Speech Prosody 2002
Aix-en-Provence, France
April 11-13, 2002

A Method for Automatic Extraction of Fujisaki-Model Parameters

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

(1) Facoltà di Ingegneria, Università degli Studi di Napoli "Federico II"
(2) Dipartimento di Ingegneria dell'Informazione, Seconda Università di Napoli
(3) C.I.R.A.S.S., Università degli Studi di Napoli "Federico II"
pierluigi@cirass.unina.it; {frapalmi,cutugno}@unina.it

Abstract

The utility of a model describing pitch profiles in speech signals is of fundamental importance in many application areas and especially in natural-sounding text-to-speech system. Fujisaki-model [1] has shown considerable accuracy on many languages, despite its simplicity. The inverse problem, i.e. the extraction of the input parameters which generated an observed pitch contour, that could be of great interest in the field of automatic extraction of prosodic parameters from a given speech signal, is a much harder task. This paper suggests a method for input parameters estimation based on two steps: an initial guessing algorithm based on relative extremes, and a refinement procedure based on a gradient optimization algorithm. Preliminary results of analysis/synthesis of pitch contours show excellent performance of the proposed method.

1. Introduction

One of the main difficulties in speech synthesis is to generate signals that have natural-sounding prosody.
Even though prosody is attributed to intensity, intonation and duration profiles, it seems that in most languages intonation is the main contribute to prosodic information. In analysing a segment of recorded speech, intonation can be rather accurately described by the estimated pitch contour.
A model describing pitch contours would then be of extreme importance in speech processing.
The Fujisaki-model [1] has shown, on tests performed on various languages [2][3][4][5] (but still more work needs to be done for Italian), a remarkable effectiveness in accounting for phrase and accent control events. This is a generative model (*Fig. 1*) that constitutes the best currently known framework to study a prosodic profile.
Even though the Fujisaki-model has been verified on hand-labelled speech, an automatic procedure for inverting it, i.e. extracting the prosodic events from recordered speech seems still to be lacking.
In this paper, after a short review of the model, we propose an inversion algorithm. The procedure starts from pitch profiles extracted from recorded speech and detects phrase and accent control excitation signals, effectively decomposing the speech profile into "elementary prosodic elements." The effectiveness of the method is tested on a number of phrases.

2. The Fujisaki-model

H. Fujisaki and his co-workers proposed, between the 70s and the 80s, an analytical model for the control of the fundamental frequency (F_0) variations [1][2]. This model, successfully tested on many languages [2][3][4][5], describes an F_0 contour (plotted in a logarithmic scale) as the superposition of two contributes obtained by filtering two signals, here named x_1 and x_2, with two linear systems. The former, which models the baseline component, accounts for speaker declination. The latter, which models micro-prosodic variations, accounts for accent components. The Fujisaki-model is depicted in the block-diagram of *Fig. 1*. A constant term referring to the minimum frequency of the speaker is ignored. *Fig. 2* shows an example of a generic F_0-contour generated by the model.

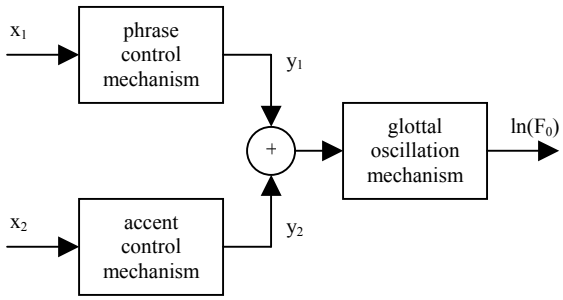

Fig. 1: Fujisaki-model.

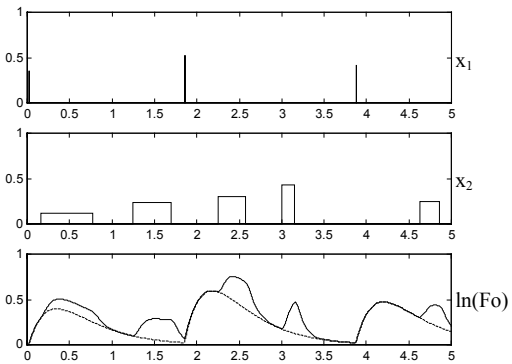

Fig. 2: Example of a generic F_0 contour generated by Fujisaki-model, with the superposition of 3 phrase commands and 5 accent commands.

The symbols in *Figs. 1* and *2* are:

x_1- phrase-command input, composed by Dirac-impulses;
x_2- accent-command input, composed by rectangle-shaped pulses;
y_1- phrase-command contribute to $\ln(F_0)$;
y_2- accent-command contribute to $\ln(F_0)$.

The system is ruled by the following equations:

$$\ln(F_0) = \ln(F_b) + \sum_{k=1}^{N_p} A_{p,k} \cdot g_p(t - T_{p,k}) + \quad (1)$$

$$+ \sum_{k=1}^{N_a} A_{a,k} \cdot \left[g_a(t - T'_{a,k}) - g_a(t - T''_{a,k}) \right]$$

where

$$g_p(t) = \alpha^2 t \cdot \exp(-\alpha t) \qquad t \geq 0 \qquad (2)$$

and

$$g_a(t) = 1 - (1 + \beta t) \cdot \exp(-\beta t) \qquad t \geq 0 \qquad (3)$$

respectively indicate the impulse response function of the phrase-control mechanism and the step response function of the accent-control mechanism.

The symbols in Eqs. (1), (2) and (3) indicate:

F_b - asymptotic value of the fundamental frequency in absence of accent-command;
N_p - number of phrase-commands;
N_a: number of accent-commands;
$A_{p,k}$ - magnitude of the *k*th phrase-command
$A_{a,k}$ - magnitude of the *k*th accent-command
$T_{p,k}$ - timing of the *k*th phrase-command
$T'_{a,k}$ - onset of the *k*th accent-command
$T''_{a,k}$ - end of the *k*th accent-command;
α - natural angular frequency of the phrase control mechanism to the phrase-commands;
β - natural angular frequency of the accent control mechanism to the accent-commands.

3. Fujisaki-model inversion

This work aims to propose an algorithm that can optimally reconstruct phrase and accent command signals for a given F_0 contour. The optimality criterion is invoked as an analysis-by-synthesis procedure: the estimated input of the Fujisaki-model is recursively fed back to the model to provide a better match to the measured contour profile.
Since the Fujisaki-model produces continuous curves a first step is to ignore the unvoiced portions of the given F_0 contour and interpolate the curve in these portions with voiced ones. It also appears appropriate to filter the obtained continuous curve with a low-pass filter just to remove high-frequency noise contribution and quick and small pitch fluctuations.
The idea on which the Fujisaki-model is based, is that phrase commands account for slow pitch variations while accent commands accounts for rapid ones. Some authors have approached the problem by differentiating or filtering F_0 contour with low-pass and high-pass filters to split the two components [6][7][8]. However, this approach has been shown to be more difficult than it seems: pitch contour spectrum is relevant only at very low frequencies, and the two contributions are not disjoint.
The proposed algorithm executes a sort of filtering in a different way. Parameters α and β are considered to be constant for sake of simplicity and their values are fixed respectively to 3 and 20, as suggested by Fujisaki's original model [1]. The glottal oscillation mechanism will be ignored.
Fig. 3 shows an example of pitch contour, after interpolation and low-pass filtering, to be analysed for model inversion.
A human observer will recognise the presence of just one or two phrase commands, giving a fast rise and a slow fall to the pitch contours showed in the figure. Phrase commands of pitch contour in *Fig. 3c* will be timed at $t \approx 0$ and eventually at $t \approx 0.6$, phrase commands of pitch contour in *Fig. 3d* will be timed at $t \approx 0$ and eventually at $t \approx 0.8$. Other fluctuations will correspond to accent commands.

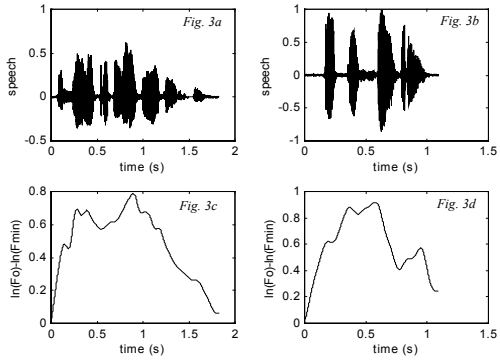

Fig. 3: Examples of interpolated and low-pass filtered pitch contour.

The initial estimation is made by the guessing algorithm which works similarly to an observer looking for significative events on pitch contours. A similar approach was assumed in [9] for speech-energy contour processing aiming at syllable parsing.

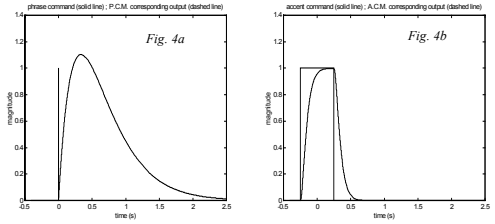

Fig. 4: Example of phrase (Fig. 4.a) and accent (Fig. 4.b) commands and their contributes to pitch contour.

Fig. 4.a (respectively *Fig. 4.b*) shows an example of phrase command (accent command) and the corresponding output to the phrase control mechanism (accent control mechanism).
Pitch contour has to be broken up into functions showed in *Fig. 4* with dashed line, the latter contibute is superimposed on the former one.

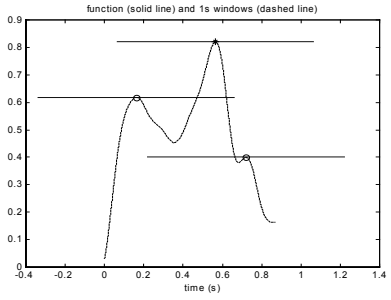

Fig. 5: Example of 1s dominant-maximum point of a curve (timing of symbol '') and relative maximum points which are not dominant (timing of symbols 'o').*

First the algorithm searches for phrase commands. From now on a point t_0 is called a *T dominant-maximum point* of the function *f(t)* if the two following conditions are verified:

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

Searching for phrase commands, the exponential shape of phrase-commands contributes to pitch curve suggests to operate as follows. Let $\{t_{D,k}\}_{k=1}^{N}$ 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 $[t_{D,k-1}; t_{D,k}]$, where $t_{D,0}$ is the beginning time of the curve. Magnitudes of phrase commands are recursively chosen by comparing contribute of phrase commands generated by the Fujisaki-model with pitch contour, remembering that the former cannot exceed the latter (this assumption was assumed by J.M. Gutierrez-Arriola et al. [10] too). Choice of T_p must be accurate, for the role this parameter plays in the procedure is somehow like the one that the inverse of 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 be described by accent commands. To locate onset and end of each accent command look at the *Fig. 4.b*. It shows a rectangle-shaped pulse (*solid line*) and the corresponding output to the accent control mechanism (*dashed line*).
The onset of the command corresponds to the point in which the output starts rising, therefore the resulting curve shows most likely a minimum point. The end of the command corresponds to the point in which the output starts falling, therefore the resulting curve shows a maximum point. Based on these remarks the algorithm searches for T_a dominant maximum points of the resulting curve, locating the end of an accent command on each one of them. The corresponding onset will be located on the first previous relative minimum point. Appropriate values for T_p and T_a was suggested by the different role that phrase commands and accent commands play; T_p order of magnitude is 1 sec. , T_a order of magnitude is 10 ms. Magnitudes of accent commands are chosen by comparing contribute of accent commands generated by the Fujisaki-model with the resulting curve.
Onset, end and magnitude estimation of accent commands could be improved by a gradient-based procedure.
Let r(t) be the residual curve of a pitch contour. Its initial estimation is:

$$\hat{r}(t) = \sum_{k=1}^{N_a} A_{a,k} \cdot \xi_a(t; T'_{a,k}; T''_{a,k}) \tag{4}$$

where

$$\xi_a(t; T'; T'') = g_a(t - T') - g_a(t - T'') \tag{5}$$

Aim of the procedure is to minimize the cost function:

$$\Re = \int e^2(t) \cdot dt \tag{6}$$

where

$$e(t) = r(t) - \hat{r}(t) \tag{7}$$

Let $\underline{p}$ be the parameter vector so defined:

$$\begin{aligned} \underline{p} &= \left[p_1, \ldots, p_{N_p}\right] = \\ &= \left[A_{a,1}, \ldots, A_{a,N_a}, T'_{a,1}, \ldots, T'_{a,N_a}, T''_{a,1}, \ldots, T''_{a,N_a}\right] \end{aligned} \tag{8}$$

The vector is updated by the following rule:

$$\underline{p}(n+1) = \underline{p}(n) - \underline{\mu}^T \cdot \overline{\nabla}_{\underline{p}} \left\{\Re(\underline{p})\right\}|_{\underline{p}=\underline{p}(n)} \tag{9}$$

The partial derivative of $\Re$ are:

$$\begin{aligned} \frac{\partial \Re}{\partial A_{a,k}} &= -2 \cdot \int e(t) \cdot \xi_a(t; T'_{a,k}; T''_{a,k}) \cdot dt \\ \frac{\partial \Re}{\partial T'_{a,k}} &= -2 A_{a,k} \cdot \int e(t) \cdot \frac{\partial \xi_a(t; T'_{a,k}; T''_{a,k})}{\partial T'_{a,k}} \cdot dt \\ \frac{\partial \Re}{\partial T''_{a,k}} &= -2 A_{a,k} \cdot \int e(t) \cdot \frac{\partial \xi_a(t; T'_{a,k}; T''_{a,k})}{\partial T''_{a,k}} \cdot dt \end{aligned} \tag{10}$$

The cost function $\Re$ is well-behaved with respect to the parameters $\{A_{a,k}\}_{k=1}^{N_a}$ and from experimental evidence it appears to be poorly convex with respect to the parameters $\{T'_{a,k}\}_{k=1}^{N_a}$ and $\{T''_{a,k}\}_{k=1}^{N_a}$.

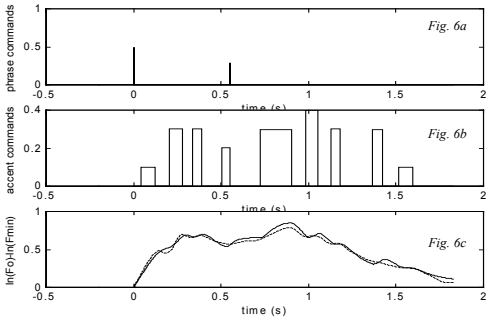

Fig. 6: Example of Fujisaki-model pitch stylization (solid line) for a given pitch contour (dashed line).

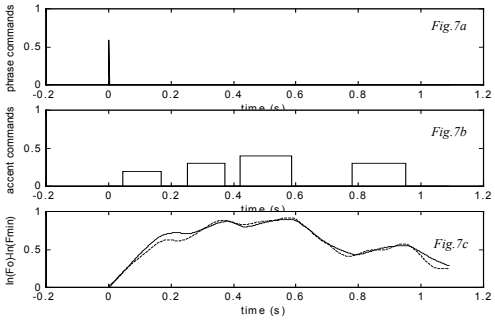

Fig. 7: Example of Fujisaki-model pitch stylization (solid line) for a given pitch contour (dashed line).

The results we report hold for an algorithm that refines only the magnitudes since we have found too critical to choose the appropriate step-size for time-parameters.
Figs. 6 and *7* show the parametric stylization of pitch contours plotted respectively in *Figs. 3c* and *3d*; in particular *Figs. 6a* and *7a* show the phrase component, *Figs. 6b* and *7b* show the accent component, *Figs. 6c* and *7c* show the pitch contour to be stylized (dashed line) and the curve given by the Fujisaki-model stylization (solid line). The utterances are

respectively “Abbiamo preparato una torta.” and “Sta più sopra?”, uttered by male speakers.

4. Results

A software based on the previously described algorithm was realized and tested on 30 utterances of spontaneous speech uttered by an Italian male speaker, chosen in the *corpus A.V.I.P.* [11].
Fig. 8 shows histograms of mean absolute error (*Fig. 8a*) and absolute error standard deviation (*Fig. 8b*) given by the tests. Pitch contour extracted from speech and Fujisaki-model stylization are compared in a halftones scale.

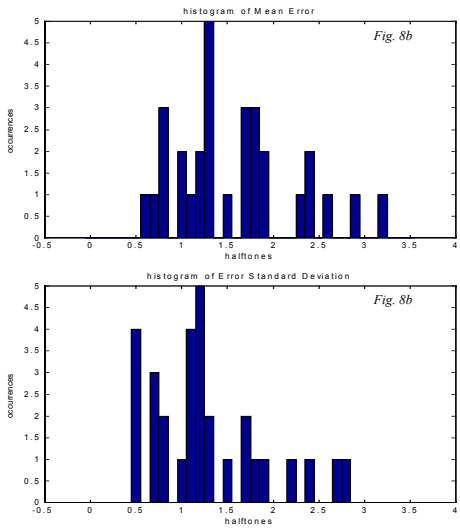

Fig. 8: Experiments on italian utterances. Histograms of Mean Error (Fig. 8a) and Error Standard Deviation (Fig. 8b).

A first analysis of the results showed that estimated phrase commands seem to locate linguistic structures as tone units or phrases, while estimated accent commands seem to correspond to syllables.

5. Conclusions

Researches for modeling Italian intonation using the Fujisaki-model are too few to compare results with manual estimation. In particular the investigation of which could be the real linguistic domain of the phrase and accent commands derived by our analysis is still missing. Future researches should clarify underlying relations between model parameters and Italian linguistic structures. This will be done by means of the comparison between prosodically hand-labelled data and output sequences of the realized program. The challenge to win is to “translate” former data (typically formatted in ToBI or other more or less similar labelling systems) in order to be comparable with the latter ones.
Despite of its initial state of evolution, the present algorithm provides encouraging results. The guessing portion of the extraction procedure indicates that, within the speech signal, a sort of “prosodic redundancy” can be extracted and used in order to mark events presenting particular significance. Even if this approach is heuristically based, its implication for a theory of categorization for the prosodic events is clear-cut. As mentioned above, here we used an approach similar to the one applied in [9] where a procedure was implemented for automatic segmentation of speech into syllabic units. Also in this case a relatively simple deterministic procedure, strongly based on the properties related to the sequence of maxima and minima within the temporal pattern of energy in the speech signal, was devised (this procedure is almost different from the one presented in [8]). The result of our work confirms that, in connected speech, prosodic features appear to be more stable and present than the segmental ones (see also [12]).

6. References

[1] Fujisaki, H., 1983. 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, 39-47.

[2] Fujisaki, H.; Ohno, S., 1998. The use of a generative model of F_0 contours for miltilingual speech synthesis. In *Proceedings of the Fourth International Conference on Signal Processing*. 714-717 vol. 1.

[3] Tams, A.; Tatham, M., 2000. Intonation for synthesis of speaking styles. *IEE Seminar on State of the Art in Speech Synthesis (Ref. No.2000/058)*. 6/1-6/11.

[4] Navas, E.; Hernaez, I.; Etxebarria, B.; Salaberria, J., 2000. Modelling Basque intonation using Fujisaki’s model and carts. *IEE Seminar on State of the Art in Speech Synthesis (Ref. No.2000/058)*. 3/1-3/6.

[5] Fujisaki, H.; Ljungqvist, M.; Murata, H., 1993. Analysis and modeling of word accent and sentence intonation in Swedish. *ICASSP-93 IEEE International Conference on Acoustics, Speech, and Signal Processing*. 211-214 vol. 2.

[6] Sakurai, A.; Hirose, K., 1996. Detection of phrase boundaries in Japanese by low-pass filtering of fundamental frequency contours. In *Proceedings of the Fourth International Conference on Spoken Language*. 817-820 vol. 2.

[7] Mixdorff, H., 2000. A novel approach to the fully automatic extraction of Fujisaki model parameters. In *Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing*. 1281-1284 vol. 3.

[8] Narusawa, S.; Fujisaki, H.; Ohno, S., 2000. A method for automatic extraction of parameters of the fundamental frequency contours. In *Proceedings of the 6th International Conference on Spoken Language Processing*.

[9] Petrillo, M., 2000. Sillabificazione dei segnali vocali: un approccio procedurale. In *AIA XXVIII Convegno Nazionale, atti*. 303-306.

[10] Gutierrez-Arriola, J.M.; Montero, J.M.; Saiz, D.; Pardo, J.M., 2001. New rule-based and data-driven strategy to incorporate Fujisaki’s F_0 model to a text-to-speech system in castillian spanish. In *Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*. 821-824.

[11] Bertinetto, P.M., 2001. Archivio delle Varietà di Italiano Parlato. Scuola Normale Superiore, Pisa.

[12] Greenberg, S., 1998. Speaking in shorthand- a syllable-centric perspective for understanding pronunciation variations. In *Proceedings of the ESCA Workshop on Modeling Pronunciation variation for Automatic Speech recognition*. 47-56.