DOI: 10.17469/O2103AISV000023

SONIA CENCESCHI, ROBERTO TEDESCO, LICIA SBATTELLA

Verso il riconoscimento automatico della prosodia

This paper presents our approach to automatic recognition of prosodic forms. In particular, we present: CALLIOPE, a multi-dimensional model aiming at categorizing all prosodic forms; SI-CALLIOPE, a sub-space for which we defined a corpus of recorded prosodic forms; and the psychoacoustic experiment we are currently planning for investigating main acoustic behaviours and features involved into the discrimination of prosodic forms. The results of the experiment will be useful for defining the acoustic/textual features to rely on for automatic recognition of prosodic forms. For that reason, we are also defining a classifier, based on Neural Nets. This study is part of the LYV project, which focuses on improving prosodic expressiveness skills of Italian speakers with autism and other cognitive disabilities.

Key words: prosody, human-computer interaction, paralinguistics, Neural Networks.

1. *Introduzione*

In questo articolo presentiamo una proposta di approccio al riconoscimento automatico di forme prosodiche. In particolare, verranno descritti:

- CALLIOPE: un modello multidimensionale che ha lo scopo di catalogare tutte le possibili forme prosodiche pronunciabili da un parlante generico.
- SI-CALLIOPE: un sottospazio di CALLIOPE, per il quale definiamo un corpus di registrazioni audio.
- Gli esperimenti psicoacustici necessari per indagarne i principali comportamenti acustici.
- Il modello di un classificatore, basato su Reti Neurali, per riconoscere automaticamente alcune forme prosodiche.

Questo studio è parte del progetto LYV[1] (Lend Your Voice), incentrato sul miglioramento delle capacità prosodiche ed espressive di parlanti italiani con disabilità cognitive, tramite l'utilizzo della tecnologia e in contesti complessi (Sbattella, 2007).

2. *Il modello CALLIOPE*

CALLIOPE è uno spazio multidimensionale, dove ogni dimensione rappresenta un fattore che influenza l'interpretazione della singola Unità Informativa (Cresti, 2000). Ogni dimensione i è rappresentata da una variabile qualitativa l_i, che assume

[1] LYV è un progetto Polisocial award 2016-2017, http://www.polisocial.polimi.it.

valore all'interno di un insieme F_i; ogni UI è quindi associata a un punto dello spazio. Più formalmente, una generica UI è associata ad una tupla T(UI) composta di dodici etichette:

$$T(UI) = (l_1, l_2, ..., l_{12}) : l_i \in F_i \, , \, 1 \leq i \leq 12$$

dove l_i è la i-esima etichetta scelta nel corrispondente insieme F_i. Diverse UI possono essere associate alla stessa T(UI), ma ogni UI è univocamente associata ad una e una sola T(UI).

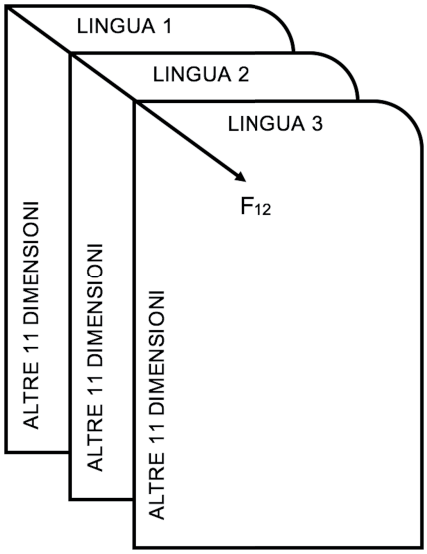
Le dimensioni di CALLIOPE sono divise *Dialogic Dimensions* (caratteristiche correlate al contesto comunicativo) e *Background Dimensions* (caratteristiche che esistono a prescindere dalla presenza di interazione tra individui).

Le Dialogic Dimensions, da F_1 a F_9, sono: *Struttura*, *Modalità Linguistica*, *Focus Intonativo*, *Forma Retorica*, *Stato Motivazionale*, *Speech Mood*, *Spontaneità, Forme di Punteggiatura,* ed *Emozioni*. La *Struttura* comprende le costruzioni verbali (Prieto, Borràs-Comes & Roseano, 2010-2014) composte da singoli elementi grammaticali e linguistici (un esempio è la forma della domanda diretta). La *Modalità* rappresenta l'intenzione comunicativa scelta dal parlante per ottenere un particolare effetto sull'interlocutore (Kratzer, 2012). Il *Focus Intonativo* cataloga le diverse motivazioni per cui una parola può essere enfatizzata rispetto alle altre all'interno del dialogo (Ouyang, Kaiser, 2012), mentre la *Forma Retorica* comprende solo le figure retoriche che influenzano la componente sonora della UI. Lo *Stato Motivazionale* (Liotti, Monticelli, 2008) indica come il suono delle UI è in relazione con il ruolo del parlante nel contesto del dialogo: la stessa frase pronunciata da un giudice o da una madre può "suonare" evidentemente differente a causa del diverso rapporto con l'interlocutore. Il *Mood* si riferisce all'intensità applicabile alle UI in un dato contesto: la prosodia di una stessa UI cambia integralmente se sussurrata o urlata. La *Spontaneità* è relativa al "livello di improvvisazione" che il parlante può, o deve, adottare in una data situazione (Nencioni, 1983): parlato letto, recitato, spontaneo assumono prosodie differenti per una stessa UI. Le *Forme di Punteggiatura* sono qui da considerarsi come un indice della presenza di pause all'interno di una UI e sono indicate, in particolare, dalla presenza di virgole. Talvolta la virgola è utilizzata come separatore di UIs, ma in questo caso consideriamo solamente un suo utilizzo interno alla singola unità intonativa. Le *Emozioni*, infine, sono un'evidente causa di modifica della prosodia di una UI (Cowie, Cornelius, 2003; Tomkins, 1984). CALLIOPE propone, per ognuno degli insiemi da F_1 a F_9, un numero finito di etichette.

Le Background Dimensions sono F_{10}, F_{11} e F_{12}: *Capacità Espressive Soggettive*, *Contesto Sociale* e *Linguaggio, Dialetto o Variazione Linguistica Locale*. La *Capacità Espressiva* può influenzare una UI modificando il modello necessario per una corretta trasmissione del messaggio. *Contesto Sociale*, e soprattutto *Linguaggio, Dialetto o Variazione Linguistica Locale*, modificano in modo non banale la prosodia dei parlanti. CALLIOPE propone, per ognuno degli insiemi F_{10}, F_{11} e F_{12}, alcune etichette che però non intendono essere esaustive.

Ciascuna etichetta in F_{12} è da immaginarsi come un *sottospazio*[2]: un livello all'interno del quale attribuire le restanti dimensioni (Figura 1). Uno di tali sottospazi è l'Italiano Standard, che sarà utilizzato nel seguito.

Figura 1 - *Diversi linguaggi*



3. *SI-CALLIOPE e validazione*

La validazione dell'intero modello CALLIOPE è molto complessa. Abbiamo quindi deciso di focalizzarci su un sotto-spazio, denominato SI-CALLIOPE (*Standard Italian* CALLIOPE). Tale modello è utile per gli scopi del progetto LYV, che l'utilizzerà l'Italiano Standard (F_{12}) come descritto da Canepari (1986), prendendo come modello la voce di parlanti normodotati (F_{10}) simulando situazioni prosodiche tipiche della quotidianità (F_{11}). L'analisi è stata poi ulteriormente ristretta a frasi non retoriche e con sospensione (F_4) e parlato recitato (F_7).

3.1 Il Corpus Prosodico: un sottospazio di SI-CALLIOPE

Grazie ai volontari di Libro Parlato Onlus[3] e ad alcuni attori professionisti, abbiamo registrato un corpus contenente una lista di frasi e pseudo-frasi. Quando l'interpretazione non era chiara, abbiamo fornito ai parlanti dei suggerimenti o una descrizione del contesto. Le UI sono così suddivise:

– F_1, Struttura:

[2] A rigore, ogni dimensione di CALLIOPE definisce un insieme di sottospazi. Tuttavia, poiché si immagina che tipicamente il modello sarà applicato a un linguaggio alla volta, visualizzare lo spazio di CALLIOPE come indicato in Figura 2 permette di comprenderne meglio la struttura senza perdita di generalità.

[3] Libro Parlato Onlus, Centro Internazionale del Libro Parlato (CILP), Feltre, Italy, www.libroparlato.org.

1. Dichiarativa
2. Interrogativa ad 1 unità tonale
3. Interrogativa a 2 o più unità tonali
4. Interrogativa disgiuntiva
5. Domanda eco
6. Esclamativa
7. Vocativo

– F_6, Speech mood:
8. Sospirato
9. Urlato
10. Standard

– F_3, Focus intonativo
11. Focus contrastivo

– F_4, Forma retorica
12. Sospensione

– F_8, Forma di punteggiatura
13. Lista

I parlanti sono 14, 7 uomini e 7 donne, con età compresa tra 33 e 48 anni. Ogni parlante ha registrato circa 1 ora di parlato, 278 UI (139 con significato, 139 pseudo-frasi) per un totale di 1946 frasi con significato e 1946 pseudo-frasi. I file audio sono stati registrati, con differenti modalità e microfoni, in formato WAV (44.1 kHz), per ottenere un modello quanto più possibile indipendente dal mezzo tecnico utilizzato.

3.2 Generazione delle pseudo-frasi

La fase sperimentale richiede la generazione e la registrazione di pseudo-frasi: frasi composte da parole senza significato, ma che rispettano la fonotassi della lingua Italiana e che quindi "suonano" come italiane. Siamo partiti dal corpus di parole italiane CoLFIS[4], dal quale sono state rimosse le parole contenenti caratteri nell'insieme {'w', 'y', 'j', 'k', 'x'}, e le parole contenenti segni diacritici diversi dall'accento grave o acuto. Le parole così rimaste sono state divise in sillabe tramite Hyphenator 0.5.1, un modulo Python che sfrutta la sillabazione del dizionario fornito da OpenOffice. Quindi, è stato addestrato un trigramma di sillabe che codifica un'approssimazione della fonotassi italiana. Il trigramma utilizzato dal nostro generatore è definito come la seguente distribuzione di probabilità condizionata:

$$P(s_i \mid s_{i-1}, s_{i-2}) : s_i \in S$$

dove s_i è la sillaba i-esima della parola da trasformare in pseudo-parola, e S è l'insieme delle sillabe ricavate dall'analisi del corpus CoLFIS.

[4] Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS), http://linguistica.sns.it/CoLFIS/Home.htm.

L'algoritmo parte da una frase italiana e genera, per ogni parola più lunga di 3 caratteri, una pseudo-frase con lo stesso numero di sillabe. In particolare, partendo da una parola composta da n sillabe, l'algoritmo sceglie n sillabe dall'insieme S, tramite n campionamenti pseudo-casuali della distribuzione di probabilità del trigramma[5], e le accosta consecutivamente, costruendo la corrispondente pseudo-parola. Per migliorare la leggibilità delle frasi ottenute, le parole più corte di 4 caratteri (per lo più articoli, preposizioni e alcune forme del verbo "essere") non sono modificate e vengono semplicemente copiate nella pseudo-frase. L'algoritmo applica quindi alle pseudo-parole un insieme di regole di concordanza che migliorano la leggibilità globale, poi ulteriormente raffinata tramite un ultimo controllo manuale.

4. *Esperimento psicoacustico*

Il test per l'esperimento psicoacustico è disponibile su web, all'indirizzo http://calliope.deib.polimi.it. Lo scopo principale dell'esperimento è comprendere quali informazioni utilizza il nostro cervello per la comprensione della componente prosodica:

- Quali forme prosodiche sono riconosciute grazie alla mera componente acustica del parlato?
- Quali forme prosodiche sono riconosciute grazie alla componente acustica e alla fonotassi?
- Quali forme prosodiche necessitano anche del significato della frase?

Il test è composto da 36 quesiti, suddivisi in tre sezioni, per ognuno dei quali un audio è estratto, con tecniche di randomizzazione, dal corpus. L'ascoltatore deve ascoltare e riconoscere la forma prosodica. Vi sono quindi:

- 13 frasi con significato
- 13 pseudo-frasi
- 10 frasi ridotte al solo inviluppo del pitch[6]

La terza sezione manca dei tre quesiti riguardanti la percezione dell'intensità, perché essa è impossibile da rilevare utilizzando il solo pitch. Ogni quesito comprende tre audio e l'ascoltatore deve segnalare, per ognuno di essi, se riconosce o no la forma prosodica suggerita dal quesito stesso (per esempio "Quali tra gli audio seguenti è una domanda diretta?"). Per ogni quesito, il sistema inserisce m_{yes} audio che l'ascoltatore dovrebbe riconoscere e $m_{no} = 3 - m_{yes}$ audio che l'ascoltatore dovrebbe scartare, dove $m_{yes} \in \{1,2,3\}$ è scelto a caso.

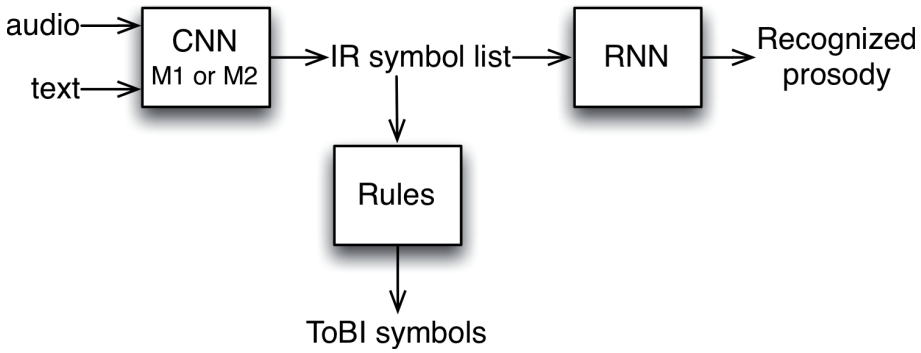
[5] Ciò significa che la probabilità di scegliere una cerca sillaba, in posizione i-esima, dipende dalle sillabe in precedenza scelte per le posizioni $i-1$ e $i-2$. È questa dipendenza che permette di approssimare la fonotassi.

[6] Per la generazione di questa versione, abbiamo utilizzato la funzione Hum del programma Praat, che genera un suono simile a una vocale (suona come una sorta di "a"), secondo l'andamento voluto del pitch.

5. *Il classificatore basato su Reti Neurali*

L'architettura generale del riconoscitore è mostrata in Figura 2.

Figura 2 - *Schema di massima del riconoscitore*



Una Convolutionary Neural Network (CNN) sarà utilizzata per trovare pattern prosodici, che saranno trasformati in una sequenza di simboli. Questa rappresentazione intermedia (Intermediate Representation – IR) è utile perché più astratta e informativa rispetto ai dati audio reali in input. Stiamo al momento definendo questi simboli, che potranno essere un'integrazione del sistema di taggatura ToBI (Beckman, Hirschberg & Shattuck-Hufnagel, 2005), ma in una versione più raffinata che possa descrivere in modo più approfondito il modello prosodico.

Una Recurrent Neural network (RNN) riconoscerà la sequenza di simboli classificandoli come una delle forme prosodiche scelte. Dalla sequenza di simboli IR, un modello a regole genererà le corrispondenti etichette ToBI. La CNN è in realtà composta da due modelli separati: il più semplice, M1, considera solo feature audio e ha lo scopo di riconoscere forme prosodiche dove le caratteristiche sonore sono condizione sufficiente per il riconoscimento. Il secondo, M2, necessita anche di feature testuali e ha lo scopo di riconoscere un più largo insieme di forme prosodiche. M1 è indicato per situazioni in cui il testo non è disponibile o l'allineamento testo-audio non è possibile. In entrambi i casi non possono essere utilizzate feature testuali.

Sia la CNN che la RNN saranno addestrate utilizzando il corpus audio derivato da SI-CALLIOPE. Il modello che genererà i simboli ToBI a partire da quelli IR sarà composto da regole definite manualmente e non necessiterà di una fase di addestramento.

6. *Conclusioni*

Questo lavoro è una proposta che potrà essere utilizzata per il riconoscimento automatico della prosodia vocale in diversi campi. Nonostante si sia deciso di partire da una serie di frasi utili nell'ambito del progetto LYV, nulla vieta che la ricerca

possa essere ampliata, seguendo la stessa metodologia, ad ulteriori forme prosodiche e applicata ad ambiti di ricerca. Potrebbe risultare utile, per esempio, come modello di ricerche inerenti altre lingue, o come guida per pianificare ulteriori studi sperimentali. Il corpus può essere ampliato con facilità, aggiungendo nuovi esempi di prosodie e creando le relative pseudo-word con il software che abbiamo scritto.

Riferimenti bibliografici

Beckman, M.E., Hirschberg, J. & Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. In Jun, S.A. (Ed.), *Prosodic typology – The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press, 9-54.

Canepari, L. (1986). *Italiano standard e pronunce regionali*. Padova: CLEUP.

Cowie, R., Cornelius, R.R. (2003). Describing the emotional states that are expressed in speech. In *Speech Communication*, 40, 5-32.

Cresti, E. (2000). *Corpus di italiano parlato*. Firenze: Accademia della Crusca.

Kratzer, A. (2012). *Modals and Conditionals: new and revised perspectives Oxford Studies in Theoretical Linguistics*. Oxford: Oxford University Press.

Liotti, G., Monticelli, F. (2008). *I sistemi motivazionali nel dialogo clinico*. Milano: Raffaello Cortina editore.

Nencioni, G. (1983). *Di scritto e di parlato*. Bologna: Zanichelli.

Ouyang, I.C., Kaiser, E. (2012). Focus-marking in a tone language: prosodic cues in Mandarin Chinese. In *Proceedings of the Linguistic Society of America, Extended abstracts of the annual meeting*, 3, 1-5.

Prieto, P., Borràs-Comes, J. & Roseano, P. (2010). *Interactive atlas of Romance intonation*. http://prosodia.upf.edu/iari/.

Sbattella, L. (2007). Le défi de la complexité: harmoniser et composer. In *La formatività del musicale*. Milano: Esagramma.

Tomkins, S.S. (1984). *Affect theory*. In Scherer, K.R., Ekman, P. (Eds.), *Approaches to emotion*. Hillsdale: Lawrence Erlbaum Associates.