A MODIFIED “PaIntE” MODEL FOR ITALIAN TTS

Piero Cosi, Cinzia Avesani

ISTC-SFD - (ex IFD) CNR
Istituto di Scienze e Tecnologie della Cognizione
Sezione diPadova “ Fonetica e Dialettologia”
(ex Istituto di Fonetica e Dialettologia)
Consiglio Nazionale delle Ricerche
e-mail: {cosi, avesani}@csrf.pd.cnr.it
www: http://nts.csrf.pd.cnr.it/
www: http://www.csrf.pd.cnr.it/

Fabio Tesser, Roberto Gretter, Fabio Pianesi

ITC-IRST
Istituto Trentino di Cultura
Istituto per la Ricerca Scientifica e Tecnologica
e-mail: {pianesi,gretter,tesser}@irst.itc.it
www: http://www.itc.it/IRST/index.htm

ABSTRACT

In this work, a slightly modified version of the original PaIntE model, based on an F0 parametrization with an especially designed approximation function, is considered. The model’s parameters have been automatically optimized using a small set of Italian ToBI labeled sentences. This method, will drive our ongoing data-based approach to intonation modeling for Italian TTS. The quality of the model has been assessed by numerical measures and preliminary tests show quite promising results.

1. INTRODUCTION

Most intonation theories hypothesize that intonation can be modeled with a set of distinct phonological entities phonetically realized as F0 movements.

The tone sequence model (TSM) introduced by Pierrehumbert [1], and its associated labeling convention ToBI [2], characterizing the intonation contour as a sequence of high (H) and low (L) tones, represent the most prominent example of such a phonological description of intonation. For different languages, pitch accents and boundaries are made up of different sets of H and L targets.

The so-called British School of intonation, with a different viewpoint, places emphasis on pitch movements instead of targets [3], and the inventory of movements basically consists of falling, rising, falling-rising movements and combinations of these.

Even if they have different ways of describing intonation, both theories share a compositional method of describing intonation. They combine, in fact, a number of distinct basic elements to make up the intonation contour.

Data-based approaches, however, in contrast to classical intonation research, often use continuous parameters for the description of F0 contours. This is mainly due to the practical reasons that the underlying functions used for the approximation are shaped by continuous parameters. F0 contours are, in fact, approximated by appropriate model functions varying a set of n continuous parameters, and the result, represented by an n-dimensional vector, is a characterization of the underlying pitch movement.

The TILT model [4-5], for example, well represents such data-based approaches. TILT models intonation events by a five-dimensional vector with a shape parameter called tilt that describes the falling or rising characteristic of F0 movements. The *tilt* parameter is continuous, assuming any value between -1 and 1, thus allowing intermediate shapes with falling and rising parts of different height. The other four parameters model the alignment of the shape relative to the accented syllable, the steepness of the curve, its amplitude and base F0 level. This approach has been successfully applied to intonation modeling, where the five parameters are predicted from appropriate features of an utterance [6].

Parametric models are quite effective in representing intonation because they well simulate true F0 movements and they can convey specific prosodic meaning to their parameters. Moreover, in comparison with speech recognition features that try to capture the most effective information needed to correctly identify different sounds discarding all redundancies, we can say that these models try to capture, in the same way, only the intonation related information.

Differently from other intonation models, the two contradictory principles expressed by classical intonation research and data-based approaches, have been simultaneously incorporated into the **PaIntE (Pa**rametric representation of **Int**onation **E**vents) model [7-8]. This model uses, in fact, an F0 parametrization with 6 continuous parametric intonation event parameters. These PaIntE parameters are derived by approximating the F0 curve with an appropriate model function that is focused, with a three syllable span, on ToBI labeled target points.

2. THE PaIntE MODEL

The **Pa**rametric representation of **Int**onation **E**vents (**PaIntE**) model [7-8] approximates stretches of F0, spaning three syllables centered on a target ToBI label, by a phonetically motivated model function consisting of a sum of two sigmoids with a fixed time delay as shown in Figure 1 for a typical “peak” F0 type (H*). Each single sigmoid can be described by four parameters modeling its floor, amplitude, alignment and steepness. Given a common upper limit *d* of the two sigmoids and a constant alignment parameter γ, the model function can be described by the following equation:

$$f(x) = d - \frac{c_1}{1+\exp(-a_1(b-x)+g)} - \frac{c_2}{1+\exp(-a_2(x-b)+g)} \qquad (1)$$

with

- a1, a2: steepness of the rising and falling sigmoids.
- b: alignment of the function (the syllable length is defined as unity).
- c1, c2: amplitudes of the rising and falling sigmoids.
- d: frequency of the peak of the function.

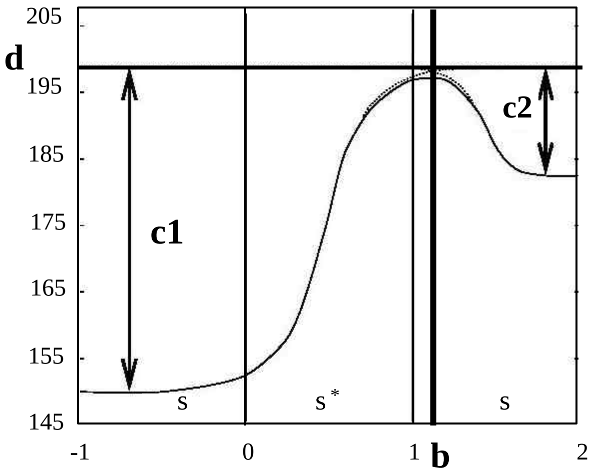

Figure 1: The PaIntE model function. A rising and a falling sigmoid with a fixed time delay are summed (the time axis is normalized to the syllables' lengths).

In order to decorrelate the F0 shape from the speaking rate and from the type of the syllables the x axis is normalized with respect to the durations of the syllables.

In our implementation the PaIntE model has been modified in order to cope with F0 "valley-like" shapes other than "peak-like" shapes, as commonly used in ToBI labeling, thus allowing c_1 and c_2 to get negative values. This has been done especially because, for Italian [9], there exist some ToBI tones such, for example L*, and some boundary tones that are better represented by this "valley-like" shape. The six parameters thus assume different meanings depending on the "peak" (c_1, $c_2 > 0$) or "valley" shape (c_1, $c_2 < 0$) as illustrated in Figure 2.

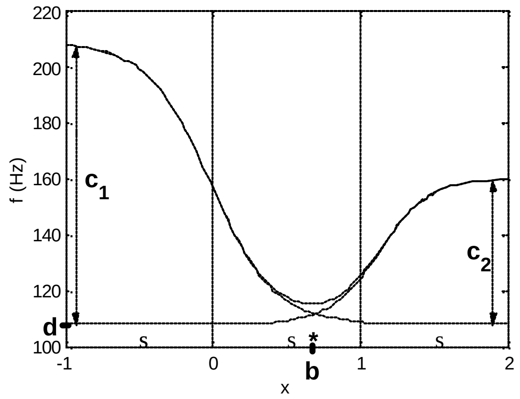

Figure 2: PaIntE model function for a "valley" shape.

Following [8] a pitch range normalization to each individual sentence has been applied in order to filter out the influence of different pitch-range levels. Taking into consideration both "peak" and "valley" shapes the upper and lower level of the pitch range are defined by:

$$\begin{aligned} UL &= \max\ [\max d, \max(d-c1), \max(d-c2)] \\ LL &= \min\ [\min d, \min(d-c1), \min(d-c2)]. \end{aligned} \qquad (2)$$

As for comparison, a PaIntE model was also designed without ToBI information. In such a case the parametrization was executed in all the stressed syllables belonging to true content words and in all phrase final syllables.

3. SPEECH DATABASE

A broadcasted news corpus, spoken by a national TV announcer (RAI-news) [10] (558 sentences, ~11500 words), was considered for training the NO-ToBI PaIntE (P*aIntE-NT*) model, while a small set of ToBI-labelled sentences of the same corpus have been used in order to compute and optimize the ToBI PaIntE (*PaIntE-T*) model's parameters.

4. EXPERIMENT

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 [1\] developed and trained on the APASCI corpus [12] using the CSLU Speech Toolkit [13].

F0, computed by Praat [14], has been interpolated in the unvoiced portions ($F0_I$) and smoothed by a 20Hz low-pass filter ($F0_{IS}$). For each ToBI-tagged syllable an $F0_{IS}$ window of three syllables centered on the target one $F0_{ISW}$ has been considered for the automatic parametrization procedure. A time normalization has been also considered with respect to the syllable length in order to discard speaking rate influence $F0_{ISWN}$. This portion of F0 is approximated by the model function introduced in (1) using a conjugate gradient method following:

$$\min_{d,b,a_1,c_1,a_2,c_2} \left\{ \frac{1}{2} \sum_{x \in [-1,2]} p(x)^2 [f(x) - F_{0-ISWN}(x)]^2 \right\} \qquad (3)$$

in order to estimate the PaIntE parameters representing the F0 movements within the window. As indicated in (3) a weight triangular shape function *p(x)*, centered on each "critical" point used in the optimization algorithm, with amplitude proportional to the "power" of each critical point and extension proportional to the corresponding influencing zone, was introduced.

As exemplified for a simple L+H* case in Figure 3, with a trial-and–error strategy, a maximum of eight critical points were positioned in:

- the ToBI events (in a bi-tonal ToBI event such as L+H* there are two critical points);
- the windows's boundary points (-1, and 2 in Figure 1,2 and 3);
- the position of F0max and F0min for each syllable;
- the position of max and min F0 derivative for each syllable.

Some of the critical points have been chosen in order to determine the initialization values of the optimization algorithm, which is executed two times with two different sets of initialization points corresponding to a "peak" and to a "valley" configuration, following (3).

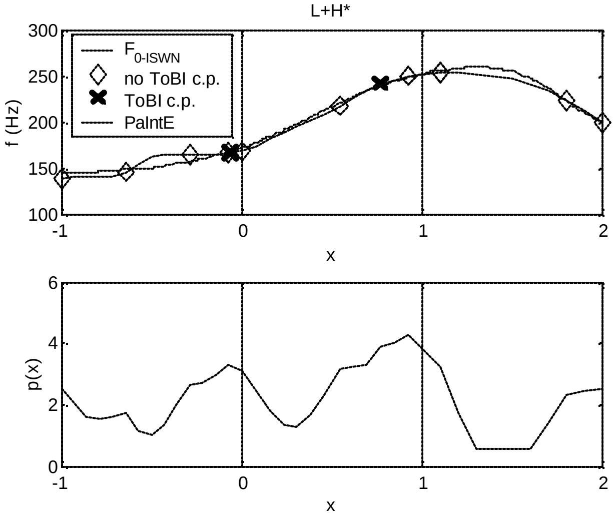

Figure 3: L+H* ToBI example and corresponding weight p(x) function (PaIntE: [a.,b,a1,c1,a2,c2] = [268.3668, 1.350618, 3.126952, 125.1043, 7.031559, 81.88443])

The reconstruction of F0 contour, starting from PaIntE parameter corresponding to ToBI events, is obtained by interconnecting all intonation events by the use of a linear interpolation procedure. Examples of an original F0 contour and of its corresponding PaiIntE reconstructed contour are given in Figure 4.

In order to test and validate the algorithm, a comparison between the original and the reconstructed F0 contours has been executed by examining the corresponding RMSE and correlation values for the whole no-ToBI labeled RAI-news corpus [10]. As shown in Table 1, the mean values of these two validation indexes for 558 sentences are quite promising.

Moreover, a comparison between the original and the reconstructed F0 contours has been executed also on three example sentences of the same RAI-news corpus in the ToBI and no-ToBI case. In Table 2, it can be noted that, in the ToBI case results are quite better than those obtained in the no-ToBI one.

RAI-news Corpus	**No-ToBI**	
	RMSE	Correlation
(558 sentences)	**12.764**	**0.8845**

Table 1: RMSE and Correlation mean values of original and PaIntE reconstructed F0 contours for the whole no-ToBI labeled RAI-news [10] corpus.

	ToBI		**No-ToBI**	
sentence	RMSE	Correlation	RMSE	Correlation
f0002	9.15	0.964	9.87	0.959
f0004	6.76	0.968	12.96	0.879
f0007	6.46	0.976	11.82	0.917
mean	**7.46**	**0.969**	**11.55**	**0.918**

Table 2: RMSEs, correlations and means of original and PaIntE reconstructed F0 contours for 3 example sentences of the RAI-news corpus [10].

However, these numerical indexes do not take into consideration the real perceptive differences between the original and the reconstructed stimuli, thus a much more complete and precise perceptual test has to be designed to cover this matter. In our preliminary tests, the perceptive differences between the original sentences and the PaIntE reconstructed ones, by using a simple PSOLA resynthesis [15] algorithm implemented in PRAAT [14], are quite small, and we believe this is due to the fact that the ToBI parameterization is executed only in those critical points perceptually more relevant while discarding all other irrelevant portions.

5. CONCLUDING REMARKS

The modified PaIntE model described in this study seems quite appropriate for describing and predicting F0 contours of Italian. We believe this is mainly because PaIntE "valley" and "peak" configurations well model the F0 contours in specific intonation points, and because perceptually relevant critical points have been chosen for the optimization algorithm.

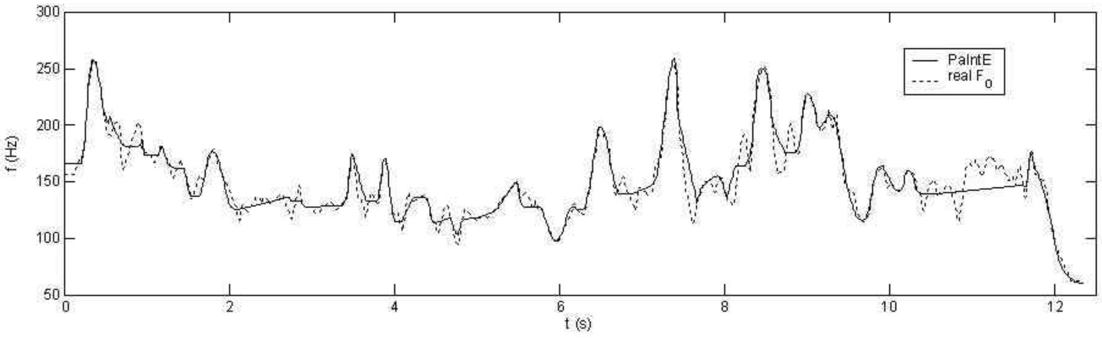

Figure 4: Example of the application of the PaIntE-T reconstruction algorithm to a simple Italian sentence.

The results of the numerical evaluation and of the preliminary perceptual tests, look quite promising and suggest that we can better predict intonation when more prosodic information is available. More testing is obviously needed for an in-depth evaluation of the model and also of the comparison of different configurations (like number of critical points) of the optimization algorithm.

6. FUTURE TRENDS

In order to better compare the ToBI and NO-ToBI PaIntE models a bigger ToBI labeled corpus will be analyzed. In particular, for that purpose, a corpus with some novels by Dino Buzzati, a well known Italian writer, spoken by a professional speaker [16], will be considered.

Moreover, inspired by some typical normalized PaIntE shapes extracted for some Italian ToBI tones, as shown in Figure 5, and motivated by those intonation theories that suggest that pitch accent and boundary phenomena can be described by a distinct number of patterns, it is easy to hypothesize that typical intonation patterns can be divided into different categories thus making the VQ-based method a promising candidate for future research on intonation modeling.

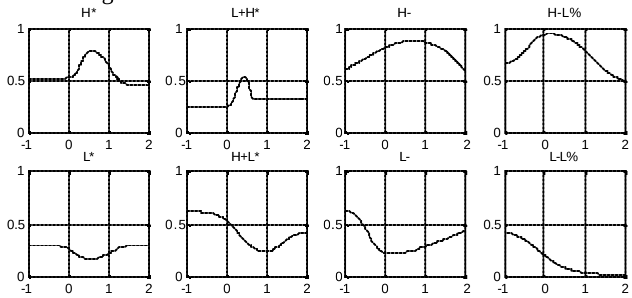

Figure 5: Typical PaIntE shapes extracted for some of the Italian ToBI tones.

Moreover, within the same ToBI tone, F0 shapes often look similar, thus supporting the idea that these patterns can be easily learned by a CART statistical procedure [17].

For these reasons, both VQ and CART will be exploited in the future, and they will be included in the final version of the Italian TTS Festival system [18].

7. ACKNOWLEDGMENTS

Part of this work has been sponsored by MPIRO (Multilingual Personalized Information Objects. European Project IST-1999-10982, http://www.ltg.ed.ac.uk/mpiro/) and TICCA (Tecnologie cognitive per l'interazione e la cooperazione con agenti artificiali) a joint “CNR-Provincia Autonoma Trentina” Project.

8. REFERENCES

[1] Pierrehumbert J., *The Phonology and Phonetics of English Intonation*, PhD thesis, MIT, Cambridge, MA,1980.

[2] Silverman K., Beckman M., Pitrelli J., Ostendorf M., Wightman C., Price P., Pierrehumbert J., Hirschberg J., “ToBI: a Standard for Labeling English Prosody”, in *Proc. of ICSLP 1992*, Vol 2, pp. 867-870.

[3] Halliday M.A.K., *Intonation and grammar in British English*, Mouton, The Hague, 1967.

[4] Taylor P., Black A.W., “Synthesizing conversational intonation from a linguistically rich input”, in *Proceedings of ESCA Workshop on Speech Synthesis*, Mohonk, NY, 1994, pp. 175-178.

[5] Taylor P., “The Tilt Intonation Model”, in *Proc.ICSLP-1998*, Sydney Australia, 30th Nov-4th Dec 1998, Paper 827, Vol. IV, pp. 1383-1386.

[6] Dusterhoff K., Black A.W. “Generating F0 Contours for Speech Synthesis Using the Tilt Intonation Theory”, in *Proceedings of ESCA Workshop on Intonation*, Athens, Greece, 1997.

[7] Mohler G., Conkie A., “Parametric Modeling of Intonation Using Vector Quantization”, in *Prooceedings of Third International Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998.

[8] Möhler G., “Improvements of the PaIntE Model for F0 Parametrization. *Research Papers, Draft version, from the Phonetics Lab, AIMS Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung* (to appear).

[9] Avesani, C., “ToBIt: un sistema di trascrizione per l' intonazione italiana”, in *Atti delle 5e Giornate di Studio del Gruppo di Fonetica Sperimentale (A.I.A.)*, 1995, Povo (TN), Italy, pp. 85-98.

[10] Federico M., Giordani D., Coletti P., “Development and Evaluation of an Italian Broadcast News Corpus”, in *Proc. LREC-2000*, Athens, Greece, 2000.

[11] Cosi P., Hosom J.P., High Performance “General Purpose” Phonetic Recognition for Italian, in *Proc. ICSLP-2000*, Beijing, Cina, 16-20 October, 2000, Vol. II, pp. 527-530.

[12] Angelini B., Brugnara F., Falavigna D., Giuliani D., Gretter R., Omologo M., “A Baseline of a Speaker Independent Continuous Speech Recognizer of Italian”, in *Proc. of EUROSPEECH 93*, Berlin, Germany, 1993.

[13] Sutton S., Cole R., Villiers J., Schalkwyk J., Vermeulen P., Macon M., Yan Y., Kaiser E., Rundle B., Shobaki K., Hosom P., Kain A., Wouters J., Massaro D., Cohen M., “Universal speech tools: the CSLU toolkit”. In *Proc. of ICSLP-98*, Sydney, Nov 30-Dec 4, 1998, Vol. 7, pp. 3221-3224.

[14] Boersma P., “Praat, a system for doing phonetics by computer”, *Glot International* 5 (9/10), pp. 341-345.

[15] Dutoit T., Leich H., MBR-PSOLA: Text-To-Speech Synthesis based on an MBE Re-Synthesis of the Segments Database, *Speech Communication*, Elsevier Publisher, December 1993, vol. 13, n° 3-4, pp. 435-440.

[16] “Il Narratore”, http://www.ilnarratore.it

[17] Breiman L., Friedman J., Stone C.J., Olshen R.A., *Classification and Regression Trees*, Chapman & Hall/CRC, 1984.

[18] Cosi P., Tesser F., Gretter R., C. Avesani C. (with Introduction by Mike Macon), Festival Speaks Italian!, in *Proc. of EUROSPEECH 2001*, Aalborg, Denmark, Sep 3-7 2001, pp. 509-512.