PROSODIC DATA DRIVEN MODELLING OF A NARRATIVE STYLE IN FESTIVAL TTS

Fabio Tesser

ITC-IRST
Istituto Trentino di Cultura
Centre for Scientific
and Technological Research
Povo (TN) ITALY

Piero Cosi, Carlo Drioli, Graziano Tisato

ISTC-CNR
Laboratory of Phonetics and Dialectology
Institute of Cognitive Sciences
and Technology
Padova ITALY

ABSTRACT

A general data-driven procedure for creating new prosodic modules for the Italian FESTIVAL Text-To-Speech (TTS) [1] synthesizer is described. These modules are based on the "Classification and Regression Trees" (CART) theory. The prosodic factors taken into consideration are: duration, pitch and loudness. Loudness control has been implemented as an extension to the MBROLA diphone concatenative synthesizer. The prosodic models were trained using two speech corpora with different speaking style, and the effectiveness of the CART-based prosody was assessed with a set of evaluation tests.

1. INTRODUCTION

The task of prosodic modules in TTS synthesizers is that of computing a set of prosodic parameters starting from the linguistic information contained in the text that has to be synthesized. Data driven techniques create the prosodic modules using statistical classification methods for learning the prosody of a real speaker. In other words, starting from a speech corpus it is possible to automatically obtain all the prosodic information needed to build the prosodic modules of the TTS synthesizer. Moreover, with respect to knowledge based approach, data driven techniques simplify the capture of the prosody of a specific speaker or of a particular narrative style, or even of an emotive attitude. However, while using these classification techniques, the serious problem of data-sparseness has to be solved. An obvious solution to this problem is that of increasing the amount of data. As an alternative, one can use specific data representations able to decrease the data-sparseness by grouping homogenous data together. The z-scores for the duration parameter [2], and the VQ-PaIntE [3] model for f0 values are two efficient examples of these transformations. We will use this approach here.

In up to date TTS technologies, synthesis control has been mainly focusing on phoneme duration and pitch, which are the two main parameters conveying the prosodic information. More recently, the speech synthesis community is showing an increasing interest in the control of a broader class of voice characteristics. As an example, voice quality is known to play an important role in emotive speech, and some recent studies have addressed the exploitation of source models within the framework of articulatory synthesis to control the characteristics of voice phonation [4, 5]. As a first step toward providing the control over a broader set of voice parameters, we experimented an extension to a diphone based synthesizer (MBROLA) that allows to control the loudness of the synthesized speech. This feature was exploited in the design of the CART-based prosody by including a loudness-specific functional block in the prosodic module.

2. TTS SYNTHESIS ENVIRONMENT

The investigation relies on the FESTIVAL text-to-speech synthesis framework developed at CSTR [6], and on the MBROLA synthesis engine [7].

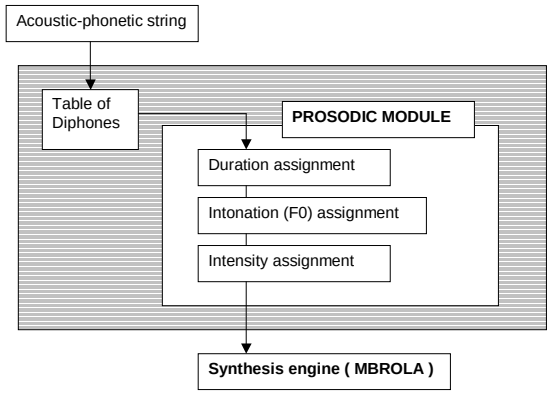

Fig. 1. Scheme of the TTS system. Prosodic module and synthesis engine.

The prosodic module of a text to speech system is aimed at computing the values of a set of prosodic variables. In a minimal configuration these variables are the phoneme duration and f0. The computation of such values can rely on rules or on machine learning methods, such as CARTs [8]. To date, Italian voices are publicly available for the FESTIVAL system for which the phoneme duration and f0 patterns are computed by rules [1, 9]. Results on the use of CARTs trained with different speaking styles have been reported in [10]. We exploit further this approach, and extend the CART-based prosodic module with the support for the intensity contour computation. To the purpose of our study, the MBROLA synthesizer has been modified so as to permit the control of intensity contours via a set of information contained in the input file. An example of an MBROLA input file and the corresponding synthesized speech is shown in Fig. 2.

(a) MBROLA input file

```
_ 25 100 143
a1 400 5 136 100 119 Intensity 0 -10 50 -5 100 -3
v 74 50 118 Intensity 50 -3
a 213 0 120 100 122 Intensity 100 0
_ 200
a1 400 5 136 100 119 Intensity 0 10 20 5 50 -5 100 -8
v 74 50 118 Intensity 50 -10
a 213 0 120 100 122 Intensity 100 -12
_ 25
```

(b) Synthesis result

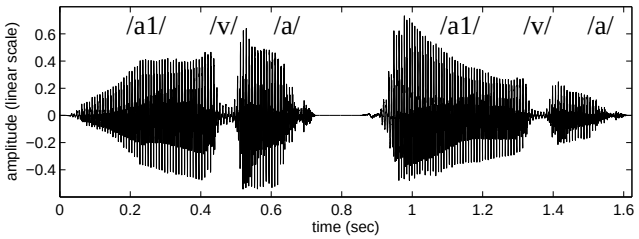

Fig. 2. MBROLA synthesis: (a) input file with extra intensity information for energy envelope control (expressed as ΔI in dB); (b) the resulting synthesis output.

Intensity, intended as the acoustical correlate of loudness, can in principle be roughly controlled by changing a gain factor uniformly across the spectrum. However, it is recognized that the result of such processing is not perceived as natural due to the lack of spectral balance modifications correlated to vocal effort variations that occur in real speech [11, 12]. A first attempt of varying the acoustical correlate of loudness in a more realistic way, i.e., by varying the spectral balance, was made by implementing a spectral weighting function that emphasizes the lower part of the spectrum when lowering the intensity level, and viceversa. This required the embedding of an online spectral processing step acting on the diphones before the overlap-and-add step performed by MBROLA.

3. SPEECH DATABASE

The CARINI database containing the speech recordings of three Italian novels read by a professional speaker was used in this study. The domain of the database is story telling and therefore the style of the database is relatively calm, relaxed and clear reading narrative style. The whole database duration is about 1 hour, for a total of 698 sentences and 7709 words.

4. DESIGN OF PROSODIC MODULES

As data-driven techniques try to model real data, first of all it is necessary to extract the real prosodic contours from the real speech signal. As regards the duration parameters, all the sentences have been automatically segmented and transcribed by using an automatic alignment procedure designed by adapting a "high-performance" Italian phonetic general-purpose speech recognition system developed and trained on the APASCI corpus [13] [14]. The intonation and intensity signals have been extracted by Praat [15].

The standard FESTIVAL heterogeneous relation graph (or HRG) [6] has been created for the whole database, and a particular set of linguistic features have been used for training the CART with the *wagon tool* [16]. The list of features used to train the CART depends on the chosen predictive units (phonemes for the duration, syllables for f0 and intensity) and also on the specific target module.

Some of the features taken into consideration are: the phrase type (declarative, interrogative...), the part of speech of the word, the relative position of the unit in the sentence, in the word and in the syllable, the accent, the level of break after the unit [10].

To have a numerical idea of how good a prosodic module is, a first indication could be given by the RMSE and Correlation between the original prosodic signal and the predicted one. 90% of the database was used for training, 10% for testing. The evaluation was based on a sample of the utterances in the test set. All the error results showed in the tables, are computed in the test set.

4.1. Duration

The duration module must give the rhythm to the sentence and essentially it predicts the duration of the phonemes in the utterance. A very efficient parameter used for building the duration model is the z-score [2]: the duration d_i of each phoneme can be calculated predicting the number of standard deviation from its average k, using the relation $d_i = \mu_i + k\sigma_i$, instead of predicting d_i directly.

The error between the natural phoneme duration and the predicted one is quantified by the values of 0.78 RMSE (in z-score units) and 0.60 Correlation.

4.2. Intonation

The automatic prediction of the fundamental frequency (f0) contour is the most difficult task. The f0 extractors are not completely reliable and data-sparseness is quite a considerable problem. The f0 value range of the same talker can vary a lot between phrases and the realizations of equal perceptive accents can have different f0 values and different pitch shapes. Moreover, more than f0 single values, that are not so important, the particular f0 movements are quite relevant, especially in some specific target syllables bringing important linguistic information. For all these reasons, a modified PaIntE model [17] has been used for creating a good model for predicting f0 trajectories. Two main improvements have been considered with respect to the original implementation: a semitones normalization and a Vector Quantization of the parameters [18].

Semitones normalization is simply a frequency axis modification that is used to transform f0 values in Hz into a feature that attempts to model how people perceive sounds, with the aim of improving the discrimination between f0 trajectories.

$$s(f) = \frac{\log_{f_m}(f) - 1}{\log_{f_m}(2)/12} \quad (1)$$

In (1) f is the frequency value in Hz, f_m is the reference frequency for the semitone and, in our case, f_m has been set to the speaker's f0 mean.

Finally, in order to reduce the data sparseness of the intonation patterns, a Vector Quantization algorithm has been considered and implemented for encoding the PaIntE parameters obtained from the PaIntE analysis and normalization. In this way it is possible to group homogeneous trajectories of the pitch in the same cluster (Fig.3); then for homogeneous cases the CART predict only the codeword rather than a different PaIntE vector. The number of possible cases is limited from the codebook size. We trained the method for different codebook sizes.

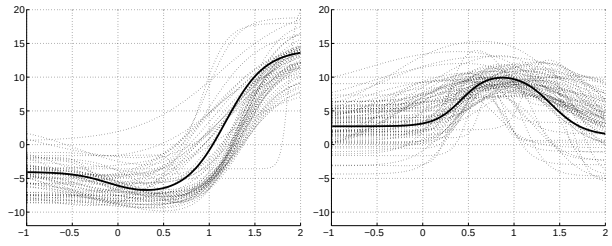

Fig. 3. Clusterization example: PaIntE-modeled f0 patterns (dotted lines) and the corresponding PaIntE codeword (continuous line). The x and y axes are normalized to the syllable length and to speaker f0 mean semitones respectively.

Table 1. RMSE (Hz) and correlation of predicted f0 contours.

CB size	RMSE	Correlation
32	39.68	0.26
64	38.58	0.33
80	36.38	0.43
128	41.67	0.27

As shown in Table 1, the best result is obtained with a codebook size of 80. For smaller codebook sizes (32,64) the performances decrease because the VQ algorithm spilts the cluster in a worst manner, while for greater codebook sizes (128) CART must distinguish between a high number of cases, and so the codebook prediction score decreases.

4.3. Intensity

A differential intensity control has been added to the MBR-OLA synthesis engine, and the differential approach has been preferred to the absolute one due to the high level of segment dependency in the last case. If $\Delta I = x$ the target frame for a particular diphone has to be synthesized with an intensity value x times greater than that of the original value stored in the database.

In Fig. 4A the intensity of a sample utterance is illustrated (**Original**). Figure 4B shows the same utterance synthesized with the same f0 and duration, but without the intensity correction (**Copy**). In Fig. 4C the above sentence is synthesized with the same f0 and duration by driving the intensity control with ΔI value computed as the difference between B and A (**Copy ΔI**). It can be observed that in the B case the intensity pattern is quite flat, while in the C case the intensity pattern is more similar to that of the A case.

In order to train an intensity CART (**Carini ΔI**), the ΔI has been calculated for the whole Carini data-base as the difference between the intensity of the natural sentences spoken by Carini (**Original**) and the corresponding synthetic ones generated by FESTIVAL with the same duration and pitch of the original ones, but without the intensity control (**Copy**).

In order to find the best unit needed to predict intensity differences, various experiments have been designed considering ΔI means computed on words, syllables and phonemes. RMSE values and correlations between the intensity of the **Original** utterance and that of the utterances synthesized with the **Copy**, **Copy ΔI**, and **Carini ΔI** modules, are illustrated in Tables 2,3,4.

As for **Copy ΔI**, it can be observed that a better estimate is obtained while looking at phonemes (Table 4) with respect to syllables and words, and this is due to the augmented resolution with which the intensity differences are computed. Considering **Carini ΔI** the CART intensity pre-

Table 2. RMSE (dB) and correlation, word case.

	RMSE	Correlation
Copy	8.55	0.64
Copy Δ**I**	5.92	0.77
Carini Δ**I**	6.81	0.68

Table 3. RMSE (dB) and correlation, syllable case.

	RMSE	Correlation
Copy	8.51	0.66
Copy Δ**I**	5.59	0.81
Carini Δ**I**	6.74	0.71

diction is similar for syllables and phonemes. This is due to the fact that, even if the size of the training database is bigger in the phoneme case, the prediction is more complex than in the syllable one. Moreover **Copy** has always lower performance, showing a clear preference for intensity control.

Even the intensity pattern shown in Figure 4D, which refers to the utterance synthesized by using the phoneme-based intensity CART, results quite similar to that illustrated in the A case.

5. SUBJECTIVE EVALUATION

To assess the effectiveness of the prosody models and of the training procedures, a set of choice tests was carried out. Four subjects with some degree of expertise in the field and eight with no expertise at all, were selected. The test session was divided in three experiments. In the first one, the listeners were asked to chose one out of two stimuli, on the basis of the following question: "Which prosodic style is the most appropriate to tell a story?". A "no preference" option was also provided. This experiment was intended to assess the preferences of the subjects with respect to the use of different prosodic modules, when listening to a brief excerpt from a literary story or a fairy tale. The two stimuli presented for each comparison were generated by synthesizing the same utterance using the rule-based prosody (**Rules**), the CART-based prosody from the story telling database CARINI (**Carini**), and the CART-based prosody from a news reading database (**Rai-news**, see [10] for details). The overall number of paired comparisons in the first experiment was 18, i.e. all the possible comparisons with 6 sentences. The preferences expressed by the subjects are reported in Table 5. The most of the listeners (88.9%) clearly preferred the narrative styled prosodic module (**Carini**) over the rule-based module (**Rules**), and the **Carini** module (83.3%) over the news reading styled module (**Rai-news**). Interestingly, the majority (52.8%) preferred the rule-based prosody over the (**Rai-news**) prosody. The percentage of "no preference" responses was quite low.

Table 4. RMSE (dB) and correlation, phoneme case.

	RMSE	Correlation
Copy	9.18	0.64
Copy Δ**I**	5.54	0.84
Carini Δ**I**	6.92	0.72

Table 5. The results of the experiment on speaking style appropriateness for (literary) story telling. **Rules** is the rule-based prosodic module, **Carini** is the narrative styled one, and **Rai-news** is the news-reading styled one.

Mod1 vs **Mod2**	**Mod1**	**Mod2**	**No pref.**
Carini vs **Rules**	88.9%	9.7%	1.4%
Carini vs **Rai-news**	83.3%	12.5%	4.2%
Rules vs **Rai-news**	52.8%	34.7%	12.5%

Table 6. The results of the experiment on prosody naturalness. **Rules** is the rule-based prosodic module, **Copy** Δ**I** is the copy prosody stimulus, and **Carini** Δ**I** is the narrative styled one. In the module names, ΔI indicates that the intensity control was used in the synthesis.

Mod1 vs **Mod2**	**Mod1**	**Mod2**	**No pref.**
Copy Δ**I** vs **Rules**	87.5%	6.9%	5.6%
Copy Δ**I** vs **Carini** Δ**I**	68.1%	23.6%	8.3%
Carini Δ**I** vs **Rules**	76.4%	13.9%	9.7%

In the second experiment, the listeners were asked to choose one out of two stimuli, on the basis of the following question: "Which stimulus sounds more natural?". The two stimuli presented for each comparison were generated by synthesizing the same utterance using the rule-based prosody (**Rules**), the prosody (including intensity contour) copied from the Carini's database (**Copy** Δ**I**), and the Carini's CART prosodic module including the intensity control (**Carini** Δ**I**). The overall number of paired comparisons in the second experiment was again 18. The results of the preferences expressed by the subjects is reported in Table 6. The most of the listeners preferred, as expected, the copy prosody (**Copy** Δ**I**) over both the rule-based prosody (**Rules**) (87.5%) and the narrative styled CART-based prosody (**Carini** Δ**I**) (68.1%), with a more pronounced preference over the rule-based. They owever judged more natural the CART-based prosody when compared to the rule-based prosody. As in the first experiment, the percentage of "no preference" responses was quite low.

Finally, in the third experiment, the listeners were asked to chose one out of two stimuli, on the basis of the same

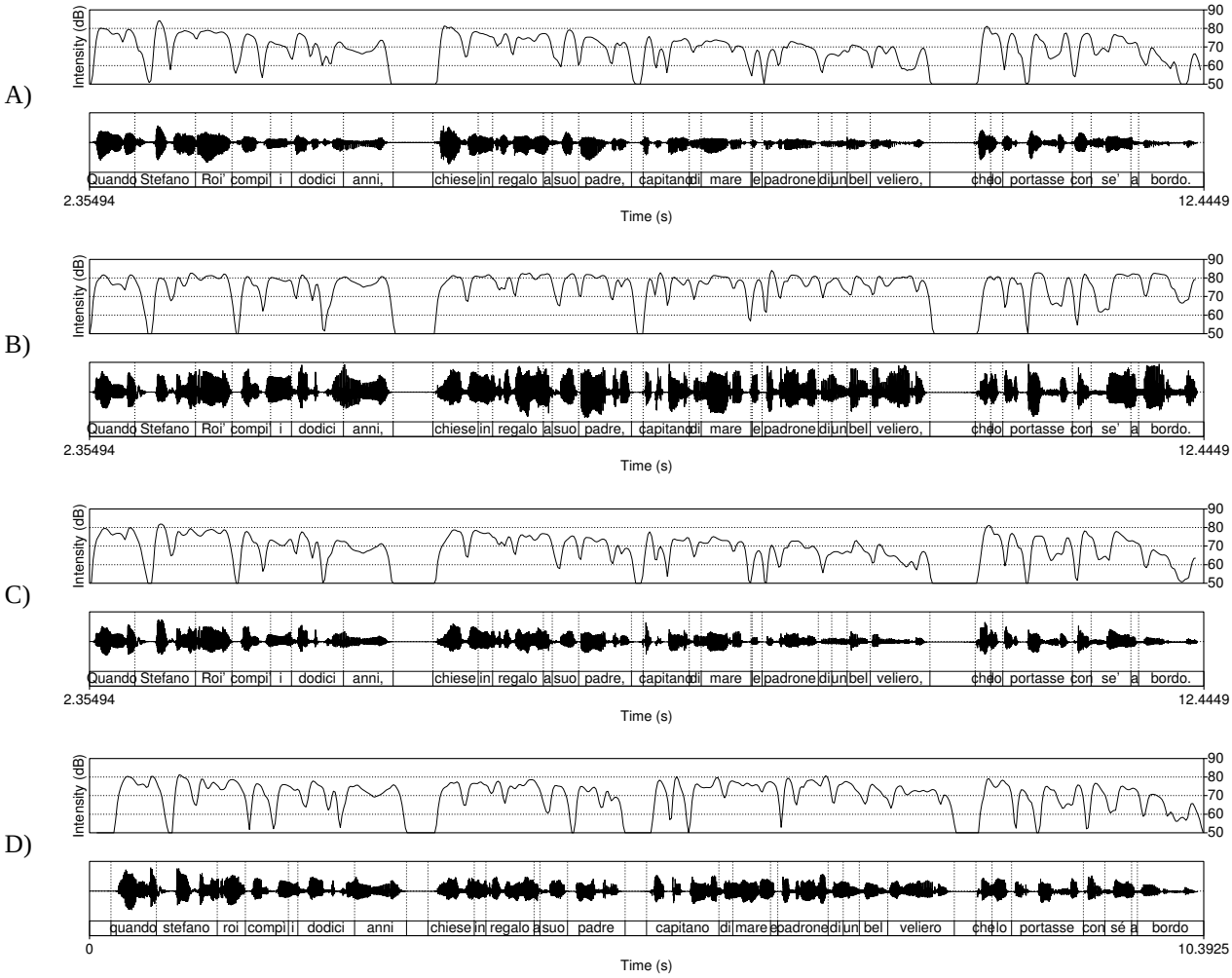

Fig. 4. A) **Original** B) **Copy** C) **Copy** Δ**I** D) **Carini** Δ**I**

question on naturalness given in experiment 2. The third experiment was intended to provide an indication on whether the subjects could appreciate the difference due to the intensity control. The two stimuli presented for each comparison were generated by synthesizing the same utterance using the Carini's CART prosodic module, with (**Carini** Δ**I**) and without (**Carini**) the intensity control. The overall number of paired comparisons in the third experiment was 6. The results of the preferences expressed by the subjects is reported in Table 7. Results show that the difference was not much appreciated, most probably because the attention of the listeners was focused on intonation. We also realized that the length of the stimuli was excessive, and this prevented the listeners from selectively compare short segments of the signal where the difference was rather appreciable.

Finally, in all the experiments, no significative differences were observed in the preferences expressed by expert and non-expert subjects.

Table 7. The results of the experiment on synthesis naturalness. **Carini** refers the CART-based prosodic module, and **Carini** Δ**I** refers to the CART-based prosodic module with intensity control.

Mod1 vs **Mod2**	**Mod1**	**Mod2**	**No pref.**
Carini Δ**I** vs **Carini**	27.8%	27.8%	44.4%

6. CONCLUDING REMARKS

The training of the intensity CART was performed on a word-, syllable-, and phoneme-basis. The results showed that, while the word-based training had the worst performance, no relevant differences were observed between the syllable-based and the phoneme-based training procedures.

The subjective tests performed showed that the CART-based approach to the modeling of a speaking style was ef-

fective in reproducing the narrative style of our CARINI database. The listeners preferred the synthesis with this prosodic style over the rule-based prosody and over a news reading CART-based prosody when listening to a story. They also judged this style more natural in general if compared with the rule-based prosody.

7. FUTURE TRENDS

In order to further investigate the modeling of the ΔIntensity trajectories, improvements are foreseen such as the use of the PaIntE for pitch.

The spectral compensation processing included in the MBROLA intensity control extension requires further refinement to improve the listening test performance. Moreover, the control on an extended set of voice characteristics, such as soft/pressed quality, breathiness, harshness, is under study at present. This should allow to change the voice quality of the speaker, so as to reproduce more effectively all the acoustic peculiarities that characterize the voice in acted or emotive speech.

8. ACKNOWLEDGEMENTS

Part of this work has been sponsored by PF-STAR (Preparing Future multiSensorial inTerAction Research, European Project IST- 2001-37599, http://pfstar.itc.it). We wish to thank the MBROLA team for providing the source code of their synthesis engine.

9. REFERENCES

[1] P. Cosi, F. Tesser, R. Gretter, and C. A. (with Introduction by Mike Macon), "Festival speaks italian!" in *Proceedings of EUROSPEECH 2001*, Aalborg, Denmark, Sept 2001, pp. 509–512.

[2] N. Campbell and S. Isard, "Segment durations in a syllable frame," *Journal of Phonetics*, pp. 37–47, 1991.

[3] G. Möhler, "Describing intonation with a parametric model," in *Proceedings of ICSLP98*, Sydney, 1998, pp. 2581–2584.

[4] C. d'Alessandro and B. Doval, "Experiments in voice quality modification of natural speech signals: the spectral approach," *in Proceedings of the 3rd ESCA/COCOSDA Int. Workshop on Speech Synthesis*, pp. 277–282, 1998.

[5] C. Gobl and A. N. Chasaide, "The role of the voice quality in communicating emotions, mood and attitude," *Speech Communication*, vol. 40, pp. 189–212, 2003.

[6] P. Taylor, A. Black, and R. Caley, "The architecture of the Festival speech synthesis system," *3rd ESCA Workshop on Speech Synthesis*, pp. 147–151, 1998.

[7] T. Dutoit and H. Leich, "MBR-PSOLA : Text-To-Speech synthesis based on an MBE re-synthesis of the segments database," *Speech Commun.*, vol. 13, no. 3-4, pp. 167–184, November 1993.

[8] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and regression trees*. Wadsworth and Brooks, 1984.

[9] Italian voice for the FESTIVAL speech synthesis system. Download page: http://www.csrf.pd.cnr.it/TTS/It-FESTIVAL.htm.

[10] P. Cosi, C. Avesani, F. Tesser, R. Gretter, and F. Pianesi, "On the use of Cart-Tree for prosodic predictions in the Italian Festival TTS," *Voce, Canto, Parlato - Studi in onore di Franco Ferrero*, pp. 73–81, 2002.

[11] A. Sluijter, V. van Heuven, and J. Pacilly, "Spectral balance as a cue in the perception of linguistic stress," *J. Acoust. Soc. Am.*, vol. 101, no. 1, pp. 503–513, 1997.

[12] W. Campbell, "Loudness, spectral tilt, and perceived prominence in dialogues," in *Proceedings ICPhS 95*, Stockholm, Sweden, 1995, pp. 676–679.

[13] P. Cosi and J. Hosom, "High performance "general purpose" phonetic recognition for Italian," in *Proceedings of International Conference on Spoken Language Processing*, vol. 2, Beijing, Cina, October 2000, pp. 527–530.

[14] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, and M. Omologo, "A baseline of a speaker independent continuous speech recognizer of italian," in *Proceedings of EUROSPEECH 93*, Berlin, Germany, 1993, pp. 847–850.

[15] P. Boersma, "PRAAT, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001, PRAAT web site: http://www.fon.hum.uva.nl/praat/.

[16] P. Taylor, R. Caley, A. W. Black, and S. King, "Edinburgh speech tools library," http://www.cstr.ed.ac.uk/projects/speech_tools/.

[17] P. Cosi, F. Tesser, R. Gretter, and F. Pianesi, "A modified "PaIntE Model" for Italian TTS," in *CDROM proceedings of IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, California, Sept 2002.

[18] G. Möhler and A. Conkie, "Parametric modeling of intonation using vector quantization," in *CDROM proceedings of 3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998.