CLOSED-FORM ESTIMATION OF THE AMPLITUDE COMMANDS IN THE AUTOMATIC EXTRACTION OF THE FUJISAKI'S MODEL

Solimar de S. Silva and Sergio L. Netto

PEE-DEL/COPPE-Poli, Federal University of Rio de Janeiro
POBox 68504, Rio de Janeiro, RJ, 21945-970, Brazil

ABSTRACT

Generation of F0 contours is required for natural-sounding text-to-speech systems. This task can be accomplished using the Fujisaki's model, proven to be very good to describe F0 contours based on simple linguistically motivated parameters. However, the extraction of the Fujisaki's model parameters is a very intricate problem. Several methods were proposed to solve this problem using iterative optimization techniques. This paper presents a new method capable of extracting the amplitude parameters of the Fujisaki's model analytically. The time-marking commands are still obtained via iterative optimization. The result is a more accurate and less computationally intensive amplitude determination due to the proposed closed-form solution. Examples are included illustrating the application of the proposed method.

1. INTRODUCTION

The generation of pitch (F0) contours is a standard procedure to produce natural-sounding speech. Such task can be successfully accomplished using the Fujisaki's model, which has been proven to be a very good model for the F0 contour in several languages [1]. The Fujisaki's model is based on simple linguistically motivated parameters. However, the automatic determination of the exact Fujisaki's model parameters constitute a very complex problem. A recent trend has been to treat such problem as an optimization problem which can be solved using iterative optimization techniques citeinflection, mixdorff, nakai, salvorossi.

This paper presents a new method for automatic extraction of the amplitude parameters in the Fujisaki's model. The proposed method is based on an analytical development of the overall optimization problem that results in a closed-form solution for the amplitude parameters. In the resulting algorithm, the time-marking commands are determined via iterative optimization, as in standard methods previously found in the literature. The result is a more accurate and less computationally intensive overall procedure for automatically determining the Fujisaki's complete prosody model.

This paper is organized as follows: In Section 2, the Fujisaki's model is briefly reviewed, with emphasis given to its mathematical description of the F0 contour. In Section 3, the automatic extraction of the Fujisaki's parameters is presented as an intricate optimization procedure. In Section 4, a closed-form procedure to determine the amplitudes parameters of the Fujisaki's model is given. Such method can greatly simplify the overall optimization problem, yielding better estimates and a reduced computational effort. The complete automatic estimation algorithm is presented in Section 5 in a step-by-step procedure. Section 6 then presents some computer experiments illustrating the interesting results achieved by the proposed algorithm. Section 7 concludes the paper by emphasizing its main contributions.

2. FUJISAKI'S MODEL

The Fujisaki's model is a superpositional model of intonation related to the physiology of the larynx [3]. The model, as shown in Figure 1, describes the F0 contour of a sentence as the sum of the responses of two critically dumped second-order linear filters. Without taking into consideration the glottal oscillation mechanism, to simplify the model without loss of generality, the Fujisaki's model can be described by [1]

$$\ln \hat{F}_0(t) = \ln F_b + \sum_{k=1}^{N_p} A_{p\,k} G_p(t - T_{p\,k}) + \sum_{k=1}^{N_a} A_{a\,k}[G_a(t - T_{a\,1\,k}) - G_a(t - T_{a\,2\,k})] \quad (1)$$

with

$$G_p(t) = \alpha^2 t e^{-\alpha t} u(t) \quad (2)$$

$$G_a(t) = \left\{ \min \left[1 - (1 + \beta t) e^{-\beta t}, \gamma \right] \right\} u(t) \quad (3)$$

where all variables of interest are described in Table 1.

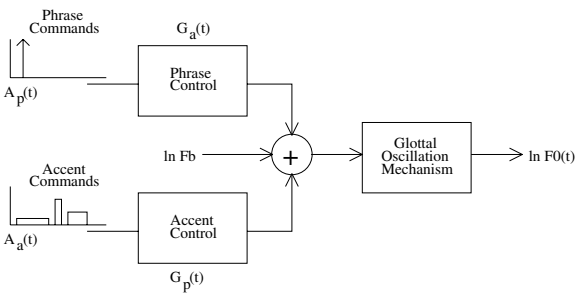

Fig. 1. Block diagram of Fujisaki's model.

Table 1. Variables Related to the Fujisaki's Model.

F_b:	base frequency
N_p:	number of phrase commands
N_a:	number of accent commands;
A_{p_k}:	amplitude of kth phrase command
$A_{a\,k}$:	amplitude of kth accent command
T_{p_k}:	onset of kth phrase command
$T_{a\,1\,k}$:	onset of kth accent command
$T_{a\,2_k}$:	offset of kth accent command
$G_p(t)$:	impulse response of phrase control mechanism
$G_a(t)$:	impulse response of accent control mechanism
α:	natural angular frequency of phrase mechanism
β:	natural angular frequency of accent mechanism
γ:	ceiling level of the accent component
$u(t)$:	unit step function

3. AUTOMATIC EXTRACTION OF FUJISAKI'S MODEL

If $F_0(t)$ is the F0 contour estimated by a pitch determination algorithm (PDA), and $\hat{F}_0(t)$ is the model F0 contour, we can state that

$$\ln F_0(t) = \ln \hat{F}_0(t) + e(t, P) \tag{4}$$

where $e(t, P)$, for $t \in [0, T]$, is the estimation error due to model inaccuracy, and P is the model-parameter vector given by

$$P = \Big\{ A_{p_1} \dots A_{p_{N_p}}, T_{p_1} \dots T_{p_{N_p}}, A_{a_1} \dots A_{a_{N_a}}, T_{a1_1} \dots T_{a1_{N_a}}, T_{a2_1} \dots T_{a2_{N_a}}, F_b \Big\} \tag{5}$$

To estimate P, we may use an analysis-by-synthesis procedure to minimize the mean-squared value of the estimation error given by

$$F(P) = \frac{1}{T} \int_0^T e^2(t, P) dt \tag{6}$$

Hence, parameter extraction of the Fujisaki's model can be regarded as an optimization problem in which one wants to determine P that minimizes $F(P)$. The complexity of such problem, however, suggests iterative solutions.

Some algorithms use linguistic information to bias the search for the parameters [2, 6]. But, for data-driven approaches, it is more convenient to employ a fully automatic algorithm that do not require a priori linguistic information. The majority of the algorithms use the fact that a slowly varying component of the F0 contour is related to the phrase commands and the rapid component is related to the accent commands. This property of the F0 contour was first exploited to detect phrase boundaries in [11]. Other algorithms were in an attempt to improve the overall parameter extraction [4, 5, 7, 8, 9, 10]. In general, most algorithms include the following three steps:

Step 1 (Pre-processing): This stage generates a continuous F0 contour, as required by the Fujisaki's model, despite the unvoiced portions of the original speech, where F0 is not defined. In addition, the Fujisaki's model only describes the macroprosodic component of the F0 contour. Hence, it is necessary to remove all small undulations associated with the microprosody, to eliminate the large errors of the PDA and to provide a smooth F0 contour to the following optimization phase [4];

Step 2 (Initial estimation): In this stage, an initial parameter estimate is obtained, by observing the critical points of the F0 contour or its smoothed versions;

Step 3 (Optimization): The initial estimate is improved by means of an iterative optimization technique in an attempt to minimize $F(P)$.

Most iterative optimization techniques differ on how the initial parameter estimation is performed. In some cases, such procedure is so involved that the second step can be regarded as part of the overall optimization procedure. This paper proposes a modification of the optimization step, by extracting the amplitude parameters by means of an analytical procedure, not requiring (for these parameters) additional iterative processing. The result is a more accurate and less computationally intensive overall procedure for automatically extracting all Fujisaki's parameters.

4. A CLOSED-FORM DETERMINATION OF AMPLITUDE PARAMETERS

To define a numerical procedure to minimize $F(P)$, we must first obtain a discrete version of it as given by

$$\varepsilon^2 = \frac{1}{m} \mathbf{e}^T \mathbf{e} = \frac{1}{m} \left(\mathbf{f_0} - \hat{\mathbf{f_0}} \right)^T \left(\mathbf{f_0} - \hat{\mathbf{f_0}} \right) \tag{7}$$

with

$$\mathbf{f_0} = [\ f_0(t_0) \quad f_0(t_1) \quad \dots \quad f_0(t_{m-1})\ ]^T \tag{8}$$

$$\hat{\mathbf{f_0}} = [\ \hat{f}_0(t_0) \quad \hat{f}_0(t_1) \quad \dots \quad \hat{f}_0(t_{m-1})\ ]^T \tag{9}$$

where

$$f_0(t_k) = \ln F_0(t_k) \tag{10}$$

$$\hat{f}_0(t_k) = \ln \hat{F}_0(t_k) \tag{11}$$

and m is the total number of time samples.

Now, let us define the auxiliary vectors

$$\mathbf{A_p} = [\ A_{p_1} \quad A_{p_2} \quad \dots \quad A_{p_{N_p}}\ ]^T \tag{12}$$

$$\mathbf{A_a} = [\ A_{a1} \quad A_{a2} \quad \dots \quad A_{a\,N_a}\ ]^T \tag{13}$$

$$\mathbf{A} = [\ \mathbf{A_p}^T \ \mid \ \mathbf{A_a}^T\ ]^T \tag{14}$$

$$\mathbf{u} = [\ 1 \quad 1 \quad \dots \quad 1\ ]_{m \times 1} \tag{15}$$

and the auxiliary matrices

$$\mathbf{G_p} = \begin{bmatrix} G_{p_{1,1}} & G_{p_{1,2}} & \cdots & G_{p_{1,N_p}} \\ G_{p_{2,1}} & G_{p_{2,2}} & \cdots & G_{p_{2,N_p}} \\ \vdots & \vdots & \ddots & \vdots \\ G_{p_{m,1}} & G_{p_{m,2}} & \cdots & G_{p_{m,N_p}} \end{bmatrix} \quad (16)$$

$$\mathbf{G_a} = \begin{bmatrix} G_{a\,1,1} & G_{a\,1,2} & \cdots & G_{a\,1,N_a} \\ G_{a\,2,1} & G_{a\,2,2} & \cdots & G_{a\,2,N_a} \\ \vdots & \vdots & \ddots & \vdots \\ G_{a\,m,1} & G_{a\,m,2} & \cdots & G_{a\,m,N_a} \end{bmatrix} \quad (17)$$

$$\mathbf{G} = \begin{bmatrix} \mathbf{G_p} & | & \mathbf{G_a} \end{bmatrix} \quad (18)$$

where

$$G_{p_{i,j}} = G_p(t_{i-1} - T_{p_j}) \quad (19)$$

$$G_{a\,i,j} = G_a(t_{i-1} - T_{a\,1\,j}) - G_a(t_{i-1} - T_{a\,2\,j}) \quad (20)$$

Then, we can rewrite equation (1) as

$$\hat{\mathbf{f_0}} = (\ln F_b)\mathbf{u} + \mathbf{GA} \quad (21)$$

The minimization of ε^2 as given in equation (7), is a complicated optimization problem, due to its highly nonlinear relationship with respect to T_{p_k}, T_{a_k}, and T_{a_k}. However, when analyzing the relationship between ε^2 and A_{p_k}, A_{a_k}, and F_b, one can readily see that this error norm is strictly convex, thus presenting a single local minimum. To make an analytical solution possible, we then consider the following subproblem: Find the parameters A_{p_k}, A_{a_k}, and F_b that minimize ε^2, when the parameters T_{p_k}, T_{a_k}, and T_{a_k} are given.

To solve this subproblem in an analytical way, consider the derivatives of ε^2 with respect to $\mathbf{A}$ and $\ln F_b$:

$$\frac{\partial \varepsilon^2}{\partial \mathbf{A}} = \frac{2}{m}\mathbf{G}^T\{\mathbf{f_0} - [(\ln F_b)\mathbf{u} + \mathbf{GA}]\} = \frac{2}{m}\mathbf{G}^T\mathbf{e} \quad (22)$$

$$\frac{\partial \varepsilon^2}{\partial (\ln F_b)} = \frac{2}{m}\mathbf{u}^T\{\mathbf{f_0} - [(\ln F_b)\mathbf{u} + \mathbf{GA}]\} = \frac{2}{m}\mathbf{u}^T\mathbf{e} \quad (23)$$

These derivatives can be made equal to zero, thus resulting in the following closed-form solution of the subproblem:

$$\ln F_b = \frac{\mathbf{u}^T\mathbf{f_0} - \mathbf{u}^T\mathbf{G}\left[\mathbf{G}^T\mathbf{G}\right]^{-1}\mathbf{G}^T\mathbf{f_0}}{m - \mathbf{u}^T\mathbf{G}\left[\mathbf{G}^T\mathbf{G}\right]^{-1}\mathbf{G}^T\mathbf{u}} \quad (24)$$

$$\mathbf{A} = \left[\mathbf{G}^T\mathbf{G}\right]^{-1}\left(\mathbf{G}^T\mathbf{f_0} - (\ln F_b)\mathbf{G}^T\mathbf{u}\right) \quad (25)$$

5. PROPOSED ALGORITHM

The following algorithm for automatic extraction of the Fujisaki's model parameters was implemented taking advantage of the optimization technique proposed in the previous section:

Step 1: The F0 contour is filtered by a third-order high-pass Butterworth filter with a cutoff frequency of 0.5 Hz. This give us an HFC component, which is subtracted from F0 to give the LFC component [8].

Step 2: LFC is searched for $T = 1\ s$ dominant points [10]. These points are used as an initial guess of the onset of the phrase commands.

Step 3: The amplitudes of the phrase commands ($\mathbf{A_p}$) and the base frequency (F_b) are determined assuming the absence of accent commands. Thus, the F0 contour generated by the model is

$$\hat{\mathbf{f_0}} = (\ln F_b)\mathbf{u} + \mathbf{G_p}\mathbf{A_p} \quad (26)$$

as $\mathbf{A_a}$ is assumed to be a null vector. The parameters $\mathbf{A_p}$ and F_b that minimize ε_p^2 for this are given by equations (24) and (25), replacing $\mathbf{G}$ with $\mathbf{G_p}$ and $\mathbf{A}$ with $\mathbf{A_p}$. This is due to the fact that (26) has the same form as (21).

Step 4: Having determined $\mathbf{A_p}$ and F_b, the phrase response can be reconstructed.

Step 5: The reconstructed phrase response is subtracted from the F0 contour, resulting in a residue, which corresponds (in the case of a perfect reconstruction) to the accent response.

Step 6: We search the residue for Ta dominant points ($Ta = 50\ ms$ was found to be very good in some tests). These points are the initial guess of the onset and offset marks of the accent commands.

Step 7: The amplitudes $\mathbf{A_a}$ of the accent commands are determined. If we define the residue vector

$$\mathbf{r} = \begin{bmatrix} r(t_0) & r(t_1) & \ldots & r(t_{m-1}) \end{bmatrix}^T \quad (27)$$

considering the accent response is given by $\mathbf{G_a}\mathbf{A_a}$, we need to minimize the functional

$$\varepsilon_a^2 = \frac{1}{m}\left(\mathbf{r} - \mathbf{G_a}\mathbf{A_a}\right)^T\left(\mathbf{r} - \mathbf{G_a}\mathbf{A_a}\right) \quad (28)$$

Since the derivative of ε_a^2 with respect to $\mathbf{A_a}$ is

$$\frac{\partial \varepsilon_a^2}{\partial \mathbf{A_a}} = \frac{1}{m}\left\{-2\mathbf{G_a}^T\mathbf{r} + 2\mathbf{G_a}^T\mathbf{G_a}\mathbf{A_a}\right\} \quad (29)$$

the objective function ε_a^2 is minimized by

$$\mathbf{A_a} = \left[\mathbf{G_a}^T\mathbf{G_a}\right]^{-1}\mathbf{G_a}^T\mathbf{r} \quad (30)$$

Step 8: An iterative search may be performed to improve the initial guess of the onset and offset of the commands.

6. COMPUTER EXPERIMENTS

The proposed algorithm was tested for modeling the F0 contour of several sentences spoken in Portuguese language (although it is suitable for any desired language). Figure 2 shows the results for a given sentence. In this figure, the solid line represents the ideal continuous F0 contour, while the 'o' and 'x' marks represent the partial (after Step 7) and final (after Step 8) estimated contours, respectively, for the proposed algorithm. From this figure, one can clearly visualize the positive results achieved with the proposed

method. Figure 3 shows all prosody commands estimated by the proposed algorithm after Step 7 (Figure 3(a)) and after Step 8 (Figure 3), respectively. Notice from these plots how the final step in the algorithm was able to improve the contour estimate by slightly modifying some onset and offset marks (in particular, the first accent mark) of the Fujisaki's model.

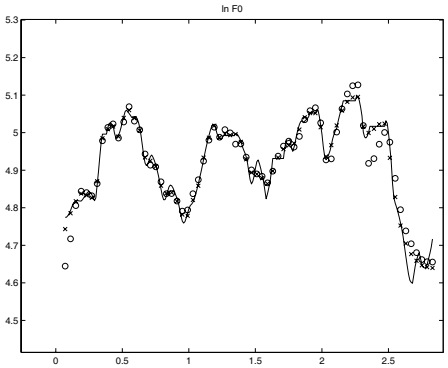

Fig. 2. F0 contour for Sentence 1: original contour (solid line), partial estimate ('o' marks), and final estimate ('x' marks) by the proposed algorithm.

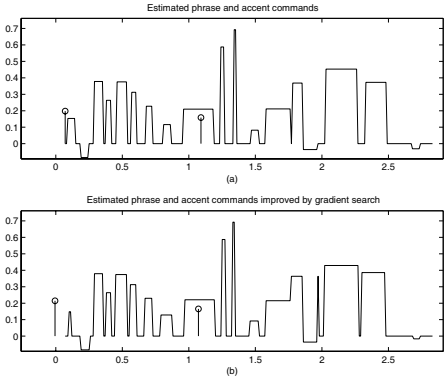

Fig. 3. Commands for Fujisaki's model for Sentence 1 using the proposed algorithm: (a) partial estimate; (b) final estimate.

The performance of the proposed algorithm was also directly compared to the algorithm described in [10], which follows a similar procedure to the one described in Section 3. The final value for the objective function ε^2 in each case is given in Table 2 for three distinct sentences. From this table, we observe that in all cases, the automatic amplitude determination used by the proposed algorithm yielded a more precise estimation of the Fujisaki's parameters. A similar result was obtained for several other sentences.

7. CONCLUSION

A new algorithm for automatic parameter estimation for the Fujisaki's model was presented. The proposed algorithm uses a closed-form analytical procedure to determine the amplitude parameters of the Fujisaki's model. All onset and offset marks are determined via an iterative optimization procedure. The accurate determination of such time positions still constitute the most complicated portion of the model extraction algorithm. The final result is a general method which has shown to be more precise and less computational intensive than previous methods presented in the literature.

Table 2. Optimized Objective Function ε^2 for Different Sentences and Different Estimation Algoritms.

	Algorithm [10]	Proposed Algorithm
Sentence 1	6.092×10^{-3}	2.030×10^{-3}
Sentence 2	6.416×10^{-3}	2.785×10^{-4}
Sentence 3	3.900×10^{-3}	1.868×10^{-3}

8. REFERENCES

[1] H. Fujisaki, *The Production of Speech*, Springer-Verlag, 1983.

[2] H. Fujisaki and S. Ohno, "Prosodic parameterization of spoken Japanese based on a model of the generation process of F0 contours," *Proc. Int. Conf. Spoken Languages Processing* vol. 4, pp. 2439-0-2442, Philadelphia, PA, 1996.

[3] H. Fujisaki, S. Ohno, and C. Wang, "A command-response model for F0 generation in multilingual speech synthesis," *Proc. 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, pp. 299-304, 1998.

[4] H. Fujisaki and S. Narusawa, "Automatic extraction of model parameters from fundamental frequency contours of speech," *Proc. 2nd Plenary Meeting and Symp. Prosody and Speech Processing*, pp. 133–138, Tokyo, Japan, Jan. 2002.

[5] E. Geoffrois, "A pitch contour analysis guided by prosodic event detection," *Proc. European Conf. Speech Communication and Technology*, vol. 2, pp. 793–796, 1993.

[6] J. M. Gutiérrez-Arriola, J. M. Montero, D. Saiz, and J. M. Pardo, "New rule-based and data-driven strategy to incorporate Fujisaki's f0 model to a text-to-speech in Castillian Spanish," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Salt Lake City, Utah, 2001.

[7] H. Kruschke and A. Koch, "Parameter extraction of a quantitative intonation model with wavelet analysis and evolutionary optimization," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Hong Kong, 2003.

[8] H. Mixdorff, "A novel approach to the fully automatic extraction of Fujisaki model parameters," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, vol. 3, pp. 1281–1284, Istanbul, Turkey, 2000.

[9] M. Nakai and H. Shimodaira, "The use of F0 reliability function for prosodic command analysis on F0 contour generation model," *Proc. Int. Conf. Spoken Languages Processing*, Sydney, Australia, 1998.

[10] P. S. Rossi, F. Palmieri, and F. Cutugno, "A method for automatic extraction of Fujisaki-model parameters," *Proc. Speech Prosody*, pp. 615–618, Aix-en-Provence, France, Apr. 2002.

[11] A. Sakurai and K. Hirose, "Detection of phrase boundaries in Japanese by low-pass filtering of fundamental frequency contours," *Proc. Int. Conf. Spoken Languages Processing*, vol. 2, pp. 817–820, Philadelphia, PA, 1996.