-irst – Trento

RIASSUNTO

L'informazione prosodica viene esplicitamente modellata negli attuali riconoscitori vocali solo in alcuni casi, per lo più legati all'individuazione di confini di frase [1]. È fuori dubbio, tuttavia, come essa sia importante per la comprensione del messaggio vocale. Per affrontare questo argomento, in ITC-irst sono stati acquisiti ed etichettati alcuni insiemi di frasi, strutturati in modo da evidenziare differenti usi della prosodia nel linguaggio.

Una parte di tale database contiene sequenze di numeri, che per un riconoscitore vocale incapace di gestire fenomeni prosodici, risultano fortemente ambigue. Ad esempio, la serie di *elementi* "87 134" verrebbe riconosciuta come "ottanta sette cento trenta quattro" la quale, a sua volta, potrebbe essere ricomposta in diversi modi: "87 134", "80 734", "80 7 100 30 4", "80 700 34" e così via. Ambiguità di questo tipo sono abbastanza frequenti nel riconoscimento automatico e si verificano nelle applicazioni che prevedono l'elaborazione di numeri di telefono o di carte di credito.

Per superare problemi di questa natura si è quindi utilizzata la prosodia [2]. Purtroppo quest'ultima non è facilmente codificabile, e può essere più o meno marcata all'interno del segnale vocale, a totale discrezione del parlatore. Evidenti strutture prosodiche sono le pause tra numero e numero, le variazioni d'intensità del suono, la differente velocità con cui vengono pronunciati gli elementi di ogni numero. La frequenza fondamentale è il parametro più difficile da trattare, sia per la variabilità tra i parlatori sia perché la sua estrazione automatica non è sempre affidabile.

Utilizzando queste misure si è cercato di disambiguare in maniera automatica sequenze di numeri fino a tre cifre. I segnali vocali registrati, pronunciati da parlatori che leggevano frasi come "123 130 575 500 70 5", sono stati etichettati sia come numeri sia come elementi che li compongono. In seguito è stata effettuata la classificazione automatica dei diversi elementi, eseguita tramite un albero binario di ricerca precedentemente addestrato usando parte dei dati disponibili.

I risultati ottenuti sono apprezzabili: con l'impiego di tutta l'informazione prosodica, codificata in otto diversi parametri, si è stati in grado di classificare correttamente circa il 98% degli elementi presenti nel database. Ottimi risultati, del tutto paragonabili ai precedenti, sono stati raggiunti anche distinguendo tra elemento finale o non finale all'interno di una sequenza: questa classificazione potrebbe essere utilizzata per potenziare un normale *start-end point detector,* al fine di migliorare la robustezza di un sistema di riconoscimento automatico della voce.

I parametri rivelatisi più efficaci sono quelli derivati dalla lunghezza delle pause tra elementi e dalla durata degli elementi stessi. Come previsto, l'interpretazione dell'informazione ottenuta dalla frequenza fondamentale del segnale si è rivelata essere la più critica. Per migliorare l'uso di quest'ultima si è provveduto a clusterizzarne gli andamenti prima della classificazione.

In un'ultima fase del lavoro tutto il sistema è stato ottimizzato in modo da poterne permettere l'integrazione in un ambiente *real-time*.

Riferimenti

[1] F. Gallwitz, H. Niemann, E. Noeth, V. Warnke, "Integrated recognition of words and prosodic phrase boundaries", *Speech Communication*, Vol. 36, pag. 81-95, 2002.

[2] Jiahong Yuan, Chilin Shih, Greg Kochanski, "Comparison of Declarative and Interrogative Intonation in Chinese", *Proceedings of Speech Prosody Conference*, Aix-en-Provence, France, 2002.

SESSIONE: Prosodia, Riconoscimento automatico del parlato