# INVERSION OF $F_0$ MODEL FOR NATURAL-SOUNDING SPEECH SYNTHESIS

*Pierluigi Salvo Rossi(1,4), Francesco Palmieri(2,4), Francesco Cutugno(3,4)*

(1) Dipartimento di Informatica e Sistemistica, Università di Napoli "Federico II"
(2) Dipartimento di Ingegneria dell'Informazione, Seconda Università di Napoli
(3) Dipartimento di Scienze Fisiche, Università di Napoli "Federico II"
(4) C.I.R.A.S.S., Università di Napoli "Federico II"

**ABSTRACT**

Natural-sounding speech synthesizers requires the information from a model quantitatively describing prosody. Fujisaki's model [1] has shown considerable accuracy on many languages [4][6]. We propose a method for Fujisaki's model parameters estimation, i.e. an inversion methods, based on relative extremes of pitch contour and a gradient algorithm refinement procedure. Preliminary results show excellent performance of the proposed method in matching the pitch contours. Preliminary results of synthesis making use of obtained features are surely encouraging.

## 1. FUJISAKI'S MODEL

Though actual synthesizers provide good segmental quality speech and synthesis techniques as PSOLA [2] are able to modify its prosodic characteristics, natural-sounding speech syntesis is a very complicate task. This is due to the difficulty of integrating a prosody model.

We focus on analysis of pitch contours as intonation is an acceptable description of prosody (a more accurate description should include duration and intensity). Fujisaki's model has shown a remarkable effectiveness in describing pitch contours. It captures the essential mechanisms involved in the speech production that confers to it a particular prosodic structure.

H. Fujisaki and his co-workers proposed, between the 70s and the 80s, an analytical model describing the fundamental frequency ($F_0$) variations [1]. The model, tested on many languages [4][6], assumes that the $F_0$ contour (in a logarithmic scale) is the superposition of two contributions, namely 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 by Dirac impulses, namely *phrase commands*, located at the onsets of phrase activities. The second input signal ($x_a$) 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 as follows: $h_p(t) = \alpha^2 t e^{-\alpha t} u(t)$ , is the impulse response of the phrase control mechanism, where $\alpha \in [2, 4]\ s^{-1}$ is its natural angular frequency, and $g_a(t) = [1 - (1 + \beta t)e^{-\beta t}]u(t)$ , is the step response of the accent control mechanism, where $\beta \in [19, 21]\ s^{-1}$ is its natural angular frequency.

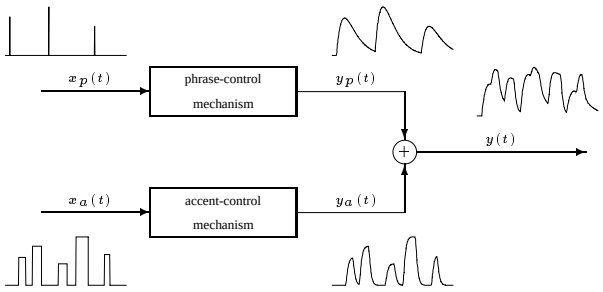



**Fig. 1**. Fujisaki's model.

The total pitch contour is expressed then as

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

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 on-

set and the end of the k-*th* accent command. A non-linear system, accounting for possible glottal effects, has been ignored as it is irrelevant to our study.

## 2. MODEL INVERSION

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).

Aim of this work is to propose a method for automatic extraction of Fujisaki's model features (phrase and accent commands) from a given $F_0$ contour. It has been tested on Italian sentences. Fujisaki's model output to the extracted features must optimally (in terms of the mean square error) match the given contour.

Since Fujisaki's model generates continuous curves, it is useful to ignore the unvoiced portions of the $F_0$ contour interpolating them with voiced ones. Moreover it appears appropriate to filter the continuous $F_0$ contour with a LPF to remove quick and small fluctuation to be considered as noise terms for the model.

Some authors [8][5][7] approached the inverse problem by differentiating or filtering $F_0$ contour with low-pass or high-pass filters to split the two contributions.

The proposed method executes this split operation by miming the hand-made inversion by a phonetist. Parameters $\alpha$ and $\beta$ are considered to be constant for simplicity. The proposed method consists of two steps:

- an estimation algorithm based on relative extremes of the pitch contour
- a refinement procedure based on a gradient optimization algorithm

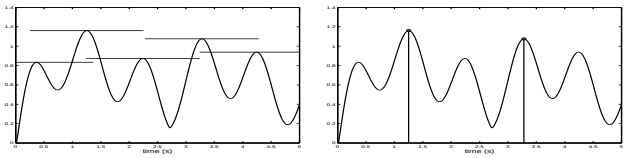
Before describing the estimation algorithm it is useful to give the following definition. Let us consider a continuous function $f(t)$ defined $\forall t \in [\tau_1, \tau_2]$ , a point $t_0 \in (\tau_1, \tau_2)$ will be named a *$T$-dominant maximum point* of $f(t)$ if the two following conditions are verified:

- $t_0$ is a relative maximum point of $f(t)$
- $f(t_0) \geq f(t) \quad \forall t \in [t_0 - T/2, t_0 + T/2]$

This definition will be used to locate contributions of every phrase command and accent command. Fig.2 shows a continuous function and its $2\ s.$-dominant maximum points.

Fujisaki's model is based on the two linear systems characterized by exponentially-shaped impulse responses shown in Fig.3.

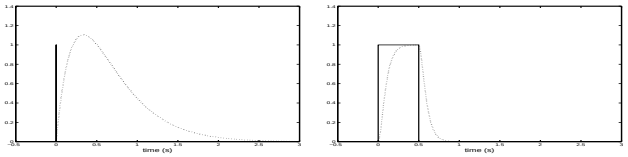
The inverse problem can be thought as the decomposition of a pitch contour onto basis functions with shapes that look like the dotted lines shown in Fig.3. The contribution of a phrase command manifests itself on a time



(a) Function, its maxima and its $2s.$-windows centered on the maxima.

(b) Function and its $2s.$-dominant maximum points.

**Fig. 2**. Example of T-dominant maximum points of a function.



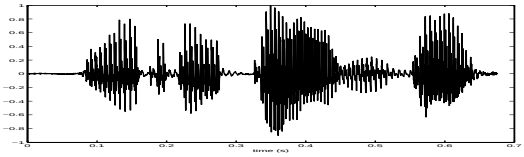
(a) Phrase command (solid) and its contribution to the pitch contour (dotted).

(b) Accent command (solid) and its contribution to the pitch contour (dotted).

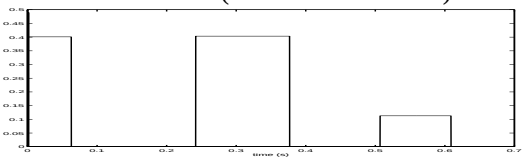
**Fig. 3**. Phrase and accent commands (solid) and their exponential-shaped contributions to the pitch contour.

interval greater than the one related to accent commands. The considerations above suggest that a $T_p$-dominant maximum point, where $T_p$ is appropriately chosen on the basis of the time constant $1/\alpha$, may be used to reveal the presence of a phrase command. Fig.3(a) shows where a phrase command is located respect to its contribution, it can be seen that likely the phrase command is located in correspondence of a local minimum of the pitch contour. Searching for phrase commands the method operates as follows. Let $\{t_{DP,k}\}_{k=1}^{N_p}$ be the set of $T_p$-dominant maximum points of pitch contour, to locate the k-*th* phrase command the algorithm chooses the minimum point of pitch contour included in the interval $[t_{DP,k-1}, t_{DP,k}]$ where $t_{DP,0}$ is the beginning time of the curve. Magnitudes of phrase commands are recursively chosen by comparing their contribute generated by Fujisaki's model with pitch contour, remembering that the former cannot exceed the latter (this assumption was also assumed in [9] ). Choice of $T_p$ must be accurate, its role in the procedure is somehow like the role that the inverse of the cut-off frequency plays in low-pass filtering.

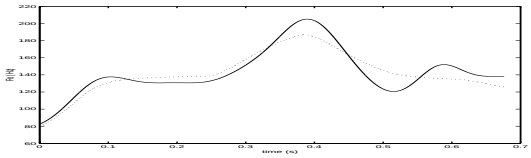
Once phrase commands have been estimated, their contributes are subtracted from global pitch contour; the resulting curve must then be decomposed to identify accent commands. Likewise phrase-commands searching, the presence of a $T_a$-dominant maximum points in the resulting curve, where $T_a$ is appropriately chosen on the basis of the time constant $1/\beta$, may be used to reveal the presence of an accent command. Fig.3(b) shows where an accent command is located respect to its contribution, it can be seen that likely



(a) Speech signal, the waveform corresponds to: "E' rotondo?" (En. "Is it rounded?").

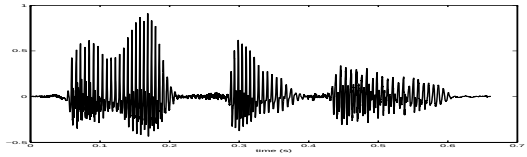


(b) Estimated phrase and accent commands.

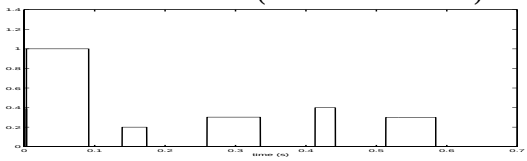


(c) $F_0$ contour (solid) and estimated one (dotted).

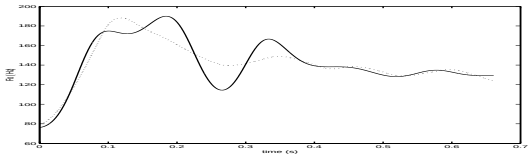
**Fig. 4**. Example of pitch analysis based on Fujisaki's model



(a) Speech signal, the waveform corresponds to: "Stammi a sentire!" (En. "Listen to me!").



(b) Estimated phrase and accent commands.



(c) $F_0$ contour (solid) and estimated one (dotted).

**Fig. 5**. Example of pitch analysis based on Fujisaki's model

the onset of the accent command is located in correspondence of a local minimum of the resulting curve, while the end of the accent command is located in correspondence of a local maximum of the resulting curve. Let $\{t_{DA,k}\}_{k=1}^{N_a}$ be the set of $T_a$-dominant maximum points of resulting curve, to locate the onset of the k-*th* accent command the algorithm chooses the minimum point of the resulting curve included in the interval $[t_{DA,k-1}, t_{DA,k}]$ where $t_{DA,0}$ is the beginning time of the curve. To locate the end of the k-*th* accent command the algorithm chooses just $t_{DA,k}$. Magnitudes of accent commands are recursively chosen by comparing their contribute generated by Fujisaki's model with the resulting curve.

$T_p$ and $T_a$ was heuristically chosen ($T_p \approx 1\ s$ and $T_a \approx 10\ ms$) according to the different role that phrase commands and accent commands play.

After the first step, in which the pitch contour is parameterized by timing and magnitudes of commands, estimations of onset time, end time and magnitude of accent commands are improved by a gradient-based procedure.

Let $r(t) = y(t) - \sum_{k=1}^{N_p} A_{p,k}\ h_p(t - t_{p,k})$ , be the residual curve of the pitch contour after the estimation of the phrase commands, its initial estimation is

$$\hat{r}(t) = \sum_{k=1}^{N_a} A_{a,k}\ [g_a(t - t^{'}_{a,k}) - g_a(t - t^{''}_{a,k})]\ , \quad (2)$$

Aim of the procedure is to minimize the cost function $\mathcal{R} = \int_0^T [r(t) - \hat{r}(t)]^2(t)dt$ , where T is the total duration of the pitch contour. By defining the parameters vector $\underline{p}^T = [A_{a,1}, A_{a,2}, \ldots, A_{a,N_a}, t^{'}_{a,1}, t^{'}_{a,2}, \ldots, t^{'}_{a,N_a}, t^{''}_{a,1}, t^{''}_{a,2}, \ldots, t^{''}_{a,N_a}]$ the gradient algorithm updates this vector, to minimize $\mathcal{R}$, as $\underline{p}(n+1) = \underline{p}(n) + diag(\underline{\mu}) \cdot \nabla_{\underline{p}}\mathcal{R}(\underline{p})\Big|_{\underline{p}=\underline{p}(n)}$ where $\underline{\mu}$ must be chosen so that the convergence of the algorithm is guaranteed. The components of the vector $\nabla_{\underline{p}}\{\mathcal{R}\}$ are the following partial derivatives

$$\frac{\partial \mathcal{R}}{\partial A_{a,k}} = -2 \int_0^T e(t)\ [g_a(t - t^{'}_{a,k}) - g_a(t - t^{''}_{a,k})]\ dt\ , \quad (3)$$

$$\frac{\partial \mathcal{R}}{\partial t^{'}_{a,k}} = 2A_{a,k} \int_0^{T - t^{'}_{a,k}} e(t + t^{'}_{a,k})\ \beta^2 t \exp(-\beta t)\ dt\ , \quad (4)$$

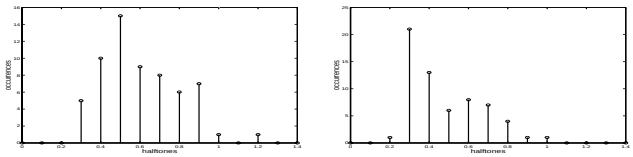
$$\frac{\partial \mathcal{R}}{\partial t^{''}_{a,k}} = -2A_{a,k} \int_0^{T - t^{''}_{a,k}} e(t + t^{''}_{a,k})\ \beta^2 t \exp(-\beta t)\ dt\ . \quad (5)$$

The exact computation of these formulas is here omitted for brevity.

Experimental evidence showed the cost function $\mathcal{R}$ be well-behaved with respect to parameters $\{p_k\}_{k=1}^{N_p}$, and significant improvements are obtained after the algorithm is applied.

## 3. RESULTS

Even though the discussion has been developed in reference to continuous time domain signals, our experiments are run on sampled signals and using digital filters. The digital filters we use to simulate the phrase-control and the accent-control mechanisms are designed using respectively the pulse-invariance and the step-invariance techniques. This



(a) Mean error. (b) Error standard deviation.

**Fig. 6**. Experimental results of pitch stylization.

has been a natural choice for the type of the input sequences assumed in the model.

The system we used to extract contours to be analyzed produces a continuous $F_0$ profile by first estimating the pitch (period by period), assigning a fixed period value to the unvoiced portions, then smoothing the curve with an interpolator followed by a low-pass filter. Before low-pass filtering, a non linear block is inserted just to change the scale to logarithmic and subtract the constant value $ln(F_{min})$.

The PDA (*Pitch Determination Algorithm*) is based on auto-correlation and follows the computation of correlation between consecutive variable-length windows of speech [3]. In the second block, isolated values of $F_0$ are corrected by using a median filter. Then a linear interpolation and a Butterworth low-pass filtering at $10\ Hz$ of order $5$ smooth the contour.

A software based on the described algorithm was realized in Matlab and tested on $62$ utterances of continuous Italian speech chosen from the corpus CLIPS [10]. Figs.45 show an example of pitch stylization based on Fujisaki's model.

The results of the tests showed the absolute error between the extracted pitch contour and the stylized one to be less than $2$ halftones.

## 4. CONCLUSIONS

The paper describes a simple method for automatic extraction of Fujisaki's model features from speech signals, it has been tested on a corpus of Italian continuous speech, giving excellent results.

The proposed technique has been used to realize a speech synthesizer that confirms the effectiveness of Fujisaki's model information in speech synthesis. Our experiments were based on manipulation of the magnitude of phrase and accent commands, and furnished very natural-sounding synthesized speech. We plan to extend our synthesizer to include manipulation of commands timing and duration with the objective of increasing our degrees of freedom in modifying significantly the intonation of an utterance without generating unnatural-sounding synthesized speech.

Future works will be a totally automatic analysis/synthesis based on clusters learned from features analysis. Moreover it will be appropriate to extend similar modeling to other prosodic parameters such as duration and energy profiles.

## 5. REFERENCES


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

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

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

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

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

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

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

[8] S. Narusawa, H. Fujisaki, and S. Ohno, "A Method for Automatic Extraction of Parameters of the Fundamental Frequency Contours," in *Sixth International Conference on Spoken Language Processing*, 2000.

[9] 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," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 821–824, 2001.

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