PROSODY MODIFICATION AND FUJISAKI'S MODEL: PRESERVING NATURAL SOUNDNESS

Pierluigi Salvo Rossi[(1)], *Patrizia Falco*[(2)], *Alessandra Budillon*[(3)], *Davide Mattera*[(2)], *Francesco Palmieri*[(3)]

[(1)] Dipartimento di Informatica e Sistemistica, Università di Napoli "Federico II"
[(2)] Dipartimento di Ingegneria Elettronica e della Telecomunicazioni, Università di Napoli "Federico II"
[(3)] Dipartimento di Ingegneria dell'Informazione, Seconda Università di Napoli

salvoros@unina.it, alebudil@unina.it, mattera@unina.it, frapalmi@unina.it

ABSTRACT

Control of prosodic characteristics is one of the most important problems in the area of speech synthesis. Fujisaki's model is probably the best model for pitch variations and its inversion is suitable for being integrated within speech synthesizres. This paper proposes a speech synthesis method based on Fujisaki's model (combined direct and inverse modeling) in order to preserve natural soundness of synthesized speech. The idea is to modify a pitch contour on the basis of Fujisaki's features and a reference contour. Experimental results have shown that using constraints related to Fujisaki's model guarantees good natural-sounding speech synthesis.

1. INTRODUCTION

The most widely used techniques for speech synthesis are concatenative [4]. The provide good segmental quality compared to other methods, such as formant synthesis [1]. On the other hand concatenative synthesis shows scarce control of prosodic characteristics, even though several techniques, such as PSOLA [4], have been developed for modifying speech prosody.

Although concatenation and prosody-modification algorithms are quite efficient, actual synthesizers cannot still provide good supra-segmental quality speech: natural-sounding speech syntesis is still hard to obtain. Storage of large databases helps Text-to-Speech systems to select the appropriate templates, but automatic systems changing the prosody of an utterance (e.g. an assertion into a question) are still quite difficult to implement if the objective is to preserve natural soundness.

In this scenario it is easy to understand how devise a robust prosody model for speech-synthesis is highly desirable. The model must be strictly connected to physical and linguistic structures of speech as implementation of good-performance natural-sounding speech synthesizers cannot leave those structures out of consideration.

We focus on analysis of pitch contours as intonation is an acceptable description of prosody even though a more accurate description should include also duration and intensity. Fujisaki's model [2] is one of the most manageable and powerful model for prosody manipulation. It has shown a remarkable effectiveness in describing pitch contours and its validity has been tested on several languages [6][8]. Several techniques have been proposed to solve its inverse problem [7][9][12][13].

In this paper we propose a method to integrate Fujisaki's model into a speech synthesizer in order to preserve natural-soundness of synthesized speech signals. An utterance is processed in order to assume a pitch contour that best matches, with appropriate constraints, a reference one.

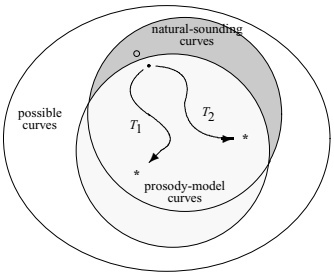


Figure 1: Prosody model for speech synthesis: pitch contour extracted from speech (○), pitch contour representation by use of the prosody model (●), possible modified pitch contours generated by use of the prosody model (∗) with appropriate prosodic-features manipulations (T_i).

2. FUJISAKI'S MODEL

H. Fujisaki and his co-workers proposed, between the 70s and the 80s, an analytical model describing the fundamental frequency (F_0) variations [2]. It captures the essential mechanisms, involved in the speech production, that are responsible of a particular prosodic structure. Subsequently the representation of speech prosody in terms of the model features, i.e. the inverse problem, has been approached with various methods [7][9][12][13]. Fujisaki's model has proven to give good overlapping between sets of model contours and natural ones (see Fig. 1).

2.1 The model

The model, shown in Fig. 2, assumes that the F_0 contour (in a logarithmic scale) is the superposition of two contributions: a *phrase component* and an *accent component*, obtained by filtering two signals. The first contribution (y_p), which models the pitch baseline, accounts for speaker declination and it is characterized by a fast rise followed by a slower fall. The second contribution (y_a), which models smaller-scale prosodic variations, accounts for accent components. The two components are superimposed to a constant value related to the minimum value of speaker's F_0 to realize a particular melodic structure. The first input signal (x_p) is composed

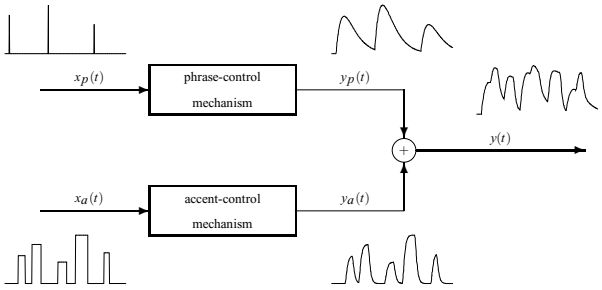


Figure 2: Fujisaki's model.

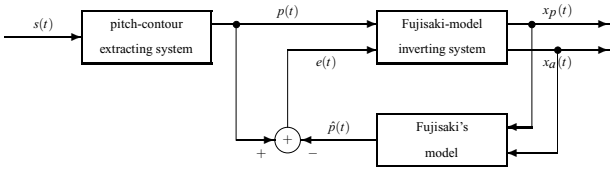


Figure 3: Block diagram of the inverse problem.

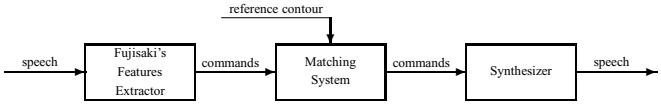


Figure 4: Reference scheme.

by Dirac impulses, namely *phrase commands*, located at the onsets of phrase activities while the second one (x_a) is composed by rectanglular pulses, namely *accent commands*. The linear systems processing x_p and x_a, namely *phrase control* and *accent control mechanisms*, are characterized by:

$$h_p(t) = \alpha^2 t e^{-\alpha t} u(t)\,, \tag{1}$$

which is the impulse response of the phrase control mechanism, where $\alpha \in [2,4]\ s^{-1}$ is its natural angular frequency, and

$$g_a(t) = [1-(1+\beta t)e^{-\beta t}]u(t)\,, \tag{2}$$

which is the step response of the accent control mechanism, where $\beta \in [19,21]\ s^{-1}$ is its natural angular frequency. The total pitch contour is then expressed as

$$\begin{aligned} y(t) &= \ln[F_0(t)] - \ln(F_{min}) = y_p(t) + y_a(t) \\ &= \sum_{k=1}^{N_p} A_{p,k}\, h_p(t - t_{p,k}) + \\ &+ \sum_{k=1}^{N_a} A_{a,k}\, [g_a(t - t_{a,k}^{'}) - g_a(t - t_{a,k}^{''})]\,, \end{aligned} \tag{3}$$

where F_{min} is the minimum value of speaker's F_0 ; N_p and N_a are the number of phrase and accent events; $A_{p,k}$ and $t_{p,k}$ are the magnitude and the timing of the k-*th* phrase command; $A_{a,k}$, $t_{a,k}^{'}$ and $t_{a,k}^{''}$ are the magnitude, the onset and the end of the k-*th* accent command. A non-linear system, accounting for possible glottal effects, has been ignored as it is rather irrelevant to our study.

2.2 The inverse problem

Integration of Fujisaki's model knowledge in a speech synthesizer requires the implementation of an automatic procedure to extract prosodic events from speech in term of model features (model inversion). Fujisaki's model output to the extracted features must optimally match the originary pitch contour.

The inverse problem is approached [13] by means of a starting procedure that guesses a first estimation of phrase and/or accent components and of a subsequent processing that refines the solution. A feedback that compares the originary pitch with the one obtained from the estimation allows recursive solution refinements (see Fig. 3). First estimations are based on high-pass filtering F_0 contour [7], or on low-pass filtering F_0 contour [9], or on differentiating F_0 contour [12], or on selecting local minima and maxima of F_0 contour [13].

3. NATURAL-SOUNDING SYNTHESIS

In this paper we focus on the problem of changing the prosody of an utterance on the basis of a "desired" reference prosodic contour. We would like to do so by preserving natural soundness. Direct superposition of a reference pitch contour on a given speech segment generally does not produce satisfactory synthesis. This is because modification should be strongly anchored to linguistic events. The challenge is to be able to do so without complicated unpacking of textual information. We found that acting on events of Fujisaki's model may be sufficient to constraint pitch modification to obtain natural-sounding speech.

The reference scheme is showed in Fig. 4. A speech signal is processed to obtain a prosody representation in a features domain in terms of phrase and accent commands. A matching system processes the features in order to minimize a cost function depending on a target reference pitch contour to obtain a synthetic pitch contour to be superimposed to the speech signal.

We consider that both the utterance to be processed and the reference contour present only one phrase command, so that the analysis can be referred only to the accent component. More specifically we consider only amplitude modifications of the accent commands.

Even though the model has been described with reference to continuous-time domain signals, our experiments are obiouvsly run on sampled signals using digital filters. The digital filters, used to simulate the phrase-control and the accent-control mechanisms, are designed via the pulse-invariance and the step-invariance techniques [3], respectively. This has been a natural choice for the kind of the input sequences assumed in the model. From here on, the discussion will be presented with reference to discrete-time domain signals, with obvious corresponding notation.

3.1 Problem formulation

Let $s(n)$ be the utterance to be processed and $y(n)$ the pitch contour extracted from it. Let $d(n)$ be the accent component of the reference contour, that has been previously normalized in amplitude and duration with respect to $y(n)$. Let the estimated phrase and accent component of $y(n)$ be

$$y_p(n) = A_p \, h_p(n - n_p) \tag{4}$$

and

$$y_a(n) = \sum_{k=1}^{N_a} A_{a,k} \left[g_a(n - n'_{a,k}) - g_a(n - n''_{a,k}) \right] , \tag{5}$$

respectively. Then our objective is to find a set of command amplitudes $\{A^*_{a,1}, A^*_{a,2}, \ldots, A^*_{a,N_a}\}$ to minimize the cost function

$$\varepsilon = \sum_{n=1}^{L} \left[d(n) - y_a(n) \right]^2 , \tag{6}$$

with respect to $\{A_{a,1}, A_{a,2}, \ldots, A_{a,N_a}\}$, where L is the number of signal samples.

Let

$$y^*_a(n) = \sum_{k=1}^{N_a} A^*_{a,k} \left[g_a(n - n'_{a,k}) - g_a(n - n''_{a,k}) \right] , \tag{7}$$

then our synthetic pitch contour is obtained as

$$y^*(n) = y_p(n) + y^*_a(n) . \tag{8}$$

Such a contour is then superimposed to the speech signal $s(n)$ by using a PSOLA synthesis technique [4]. The procedure results in an utterance whose intonation resembles the reference contour and whose natural soundness is generally preserved.

3.2 Synthesis parameter calculation

Let $h_a(n)$ be the impulse response of the digital filter that simulates the accent control mechanism, let $\mathbf{x}_a$, $\mathbf{y}_a$ and $\mathbf{d}$ be the column vectors whose elements are the samples of the signals $x_a(n)$, $y_a(n)$ and $d(n)$ respectively, and let $\mathbf{a}$ be the column vector whose k-*th* element is the amplitude $A_{a,k}$ of the k-*th* accent command. Then the problem is

$$\mathbf{a}^* = arg \min_{\mathbf{a}} \, \varepsilon . \tag{9}$$

Consider the truncated version with M samples of the infinite exponentially decaying impulse response of the accent control mechanism, and let

$$\mathbf{H}^T = \begin{pmatrix} h_a(0) & 0 & \ldots & 0 \\ h_a(1) & h_a(0) & \ldots & 0 \\ h_a(2) & h_a(1) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ h_a(M-1) & h_a(M-2) & \ldots & 0 \\ 0 & h_a(M-1) & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & h_a(0) \end{pmatrix} , \tag{10}$$

then

$$\mathbf{y}_a = \mathbf{H}^T \mathbf{x}_a , \tag{11}$$

and

$$\begin{aligned} \min_{\mathbf{a}} \varepsilon &= \min_{\mathbf{y}_a} \{ \mathbf{y}_a^T \mathbf{y}_a - 2\mathbf{d}^T \mathbf{y}_a \} \\ &= \min_{\mathbf{x}_a} \{ \mathbf{x}_a^T \mathbf{H}\mathbf{H}^T \mathbf{x}_a - 2\mathbf{d}^T \mathbf{H}^T \mathbf{x}_a \} . \end{aligned} \tag{12}$$

The solution of the uncostrained problem

$$\mathbf{x}^*_a = (\mathbf{H}\mathbf{H}^T)^{-1} \mathbf{H}\mathbf{d} , \tag{13}$$

is not suited to the problem because $x^*_a(n)$ is costrained to be a sequence of rectangular pulses located in correspondence of the accent commands of $y(n)$. To take into account for this constraint, it is useful to consider the following expression

$$\mathbf{x}_a = \begin{pmatrix} \mathbf{z}_1 \\ A_{a,1}\mathbf{o}_1 \\ \mathbf{z}_2 \\ A_{a,2}\mathbf{o}_2 \\ \vdots \\ \mathbf{z}_{N_a} \\ A_{a,N_a}\mathbf{o}_{N_a} \end{pmatrix} , \tag{14}$$

where $\mathbf{z}_k$ (resp. $\mathbf{o}_k$) is a vector of zeros (resp. ones) whose length L_{0k} (resp. L_k)is equal to the number of samples of $y(n)$ between $n''_{a,k-1}$ and $n'_{a,k}$ (resp. $n'_{a,k}$ and $n''_{a,k}$). Therefore $\mathbf{H}$ can be divided into $2 * N_a$ submatrices

$$\mathbf{H}^T = \begin{pmatrix} \mathbf{Z}_1 & \mathbf{H}_1 & \mathbf{Z}_2 & \mathbf{H}_2 & \ldots & \mathbf{Z}_{N_a} & \mathbf{H}_{N_a} \end{pmatrix} , \tag{15}$$

where $\mathbf{Z}_k$ is a matrix $L \times L_{0k}$ and $\mathbf{H}_k$ is a matrix $L \times L_k$.

From Eqs. (11),(14) and (15) it follows that

$$\mathbf{y}_a = \sum_{k=1}^{N_a} A_{a,k} \mathbf{H}_k \mathbf{o}_k , \tag{16}$$

and therefore the solution to the constrained problem is

$$\mathbf{a}^* = (\mathbf{P}\mathbf{P}^T)^{-1} \mathbf{P}\mathbf{d} , \tag{17}$$

where

$$\mathbf{P}^T = \begin{pmatrix} \mathbf{p}_1 & \mathbf{p}_2 & \ldots & \mathbf{p}_{N_a} \end{pmatrix} , \tag{18}$$

and

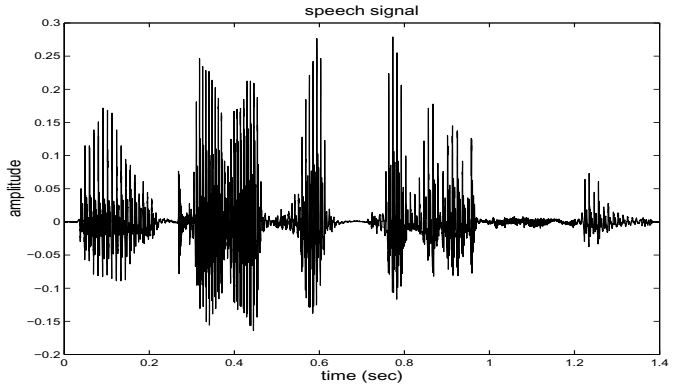
$$\mathbf{p}_k = \mathbf{H}_k \mathbf{o}_k . \tag{19}$$

3.3 Experiments

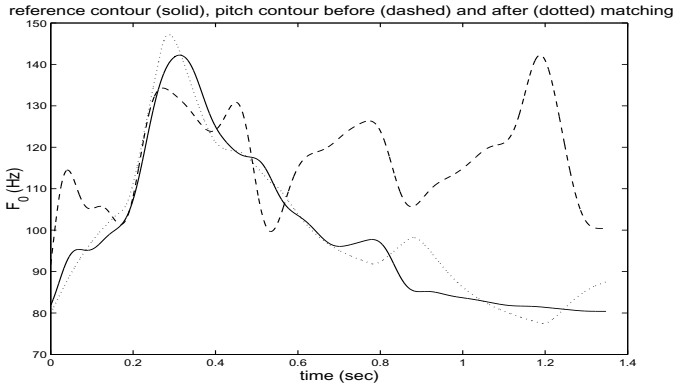
A software based on the described algorithm has been implemented in MATLAB and tested on a corpus of 50 utterances of continuous Italian speech partially chosen from the corpus CLIPS [11]. The results were encouraging. From a perceptive point of view the synthesized speech presents the desired intonation characteristics on 50 percent of cases, but it never resulted in an unnutural-sounding speech. Therefore we can say that natural-soundness is preserved by the model.

Fig. 5 shows an example of an interrogative utterance processed to become declarative by using a pitch contour obtained by another reference declarative utterance. It can be observed that there is a close matching between the reference pitch contour and the pitch contour obtained by using the proposed method. Synthesis results are quite natural sounding.

Some improvements are still to be implemented to allow greater variability as large mismatches tend to cause critical problems in PSOLA synthesis [4].



(a) Speech signal to be processed. The utterance corresponds to "E' ancora troppo presto?" (en. "Is it still too much soon?").



(b) Originary interrogative pitch contour (dashed line), reference declarative pitch contour (solid line) and matched declarative pitch contour (dotted line).

Figure 5: Example of matching the intonation of an utterance to a reference one preserving natural-soundness of synthesized speech.

4. CONCLUSIONS

The paper describes a simple method for automatic matching of intonation characteristics of speech. Basing on the Fujisaki's model the pitch contour extracted from on utterance is matched to a reference contour. The Fujisaki's model is used to introduce appropriate constraints for the matching problem in order to preserve natural-soundness of the synthesized speech.

The proposed technique has been used to realize a speech synthesizer that confirms the effectiveness of Fujisaki's model in speech synthesis. Our experiments have been based on manipulation of the magnitude of accent commands. They have provided very natural-sounding synthesized speech. Future works include the extension of the synthesizer so that manipulation of commands timing and duration is allowed.

The results in this paper represent an important step toward the implementation of a totally automatic analysis/synthesis based on clusters learned from features analysis. We are currently considering the extension of similar modeling to other prosodic parameters such as duration and energy profiles.

REFERENCES

[1] D.H. Klatt, *Software for a cascade/parallel formant synthesizer*. Journal of Speech and Hearing Research, pp. 287–299, 1980.

[2] H. Fujisaki, *Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing*. The Production of Speech, P.F. MacNeilage (ed.),Springer-Verlag New York Heidelberg Berlin, pp. 39–47, 1983.

[3] A.V. Oppenheim and R.W. Schafer, *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice Hall, 1989.

[4] E. Moulines and F. Charpentier, *Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones*. Speech Communication, Vol. 9, pp. 453–467, 1990.

[5] Y. Medan, E. Yair and D. Chazan, *Super Resolution Pitch Determination of Speech Signals*. IEEE Transaction on Signal Processing, pp. 40–48, 1993.

[6] H. Fujisaki, M. Ljungqvist and H. Murata, *Analysis and Modelling of Word Accent and Sentence Intonation in Swedish*. IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 2, pp. 211–214, 1993.

[7] A. Sakurai, and H. Hirose, *Detection of Phrase Boundaries in Japanese by Low-Pass Filtering of Fundamental Frequency Contours*. Fourth International Conference on Spoken Language, Vol. 2, pp. 817–820, 1996.

[8] H. Fujisaki, and S. Ohno, *The Use of a Generative Model of F_0 Contours for Multilingual Speech Synthesis*. Fourth International Conference on Signal Processing, Vol. 1, pp. 714–717, 1998.

[9] H. Mixdorff, *A Novel Approach to the Fully Automatic Extraction of Fujisaki Model Parameters*. IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 3, pp. 1281–1284, 2000.

[10] J. M. Gutierrez-Arriola, J. M. Montero, D. Saiz and J. M. Pardo, *New Rule-Based and Data-Driven Strategy to Incorporate Fujisaki's F_0 Model to a Text-to-Speech System in Castillian Spanish*. IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 2, pp. 821–824, 2001.

[11] CLIPS (Corpora di Lingua Italiana Parlata e Scritta) working process. (ref. F. Albano Leoni, *Tre Progetti per l'Italiano Parlato*. VI Convegno Internazionale della Societ Internazionale di Linguistica e Filologia Italiana, 2000).

[12] S. Narusawa, N. Minematsu, K. Hirose, and H. Fujisaki, *A Method for Automatic Extraction of Model Parameters from Fundamental Frequency Contours of Speech*. IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 1, pp. 509–512, 2002.

[13] P. Salvo Rossi, F. Palmieri, and F. Cutugno, *Inversion of F_0 Model for Natural-Sounding Speech Synthesis*. IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 1, pp. 520–523, 2003.