Prosodic prominence detection in Italian continuous speech using probabilistic graphical models

Fabio Tamburini[1], *Chiara Bertini*[2], *Pier Marco Bertinetto*[2]

[1] FICLIT, University of Bologna, Italy
[2] Scuola Normale Superiore, Pisa, Italy

fabio.tamburini@unibo.it, c.bertini@sns.it, p.bertinetto@sns.it

Abstract

Prosodic prominence, a speech phenomenon by which some linguistic units are perceived as standing out from their environment, plays a very important role in human communication. In this paper we present a study on automatic prominence identification using Probabilistic Graphical Models, a family of Machine Learning Systems able to properly handle sequences of events. We tested the most promising members of such models on utterances selected from a manually annotated Italian speech corpus, obtaining very good recognition results crucially converging with the prominence detection responses provided by a pool of native speakers.

Index Terms: prosody, prominence, probabilistic graphical models, prominence annotation.

1. Introduction

A fairly uncontroversial definition of *prosodic prominence* due to Terken [29: 1768] states: "*prominence is the property by which linguistic units are perceived as standing out from their environment*". These prominent units typically contain relevant information for discourse and their correct perception is crucial for a successful communication strategy. Speakers use prominence to draw the listener's attention on specific point of the utterance, to express their emotion or attitude about the topic being discussed, to indicate the focus of an utterance, to mark the introduction of new topics, to indicate the information status of a word (new or given), to change speaking style, etc.

For all these reasons, the automatic management of prosodic prominence is crucial for both recognition and synthesis in order to build systems able to properly handle information in speech.

There is a long-standing agreement among scholars to consider the syllable as the prominence-bearing unit in connected speech. This position is not uncontroversial, however, and various studies analyse prominence at word level, especially if they mainly concern information extraction from speech utterances. In this paper we will consider the syllable, and its constituent units, as the relevant domain for prominence computation.

Several recent contributions handle prosodic prominence from a computational point of view, proposing different models (both Rule-Based and Machine Learning Systems - **MLS**) for the automatic detection of prominence in various languages, e.g. [2, 5, 11, 12, 15, 24, 25, 26]. Some of them are specifically devoted to Italian, or handle the identification of prosodic prominence in Italian among other languages [1, 8, 17, 26].

In this paper we present a procedure for the identification of prosodic prominence in Italian in the framework of MLS, based on training procedures that extract data and models from annotated corpora. These systems only take acoustic features into consideration, such as nucleus / syllable duration, energy measures in the nuclei / syllables and analysis of specific pitch profiles in the utterance.

Adopting the above reported definition [29], we consider prominence as a phenomenon establishing precise syntagmatic relations with respect to the neighbouring syllables. Its identification requires MLS able to properly model sequences of events, because the immediate context information, both in the feature sequence of the input and in the label sequence of the output, are crucial for the correct identification of syllable prominence. A syllable can be defined as prominent only by considering the relationships with other syllables, in line with the classical figure – ground contrast proposed by Gestalt psychology, rather than by considering it – and its features – in isolation.

Probabilistic graphical models (**PGM**) represent a class of MLS that, taking advantage of discriminative stochastic models, can successfully handle recognition problems that heavily depend on sequences. PGM, in some of their various configurations / models, have been applied to the task at stake with encouraging results [8, 21]. Despite these findings, PGM and their complex family of model variations have not yet been extensively applied to this problem, especially considering hidden or latent dynamics detection in speech data and the possibility of extracting and using high order relations among the acoustic features.

2. Probabilistic graphical models

PGM are powerful frameworks for representation and inference in multivariate probability distribution. They use a graph-based representation as the basis for compactly encoding a complex distribution over a high-dimensional space representing the conditional dependence structure between random variables.

In this paper we considered some of the most powerful and widespread discriminative models to identify prosodic prominence in continuous speech, as well as some recently presented new models.

PGM consists of a large family of different methods that constrain the graph structure in specific ways. Conditional Random Fields (**CRF** - see [10, 22] for general introductions) are no doubt the most used PGM in various fields. However, most CRF models use linear functions to represent the relationships between input features and the classification output and a simple graph structure for the entire model. This way of coding relations presents severe limitations for real-world applications, because: (a) in many cases the relationships between inputs and outputs are complex and nonlinear, and (b) some problems require modelling relevant sub-structures in the label sequence.

In this work we used different PGM, each addressing in its own way the shortcomings of CRF models, considered as a baseline. Conditional Neural Fields (**CNF**) [20], inserting a small neural network between input and output, are able to capture the nonlinearities required by constraint (a) above; Latent-Dynamic Conditional Random Fields (**LDCRF**) [18], in turn, can learn latent sub-structures in output class labels. Latent-Dynamic Conditional Neural Fields (**LDCNF**) [16] can combine the advantages of both previous approaches in a single model.

PGM are able to manage sequences of input-output data predicting the output sequence considering both the input feature configuration, in a specific window centered on the generic input vector of features x_j, and the previous output sequence. Figure 1 outlines the different structures of the PGM used in this work.

Given the input sequence of local features $x_1,...,x_n$, typically consisting of vectors of features, and given the output sequence $y_1,...,y_n$, linear CRF assign the most likely label to output y_j conditioned by the feature vectors belonging to the local window and the previous output label y_{j-1}.

CNF extends CRF by adding one level of gate units, acting as a neuron tier (more precisely a perceptron), between the input and the output layers. These gate neurons are a sort of feature extractor able to capture nonlinear relationships between input and output.

A further, completely different way of extending CRF is implemented in LDCRF adding a layer of hidden-state units between input and output layers: these units are able to model the sub-structure of the label sequence and can learn complex dynamic behaviours between output labels.

Finally, LDCNF take advantage from both approaches, combining them in a unique structure. The LDCRF model is modified by adding the neural network introduced in CNF between input and hidden units; in this way, LDCNF models can both identify sub-structures in the output sequence and learn nonlinear relationships between input feature vectors and output class labels.

For lack of space, we refer to the cited papers for all mathematical and algorithmic details that define these approaches, especially for what concerns the learning algorithms, inference and parameter setting.

3. Corpus building and data collection

The materials used in this study were utterances extracted from the API/AVIP corpus [3] and from a selection of sentences read by a subset of the same speakers. The corpus consisted of semi-spontaneous conversations between native Italian speakers elicited with the map-task method for different Italian language varieties. The speakers used for the present purpose were from the Pisa area (central Italy).

The selected utterances presented a neutral intonation contour, without emphatic stress or pauses, and presenting at least 8 syllables. Care was taken to avoid any disturbing phenomena such as speakers' overlap, laughters, background noise, etc.

A perception experiment, divided into two different sessions (sets A-C vs D-G in table 1), was carried out using 120 selected corpus utterances (90 spontaneous and 30 read), produced by female and male speakers. The average utterance length was 18 syllables, ranging from 9 to 35. The task was performed by 35 Italian native speakers.

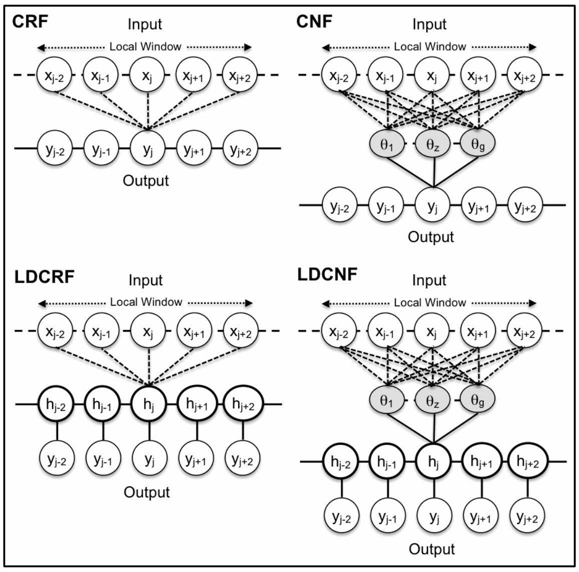
CRF
Input
Local Window
x_{j-2} x_{j-1} x_j x_{j+1} x_{j+2}
y_{j-2} y_{j-1} y_j y_{j+1} y_{j+2}
Output
CNF
Input
Local Window
x_{j-2} x_{j-1} x_j x_{j+1} x_{j+2}
θ_1 θ_z θ_g
y_{j-2} y_{j-1} y_j y_{j+1} y_{j+2}
Output
LDCRF
Input
Local Window
x_{j-2} x_{j-1} x_j x_{j+1} x_{j+2}
h_{j-2} h_{j-1} h_j h_{j+1} h_{j+2}
y_{j-2} y_{j-1} y_j y_{j+1} y_{j+2}
Output
LDCNF
Input
Local Window
x_{j-2} x_{j-1} x_j x_{j+1} x_{j+2}
θ_1 θ_z θ_g
h_{j-2} h_{j-1} h_j h_{j+1} h_{j+2}
y_{j-2} y_{j-1} y_j y_{j+1} y_{j+2}
Output

Figure 1. *The various PGM considered in this study. The gate units $\theta_1...\theta_g$ are in gray while the hidden units $h_1...h_n$ present a thick border. For clarity, only the region of the model net surrounding the generic input feature vector x_j is represented in the pictures.*

Table 1*: The latin square scheme applied to the 120 utterances composing the corpus.*

Utterance Sub-list	Annotators IDs
A (20 utt.)	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
B (20 utt.)	1, 2, 3, 4, 5, 11, 12, 13, 14, 15
C (20 utt.)	6, 7, 8, 9, 10, 11, 12, 13, 14, 15
D (15 utt.)	16, 17, 18, 19, 20, 21, 22, 23, 24, 25
E (15 utt.)	16, 17, 18, 19, 20, 26, 27, 28, 29, 30
F (15 utt.)	31, 32, 33, 34, 35, 21, 22, 23, 24, 25
G (15 utt.)	26, 27, 28, 29, 30, 31, 32, 33, 34, 35

The experimental task was to identify the sentence prominences. The participants could listen to each sentence as many time as they wanted. To reduce the task difficulty, participants were presented with a transcription of the given sentence whereby each syllable (as the possible prominence-carrying unit) was separately indicated; in addition, the lexically stressed syllables were explicitly pointed out. However, participants were warned during the training phase that not all lexically stressed syllables were actual targets of sentence prominence, while prominence could also land on lexically unstressed syllables. As soon as the participant had made her/his own choice by clicking on the square corresponding to the intended syllables, s/he was immediately presented with another sentence.

Since this task is very demanding in terms of attention, the sentences were divided into sub-lists according to a latin square scheme, so that each utterance was judged by 10 speakers. As a consequence, no participant heard/read all the sentences. Table 1 depicts the annotation scheme: the total number of utterances was divided into seven sub-lists each assigned to 10 annotators.

For each test utterance extracted from the AVIP corpus, the phonetic transcription and the phoneme level segmentation were available in the source. The selected utterances were further segmented manually in order to identify the syllable boundaries.

3.1. First-step: data annotation and overall judgment convergence

As a first step, the participants judgments were pooled together and evaluated with respect to the degree of convergence relative to the identification of any given syllable as prominent. The convergence level was assessed with respect to four criteria (60%, 70%, 80% and 90%), indicating the percentage of shared prominence identification. In practice, considering that each sentence was judged by 10 participants, the 60% criterion implied the convergence of 6 out of 10 listeners (and similarly for the other levels).

Needless to say, the 90% agreement level involves a smaller number of prominent syllables within any given sentence, since almost all participants have to agree on their judgment concerning the given syllable. By contrast, the more generous 60% level concerns a larger number of syllables. Interestingly, there was full agreement as for the last prominence of each sentence, evidently due to unequivocal durational cues, although the energy and frequency levels at the end of an utterance are usually fairly low.

As is well known, while the identification of emphatic prominences is undisputable, there usually is considerable divergence among human judges on the identification of non-emphatic prominences. As a consequence, the 80% level was selected as bench-mark for the automatic detection of prominences as a first approximation. The dataset contains 480 prominent syllables out of 2037, thus close to one prominent syllable out of four (23.56%).

3.2. Second-step: best annotators selection

Capitalizing on the well-known lack of overwhelming convergence among human judges as for the localization of sentence prominences, a second type of comparison between the automatic detector and the human judges was adopted. For each of the two sub-lists of utterances, the three most reliable judges were selected, i.e. the three participants presenting the highest level of mutual agreement according to the Fleiss K index (this schema will be referred to as the "best-3" annotation agreement). Subsequently, the syllables judged as prominent by the majority of the "best-3" were considered prominent. The correlation data are reported in Table 2. In this dataset, 33.46% of the syllables are prominent (one out of three).

4. Acoustic features

The acoustic features used in this study are the same used in some previous studies of one of the authors [25, 26]. These works proposed a rule-based system resting on four acoustic features that exhibited good performances in prominence detection. One of the major challenges in predicting syllable prominence is the correct identification of the various sources of influence, such as: fundamental frequency excursions, duration, intensity-related parameters and listeners' linguistic expectancies.

The automatic prominence detection system described in [25, 26] is based on the global prominence model proposed by Kohler [13, 14]. In his view, there are two main 'actors', at the linguistic-prosodic level, playing a relevant role in supporting sentence prominence. The first, *pitch accent*, coincides with a concept first introduced by Bolinger [4] and concerns specific movements in F0 profile. The second, *force accent*, is completely independent from the intonational profile and is connected with different acoustic phenomena, such as intensity (or spectral emphasis), segmental durations and possibly others. Both 'actors' seem to play a relevant role in supporting prominence perception at utterance level, mutually reinforcing each other.

Table 2*: The "best-3" annotators for each utterance sub-list.*

Utterance Sub-list	Best 3 annotators	Fleiss-K
A	2, 3, 6	0.875631
B	1, 2, 13	0.876340
C	6, 9, 13	0.859735
D	16, 18, 20	0.836450
E	16, 19, 20	0.857646
F	31, 33, 25	0.863555
G	31, 33, 34	0.812559

In the present study, we considered the four features used in the cited work (reported in Table 3 with brief reference to their actual computation) and added one further acoustic feature, namely syllable duration, following the good results obtained in [8]. All these features, except syllable duration, are computed within the syllable nucleus domain. Thus, using the phonetic and syllabic segmentation provided in the source corpus, all we had to do was to define the duration of the syllabic nuclei, deriving it automatically from the other two measures.

5. Results and discussion

We made a number of experiments, considering various PGM and different parameter configurations in order to maximize the agreement between the automatic procedure and the human annotators. We tested the best system on the above-described corpus, applying a random sub-sampling validation to define the training and test set (respecting a 5/1 proportion, 100/20 utterances), repeating this procedure 20 times and averaging the obtained results.

The best performances so far obtained, in comparing the automatic classifications with the gold standard, are depicted in Table 4 and 5. The first table refers to the "80%" level of annotation agreement described in section 3.1, the second refers to the "best-3" agreement described in section 3.2. There is a clear performance improvement considering the "best-3" annotation schema: the larger inter-human agreement produces a more consistent annotation and, thus, better performances of the MLS trained on these data.

The accuracy obtained by our automatic systems is very high, considering that the typical inter-human agreement accuracy reported in the literature, when annotating prominence by means of two levels (prominent vs non-prominent) is in the range 70-90%.

However, considering that the distribution between the two prominence classes is rather skewed, one should best adopt the

F-measure as a more reliable metric of the actual system performance. An F-measure of 0.770 is quite high, considering that:

a) The training corpus used to set up the model, in the various permutations composing the 20-random-sub-sampling validation, is rather small to properly train a MLS, as it only contains 100 utterances (1800 syllable, on average);
b) We only used acoustic information and did not consider any linguistic feature that might improve the system's behaviour, as showed, for example, by [21].

Table 3. *Acoustic features used to set up the PGM models for prominence identification.*

Acoustic Feature	Description
Nucleus Duration	Duration of the syllable nucleus normalised w.r.t. mean and variance duration of the syllable nuclei in the utterance (z-score), as based on the manual segmentation available in the database.
Spectral emphasis	Normalised SPLH-SPL parameter [9] (z-score).
Pitch movements	Computed as the product of A_{event} and D_{event} parameters of the TILT model representation [28] of pitch movements. The raw pitch contour is the median of three pitch tracking algorithms [27]: RAPT [23], SWIPE' [6] and YAAPT [31]. The raw pitch profile was stylised by using a quadratic spline function, interpolating the control points derived from the OpS algorithm proposed in [19].
Overall intensity	RMS energy computed in the frequency band 50-5000 Hz, normalised to mean and variance of intensity inside the utterance (z-score).
Syllable Duration	The same as the nucleus duration but referred to the entire syllable.

Table 4. *Results obtained by the PGM tested, in terms of Accuracy, Precision & Recall and F-Measure, as referred to the 80% level of annotation agreement described in section 3.1. The parameters considered with the various PGM are:* w – *local window size (symmetric),* h – *number of hidden units,* g – *number of gate neurons and* α – *regularization factor.*

Model	Parameters	Results			
		Acc.	Prec.	Rec.	F
SVM	w=2, C=0.5	0.858	0.765	0.592	0.665
CRF	w=1	0.856	0.735	0.609	0.665
LDCRF [18]	w=1, h=2	0.856	0.720	0.640	0.676
CNF [16]	w=2, g=40 α=0.5	0.871	0.784	0.642	0.705
CNF [20]	w=1, g=20	0.872	0.769	0.667	0.713
LDCNF [16]	w=1, h=4, g=40, α=0.5	**0.875**	**0.788**	**0.658**	**0.716**

Table 5. *Results obtained by the PGM tested, as referred to the "best-3" level of annotation agreement described in section 3.2. The parameters are described in the caption of Table 4.*

Model	Parameters	Results			
		Acc.	Prec.	Rec.	F
SVM	w=1, C=50	0.833	0.791	0.681	0.732
CRF	w=2	0.838	0.795	0.695	0.741
LDCRF [18]	w=1, h=2	0.842	0.792	0.713	0.750
CNF [16]	w=1, g=20 α=0.5	0.845	0.803	0.712	0.754
LDCNF [16]	w=1, h=4, g=20, α=0.5	0.851	0.823	0.706	0.759
CNF [20]	w=1, g=20	**0.855**	**0.831**	**0.718**	**0.770**

Our results cannot be directly compared with other similar studies, because there are no standard corpora for evaluation, both in general and for Italian in particular, nor specific standardised metrics. In any case, it is worth observing that the best-obtained results are equivalent or better than those of the already cited studies (e.g. [8, 15, 17]).

The best PGM for the problem at hand seems to be CNF in the implementation proposed by [20]. LDCNF and CNF from [16] obtained slightly lower performances, especially using the "best-3" annotation schema. This is probably due to the small set of utterances used to train these models, since while performing some other tests not reported here, using more utterances and a different corpus, LDCNF performed best.

In order to compare the PGM results with standard non-sequential MLS, we included in our experiments the results obtained using classical Support Vector Machines (**SVM**). All PGM exhibit significant performance improvements when compared with SVM, confirming their superiority when applied to intrinsically sequential problems.

6. Conclusion

This paper presents some experiments on the automatic detection of prosodic prominence in continuous Italian speech. Considering that in order to properly define prosodic prominence one needs to take contextual information into account, we tested a number of versions of MLS, able to correctly manage problems involving sequences of input features and sequences of output label classes, to be related in a complex way. In particular, we tested MLS belonging to the large family of PGM.

We thus performed several experiments with CRF, CNF, LDCRF and LDCNF models, obtaining very good classification results (F-measure = 0.770) despite using a small Italian corpus consisting of only 120 utterances.

Considering that, as outlined by [7, 30], prominence perception is highly influenced by the listener's linguistic expectations, there is room for large improvements in the system's performance by including linguistic features in the automatic system.

We are planning to test these models on different corpora and different languages, in order to verify the effectiveness of the proposed approach to automatic prominence detection.

7. References

[1] Abete, G., Cutugno, F., Ludusan, B., Origlia, A., "Pitch behavior detection for automatic prominence recognition", in *Proc. of Speech Prosody 2010*, 2010, Chicago.

[2] Al Moubayed, S., Beskow, J., "Prominence detection in Swedish using syllable correlates", in *Proc. of Interspeech 2010*, Makuhari, Japan, 2010.

[3] Albano Leoni F., "Tre progetti per l'italiano parlato: AVIP, API, CLIPS", in Maraschio N., Poggi Salani T., (eds.), *Italia Linguistica. Anno mille, anno duemila, Atti del XXXIV Congresso della Società di Linguistica Italiana (SLI)*, Roma, Bulzoni, 675-683, 2003.

[4] Bolinger, D., "A theory of pitch-accent in English", *Word*, 14, 1958, 109-149.

[5] Brenier, J.M., Cer, D.M., Jurafsky, D., "The detection of emphatic words using acoustic and lexical features", in *Proc. of Interspeech 2005,* Lisbon, 2005, 3297–3300.

[6] Camacho A., *SWIPE: A sawtooth waveform inspired pitch estimator for speech and music*, PhD Thesis, University of Florida, 2007.

[7] Cole, J., Mo, Y., Hasegawa-Johnson, M., "Signal-based and expectation-based factors in the perception of prosodic prominence". *Laboratory Phonology*, 1, 2010, 425–452.

[8] Cutugno, F., Leone, E., Ludusan, B., Origlia, A., "Investigating syllable prominence with conditional random fields and latent-dynamic conditional random fields", in *Proc. of Interspeech 2012*, Portland (OR), 2012.

[9] Fant G., Kruckenberg A., Liljencrants, J., "Acoustic-phonetic Analysis of Prominence in Swedish", in Botinis, A. (Ed.), *Intonation*, Kluwer Academic Publisher, 2000, 55–86.

[10] Fosler-Lussier, E., Yanzhang, H., Preethi, J. and Prabhavalkar, R., "Conditional Random Field an Speech, Audio and Language Processing", in *Proc. IEEE*, 101(5), 2013, 1054–1075.

[11] Goldman, J-P., Avanzi, M., Auchlin, A., Simon, A.C., "A continuous prominence score based on acoustic features", in *Proc. of Interspeech 2012*, Portland (OR), 2012.

[12] Kocharov, D., "Automatic detection of prominent words in Russian Speech", in *Proc. of the IEEE International Multiconference on Computer Science and Information Technology*, 2010, 435–438.

[13] Kohler, K.J., "Neglected categories in the modelling of prosody - Pitch timing and non-pitch accents", in *Proc. of ICPhS'03*, Barcelona, 2003, 2925-2928.

[14] Kohler, K.J., "Form and Function of Non-Pitch Accents", *Prosodic Patterns of German Spontaneous Speech, AIPUK*, 35a, 2005, 97-123.

[15] Li, K., Zhang, S., Li, M., Lo, W-K., Meng, H., "Prominence model for prosodic features in automatic lexical stress and pitch accent detection", in *Proc. of Interspeech 2011*, Florence, 2011, 2009–2013.

[16] Levesque, J.C., Morency, L.P. and Gagné, C., "Sequential emotion recognition using Latent-Dynamic Conditional Neural Fields", in *Proc. 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, Shanghai, 2013, 1–6.

[17] Ludusan, B., Origlia, A., Cutugno, F., "On the use of the rhythmogram for automatic syllabic prominence detection", in *Proc. of Interspeech 2011*, Florence, 2011, 2413–2417.

[18] Morency, L., Quattoni, A. and Darrell, T., "Latent-dynamic discriminative models for continuous gesture recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, Minnesota, 2007, 1–8.

[19] Origlia, A., Abete, G. and Cutugno, F., "A dynamic tonal perception model for optimal pitch stylization", *Computer Speech & Language*, 27(1), 2013,190–208.

[20] Peng, J., Bo, L. and Xu, J., "Conditional Neural Fields", in *Proc. Neural Information Processing Systems (NIPS)*, Vancouver, Canada, 2009.

[21] Shridar, V.K.R, Nenkova, A., Narayanan, S., Jurafsky D., "Detecting prominence in conversational speech: pitch-accent, givenness and focus", in *Proc. of Speech Prosody 2008,* Campinas, Brazil, 2008.

[22] Sutton, C. and McCallum, A., "An Introduction to Conditional Random Fields", *Foundations and Trends in Machine Learning*, 4(4), 2011, 267–373.

[23] Talkin, D., "A robust algorithm for pitch tracking (RAPT)", in W.B. Kleijn & K.K. Paliwal (Eds.), *Speech coding and synthesis*, New York: Elsevier, 1995, 495–518.

[24] Tamburini F., "Reliable Prominence Identification in English Spontaneous Speech", in *Proc. of Speech Prosody 2006*, Dresden, 2006, PS1-9-19.

[25] Tamburini F., Wagner P., "*On Automatic Prominence Detection for German"*, in *Proc. of InterSpeech 2007*, Antwerp, 2007, 1809–1812.

[26] Tamburini F., "Prominenza frasale e tipologia prosodica: un approccio acustico", in *Proc. Linguistica e modelli tecnologici di ricerca, XL congresso internazionale di studi*, Società di Linguistica Italiana, Vercelli, 2009, 437–455.

[27] Tamburini F., "Una valutazione oggettiva dei metodi più diffuse per l'estrazione automatica della frequenza fondamentale", in *Atti del IX Convegno dell'Associazione Italiana Scienze della Voce*, Venice, in press.

[28] Taylor, P.A., "Analysis and Synthesis of Intonation using the Tilt Model", *J. Acoust. Soc. Amer.*, 107(3), 2000, 1697–1714.

[29] Terken, J., "Fundamental Frequency and Perceived Prominence", *Journal of the Acoustical Society of America*, 89, 1991, 1768–1776.

[30] Wagner, P., "Great Expectations – Introspective vs Perceptual Prominence Ratings and their Acoustic Correlates", in *Proc. of Interspeech 2005*, Lisbon, 2005, 2381–2384.

[31] Zahorian, S.A., Hu, H., "A Spectral/temporal method for Robust Fundamental Frequency Tracking", *Journal of the Acoustical Society of America*, 123(6), 2008, 4559–4571.