Novel Estimation Method for the Superpositional Intonation Model

Humberto Torres and Jorge Gurlekian

***Abstract*—Fujisaki's intonation model parameterizes the F0's contour efficiently and becouse of its strong physiological basis has been successfully tested in different languages. One problem that has not been fully addressed is the extraction of the model's parameters, i.e., given a sentence, which model's parameter values best describe its intonation. Most of the proposed methods strive to optimize the parameters so as to obtain the best fit for the F0 contour globally. In this paper we propose to use text information from the sentence as the main guide or reference for adjusting the parameters. We present a method that defines a set of rules to fix and optimize the model's parameters. Optimization never loses sight of the text structure events that arouse it. When text information is not enough, the algorithm predicts parameters from F0 contour and tie it to the text. The process of parameter estimation can be seen as a way to go from text information to the F0 contour. Parameter optimization is carried out to fit the F0 contour locally. Our novel approach can be implemented manually or automatically. We present examples of manual implementation and the quantitative results of the automatic one. Tested on three corpora in Spanish, English and German, our automatic method shows a performance of 34% better than other tested methods.**

***Index Terms*—Superpositional Intonational Model, Fujisaki's Intonational Model, Model Estimation.**

I. Introduction

THE human voice is more than a simple sequence of words. Expression is a constitutive part of speech, and we expect to find it there. Prosody is a form of expression that adds emotional states, differences of gender, age, attitude, intention, dialectal, among others, to the word sequence. Without expression, speech sounds lifeless and artificial.

Intonation is one of the most important prosody attributes of natural speech. Fundamental frequency (F0) contour and pauses are the two most important physical correlates of intonation. Furthermore, F0's contour models have multiple potential technological applications in fields such as emotion recognition, speaker recognition, speech synthesis, among others, where components of prosody are highly appreciated.

Fujisaki's model of intonation [1] has been successfully tested for different languages [2]–[8], it stands out for its simplicity and strong physiological basis. Currently it is widely used in different application areas [9]–[13]. This model parameterizes F0 contours in an efficient manner: with a small number of parameters we can achieve a desired level of fitting accuracy. A task not satisfactorily addressed is the automatic model parameter extraction, that is, parameter estimation from F0 contours, since it is not directly reversible and hence there is no unique representation [14].

Fujisaki's model describes the vibration of the vocal cords, and hopefully it accounts for the form and function of intonation. Given a speech utterance, existing parameter extraction methods focus on the form of the F0 contour: achieving an F0 contour parameterization as accurate as possible, forgetting and/or putting off the relationship between model parameters with utterance text and intonation function alike. In this paper we present a method to extract the model's parameters with the text utterance as the main policy.

A. Fujisaki's Model

This model –called superpositional and command-response– is hierarchical, additive, parametric and continuous in time.

It allows the efficient and automatic calculation of a reduced parameter set that represents real intonation contours. This model analytically describes the F0 contour in a log scale, as the superposition of three components [15]: a base frequency, accent and phrase, as shown in Fig. 1. The model is based on the anatomical structures and physiology of the vocal cords. Base frequency is related to the basal stress of vocal cords. Phrase components are generated by a set of muscles which tense the vocal cords slowly. Accent components are the answer to the contraction of a muscle set of fast action.

Phrase components are calculated as the response to a critically damped second order linear filter excited with a delta function called phrase command. Accent components result from the response to a similar filter, excited with a step function called accent command.

The F0's contour can be expressed by:

$$\begin{aligned}\ln(F0(t)) = \ln(Fb) + \sum_{i=1}^{N_f} Ap_i Gp_i(t - T0_i) \\ + \sum_{j=1}^{N_a} Aa_j \left\{ Ga_j(t - T1_j) - Ga_j(t - T2_j) \right\} \end{aligned} \tag{1}$$

$$Gp_i(t) = \begin{cases} \alpha_i^2 t \exp -\alpha_i t ; t \geq 0 \\ 0; \qquad\qquad t < 0 \end{cases} \tag{2}$$

$$Ga_j(t) = \begin{cases} \min\left\{1 - (1 + \beta_j t) \exp -\beta_j t, \gamma_j \right\} ; t \geq 0 \\ 0; \qquad\qquad\qquad\qquad t < 0 \end{cases} \tag{3}$$

where:

- Fb: baseline value of fundamental frequency,

The authors are with the Laboratorio de Investigaciones Sensoriales, INIGEM, CONICET-UBA, Av. Córdoba 2351, 9 Piso Sala 2,1120 Ciudad Autónoma de Buenos Aires, Argentina (e-mail: hmtorres@conicet.gov.ar, jag@fmed.uba.ar)

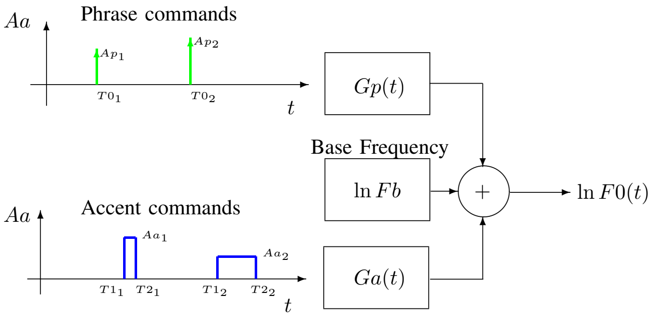

Fig. 1. A scheme of the superpositional intonation model (Adapted from [15]).

- Gp_i: impulse response of the ith phrase control mechanism,
- Ap_i: magnitude of the ith phrase command,
- $T0_i$: timing of the ith phrase command,
- Ga_j: step response of the jth accent control mechanism,
- Aa_j: amplitude of the jth accent command,
- $T1_j$: onset of the jth accent command,
- $T2_j$: end of the jth accent command,
- α_i: is the eigenvalue of the ith phrase control mechanism,
- β_j: is the eigenvalue of the jth accent control mechanism,
- γ_j: is the maximum value of the jth accent component.

The parameters α and β characterize the dynamic properties of the laryngeal mechanisms of phrase and accent control. Together with γ they can be considered practically constant for all speakers. Fb must be estimated for each utterance, but it is assumed to be constant for individual speaker [15].

B. Extraction of Model's Parameters

The model's parameters can be estimated manually. Tools have been designed to facilitate this task[1], but no algorithm has been reported to show how it can be done. In addition, the task of estimating the model's parameters for a corpus can be incredibly time consuming, and the consistency of the value set obtained cannot be ensured.

Several approaches have been proposed to automatically estimate model's parameters. One is the Mixdorff method [16], called here A-ME, successfully tested in different languages. Although this method is completely automatic, the author proposes a *post-hoc* manual correction to eliminate spurious commands which cannot be justified linguistically [17].

Given the characteristics of the F0 signal, it must be pre-processed before attempting to extract the model's parameters [18]. In general, the pre-processing includes four steps. First, we must remove gross errors generated in the estimation process of F0 contour. For example, the effects of edges and the estimation error of frequency doubling/halving. A median filter can be used to eliminate these spurious data [19]. Secondly, we must mitigate the effect of microprosody, which is not reached by the model. Searching abrupt changes in the boundaries of voice-voiceless phonemes has been successfully applied to detect micro-prosody [18]. In third place, the segments corresponding to voiceless phonemes are interpolated; and finally, overall F0 contour smoothing is performed. Momel [20], piecewise cubic interpolation [19], quadratic spline [16] or median filter [21] methods, among others, have been used for the curve stylization.

With the pre-processed signal, a first-order approximation to the model's parameters is obtained. In general, the minimum value of F0 contour is assigned to Fb. Fujisaki et al. [19] propose to define the first-order model from the maximum and minimum of F0 contour, iteratively, first accent commands and then phrase commands. Mixdorff [16] performs a low pass filter with a cutoff frequency of 0.5 Hz [22] to separate the phrase and accent components. Kruschke and Lenz [23] propose the alternative of Wavelet Transform to carry out the separation of both components. These authors support their use because the two components are not stationary, close in frequency, and below 10 Hz. In this scenario it is not easy to separate the phrase and accent components, including the uncertainty in the threshold value, and low frequency resolution in the estimation methods. As before, both the minimum and maximum values of F0 contour are used to obtain a first approximation to the model's parameters.

The first order model could be optimized, in one or more steps, in order to minimize the fitting error of estimated F0 regarding the smoothed F0 contour. Hill Climb Search [16] is used to minimize the overall mean-square-error. Parameters are sequentially optimized in *a priori* order. The order in which the parameters are optimized is one of the disadvantages of this technique: different orders of optimization give different values to the parameters. As an alternative, it has been proposed that the Evolutionary Algorithms [23] and Genetic Algorithms [11] be used to optimize all parameters simultaneously. The ability to perform a global search in an n-dimensional solution space is the main justification to use these methods, even more when the error surface has many local minima.

Silva and Netto [24] have proposed a closed-form estimation method for command amplitudes, assuming as previously known the command positions. Using the first-order model proposed by Mixdorff [16], they iteratively optimize the model's parameters by estimating the amplitudes in closed-form and then correcting the command positions until convergence.

Fujisaki's model has also been formulated as discrete-time stochastic process [25], resulting in a F0 contour statistical model. The aim is to introduce statistical methods to learn the Fujisaki's model parameters from speech F0 contours, and apply it to speech synthesis based on hidden Markov models.

Pfitzinger and Mixdorff [14] discuss the accuracy of the current methods to estimate the model's parameters solely on the basis of the extracted natural F0 contours. The authors emphasize the importance of F0 contour stylization, as a way to ensure the elimination of micro-prosody. Algorithm initialization is a critical issue, given that different sets of initial parameter values produce different model estimations with varying accuracy in fitting the F0 contour.

Hirose et al. [26] have proposed to introduce linguistic information in the estimation process for a Japanese corpus. The model estimation is performed in two stages: first an

[1] FujiParaEditor: http://www.tfh-berlin.de/~mixdorff/thesis/fujisaki.html

automatic estimation trying to fit the F0 contour, and second a correction of the parameters using *ad hoc* rules based on linguistic hypotheses. In a later work [27], they used linguistic information to obtain a first approximation of the location of the command, which is then adjusted by an iterative analysis-by-synthesis process. In this process the linguistic information is automatically extracted using binary regression trees, which were created automatically from a portion of the corpus manually analyzed beforehand.

For TTS application, Agüero [28] proposes to set the same parameters for all commands that share the same linguistic features. The search is performed on the whole corpus of data. The purpose is to facilitate the prediction of commands from the text, but does not ensure a proper optimization of the model's parameters.

For Spanish, Torres and Gurlekian [11] consider linguistic aspects, such as the positions of pauses and syllables with lexical stress. In a later paper [29], the authors extended their approach for English and German.

C. *Outline*

The purpose of this paper is to introduce a new approach for estimating Fujisaki model parameters, primarily guided by the linguistic content of sentences, performing the capture of speech events by local optimization of the parameters. First, we present our assumptions about how the model should be considered. Then we outline an approach for extraction of model's parameters, and finally we layout a computational algorithm implementing our new estimation method. In this paper we present the results of applying the new automatic estimation method to German, English and Spanish corpora.

This paper is organized as follows: In Section II, we present our approach to estimate of model's parameters. The experiments and the results obtained are presented and discussed in Section III. Finally the conclusions are summarized in Section IV.

II. New estimation method

A. *Motivation and Background*

In a previous work [29], we have presented a method to initialize the Mixdorff algorithm for parameter extraction of Fujisaki's model. Our approach sought to introduce linguistic information in the estimation process. Our main assumptions were: 1) the position of accent commands will be close to the location of syllable with lexical stress in content words; 2) it is reasonable to expect accent commands occurring at or near the end of intermediate intonational phrases; these are model approximations to ”boundary tones” in some linguistic transcription methods [30]; and 3) phrase commands will be near at intonational phrase beginnings, as has been reported in previous studies [11], [31]. Under these assumptions we initialize the estimation method, obtaining a first order model, then it is optimized with the aim to minimize the overall fitting error between the original F0 and F0 estimated by the Fujisaki's model. We call this method L-ME. Unfortunately in the process of parameter optimization, many commands lose their relation regarding the structures which gave them their origin. This prevents some kind of linguistic meaning being assigned to the estimated commands.

Figs. 2 and 3 show examples of model's parameter estimation using both A-ME (Fig. 2.b) and L-ME (Fig. 2.c) methods, and where some of the aforementioned drawback are shown. Fig. 2 shows how the optimization method fails at the second phrase command. The insertion point is delayed in Mixdorff's method, and movement of F0 contour is compensated for an accent command with large amplitude. L-ME method does the opposite, the insertion point moves backward and absorbs the first accent command. In addition, subsequent accent commands have a long duration. Fig. 3 shows that one intermediate intonative phrase starts at the beginning of the third word. Both methods fail to insert a phrase command to model this movement in the F0 contour. Again, previous accent commands have long durations to compensate a bad estimation of phrase commands.

In this work, we hold the parameter initialization and we propose a new method to the parameter extraction. Our method is based on the following guidelines:

- α, β and γ model's parameters are constant and they are fixed *a priori*.
- Commands are developed from left to right, representing running speech events. Parameter optimization must be local, associated with the text that gave rise to it, leaving in second place the global error minimization in the fit of the F0 contour. Throughout the estimation process we should always maintain a link between the commands and the text events. In our algorithm we have not considered negative amplitude commands. In spite of the fact that some languages use commands with negative amplitude [32], our reluctancy to allow negative commands is because it is difficult to find them in seminal Fujisaki physiological motivation, since negative commands go against to its assigned role. Subsequent works have shown that the muscles responsible for downfall of vocal cords stress have to be consider in the formulation of Fujisaki model to increase language coverage [33], [34].
- Fb allows to eliminate the constant component of F0, and its value should be slightly lower than the minimum value of F0, leaving a gap for the action of phrase components. Unlike our previous approach [29], this value should be optimized for each sentence. Given their physiological correlate, we expect that their values will have a small scattering.
- Phrase components shape the smooth movement of the F0 contour, giving rise to the intonative phrases. The intonative phrases start after a pause, or in a tonal change without pause. So, we have to put one command phrase before each pause. If the components of the commands associated with the pauses are not enough to model the smooth movement of F0, new phrase commands must be inserted. A possible insertion point is at the onset of each accentual phrase. Phrase commands are performed from left to right. After subtracting phrase components from the F0 contour, only rapid movements should remain, which could be associated with the accent components.
- Accent commands are intended to model the abrupt

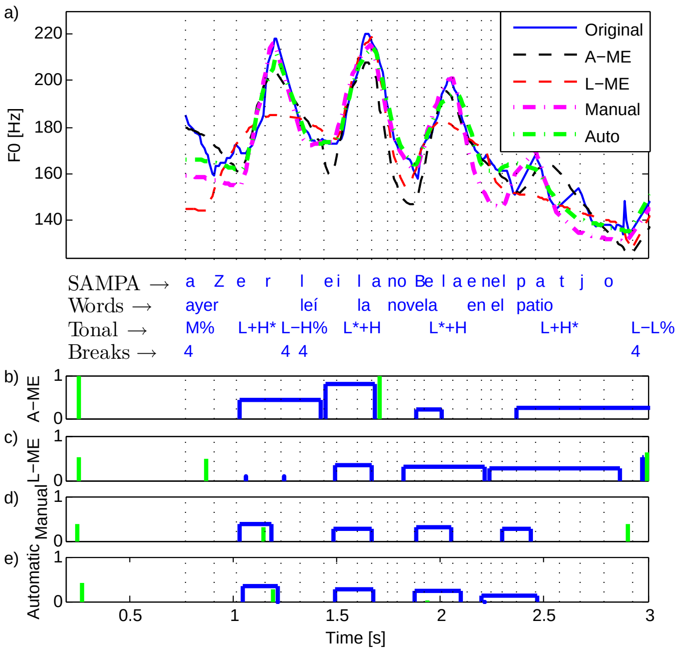

Fig. 2. Example of F0's contour and parameter extraction methods, in Spanish. a) F0's and phones, words, and ToBI labels. Commands for: b) Mixdorff (A-ME), c) lexical initialized (L-ME), d) *manual*, and e) *automatic* implementations of the new extraction method. The text sentence is *"Ayer leí la novela en el patio, ..."* (*"Yesterday I read a romance in the yard, ..."*).

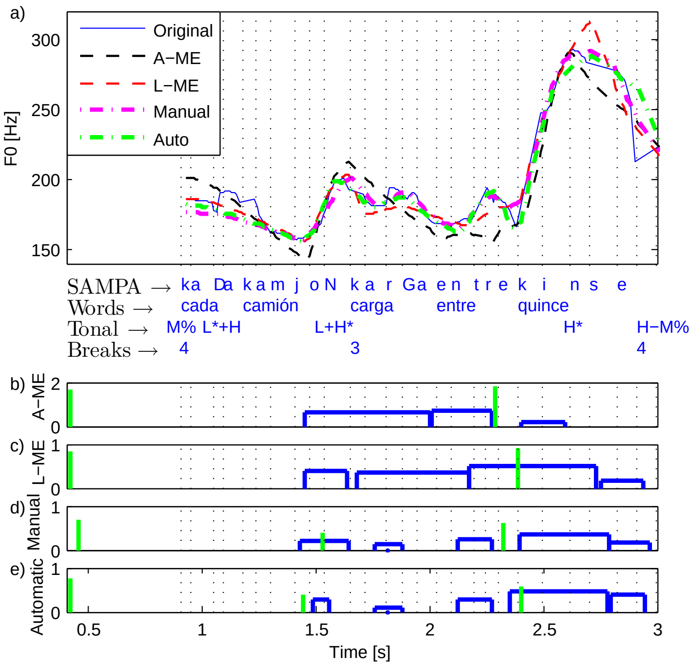

Fig. 3. Example of F0's contour and parameter extraction methods, in Spanish. a) F0's and phones, words, and ToBI labels. Commands for: b) Mixdorff (A-ME), c) lexical initialized (L-ME), d) *manual*, and e) *automatic* implementations of the new extraction method. The text sentence is *"Cada camión carga entre quince ..."* (*"Each truck loads between fifteen ..."*).

changes in the F0 contour. Positions of syllables with lexical stress at content words and pause onset are natural candidates for insertion points of accent commands. If there are still abrupt movements without modeling extra accent commands should be added. These new command will be associated with syllables with lexical stress at the closest function words.

B. Model parameter extraction algorithm

We suppose that we have a properly estimated F0 contour without gross errors and without considering the effects of micro-prosody. We also assume that we have the phonetic and orthographic labeling of sentences. Values of the F0 contour have to be in a log scale. After estimating a component and/or command, its contribution to the F0 contour must be subtracted before estimating other parameters. This means that the reference for estimation of the next command is the residual F0 contour.

The five steps of proposed algorithm are listed below:

- *Step 1: Estimation of Fb.* The value of Fb will be associated with the minimum value of $F0$. It will be optimized for each sentence taking care of the tips given in the previous section. In our experiments, a value of $Fb = min(F0) - 15Hz$, it was a good starting point, allowing the realization of the other model's components.
- *Step 2: Estimation of phrase commands associated with the pauses.* Insert one phrase command at the end of each pause. The first command will be placed $1/\alpha$ s before of sentence onset, getting the maximum of their component at sentence onset. Its amplitude will be fixed in order to maximize the contribution of the component, but without exceeding the $F0$ contour. For the following commands, from left to right, the amplitude and position have to be optimized. The positions may take a value in a neighborhood of its associated pause end. As before, the aim of the optimization is to maximize the component contribution, but without exceeding the residual $F0$ contour.
- *Step 3: Insertion of phrase commands.* From left to right, we must first find a candidate for an accent command, then we have to search the peak in the residual $F0$ that will be generated for this command, and their closest local minimums, at left and right of this peak. If the minimum minimorum is not near to zero, then the end of the left content word is a candidate point to insert a phrase command. The final position and amplitude of the inserted command have to be optimized, as in the previous step. After each command insertion, all commands on the right side have to be optimized again.
- *Step 4: Estimation of accent commands.* First, from left to right, we take the midpoints of the stressed vowels in content words as possible positions of accent commands. The position of the closest peak in the residual $F0$ is candidate for the end point of the command, and the left local minimum of this point is fixed as the starting point of the command. Finally, the end of each intonation phrase, delimited by phrase commands, will also be a candidate for an accent command. If at that point the residual $F0$ value is high, we have to look in a left neighborhood for a peak, and take this point as the end time of the command. From this point, the left local

Algorithm 1 Main algorithm for estimating the values of the Fujisaki's model parameters.

```
Require: $\boldsymbol{F0}$: F0 contour, in $Hz$.
Require: $\{\alpha,\ \beta,\ \gamma\} \leftarrow$ model's parameters.
Require: $F0_{offset}$.
Require: $\theta_{F0_k}$, $k = 1; 2$ $\leftarrow$ $F0$ thresholds to insert a PC.
Require: $\theta_{F0_k}$, $k = 3 \dots 8$ $\leftarrow$ $F0$ thresholds to allow a AC.
Require: $\theta_{T0_{hyp}}$ $\leftarrow$ minimum allowed hyphenation of PCs.
Require: $\theta_{T0_k}$, $k = 1; 2$ $\leftarrow$ thresholds for distance from end
    of sentence to last $T0$.
Require: $\theta_{T2T1_{min}}$ $\leftarrow$ minimum allowed AC duration.
Require: $\theta_{T2T1_k}$, $k = 1; 3$ $\leftarrow$ threshold for insert a PC.
  EstimateFb($\boldsymbol{F0}$, $Fb$)
  EstimatePCs($\boldsymbol{T0}$, $\boldsymbol{Ap}$, $\boldsymbol{F0}$, $\boldsymbol{F0}_{PCs}$)
  InsertPCs($\boldsymbol{T0}$, $\boldsymbol{Ap}$, $\boldsymbol{F0}$, $\boldsymbol{F0}_{PCs}$)
  EstimateACs($\boldsymbol{T0}$, $\boldsymbol{T1}$, $\boldsymbol{T2}$, $\boldsymbol{Aa}$, $\boldsymbol{F0}$)
  InsertACs($\boldsymbol{T1}$, $\boldsymbol{T2}$, $\boldsymbol{F0}$)
```

Algorithm 2 Estimate Fb. Step 1 in manual version.

```
procedure EstimateFb($\boldsymbol{F0}$, $Fb$)
    $Fb \leftarrow min(\boldsymbol{F0}) - F0_{offset}$.
    $\boldsymbol{F0} \leftarrow Ln(\boldsymbol{F0}) - Ln(Fb)$
end procedure
```

minimum will be fixed as a possible onset of command. Amplitude, start and end of commands will be optimized in a similar way as for optimizing phrase commands.

- *Step 5: Insertion of accent commands*. We seek the maximum peak in the residual $F0$ into segments that correspond to function words that did not have accent commands. If peak value is high, this point is candidate to a new accent command. We can process this point as belonging to a content word, and estimate its command parameters just like in the previous step. We repeat this procedure until we do not find more peaks that meet the required conditions.

C. Manual implementation

The algorithm introduced in the previous section can be implemented manually. Figs. 2.d and 3.d show examples of model's parameter extraction with this approach. Estimated F0 contours fit the original F0 better than A-ME and L-ME. Also commands have lower amplitudes and durations. In our experiments, the graphical tool FujiParaEditor has been helpful, wherein model's parameters are extracted from the F0 contour using an analysis-by-synthesis method.

D. Automatic implementation

If we want to estimate the model for a speech database, for *a posteriori* comparison, manual estimation is not feasible nor advisable. Estimated parameter values will have a bias because of the perceptual assessment, and besides, it is very possible that we fail to get the best parameter values in the optimization stage. Therefore, in this section we present a computational implementation.

Algorithm 3 Phrase commands estimation. Step 2.

```
procedure EstimatePCs($\boldsymbol{T0}$, $\boldsymbol{Ap}$, $\boldsymbol{F0}$, $\boldsymbol{F0}_{PCs}$)
    $\boldsymbol{F0}_{PCs} \leftarrow \boldsymbol{F0}$
    $\boldsymbol{T0} \leftarrow$ onset of first sentence - $(1/\alpha)$s.
    OptimizePCs($\boldsymbol{T0}$, $\boldsymbol{Ap}$, $\boldsymbol{F0}_{\boldsymbol{PCs}}$)
    for all next sentence delimited by a pause, $k = 2, ..$ do
        $t_{new} \leftarrow$ onset of $kth$ sentence.
        $\boldsymbol{T0} \leftarrow [\boldsymbol{T0};\ t_{new}]$.
        $\boldsymbol{F0}_{PCs} \leftarrow \boldsymbol{F0}$
        OptimizePCs($\boldsymbol{T0}$, $\boldsymbol{Ap}$, $\boldsymbol{F0}_{PCs}$)
    end for
end procedure
```

Algorithm 4 Phrase commands optimization

```
procedure OptimizePCs($\boldsymbol{T0}$, $\boldsymbol{Ap}$, $\boldsymbol{F0}$)
    for all $\{T0_k, Ap_k\}$, $k = 1, 2, \dots$ do
        Optimize $\{T0_k, Ap_k\}$, by:
            $\begin{cases} min(\boldsymbol{F0} - \sum_{i=1}^{k} Ap_i * Gp_i) \\ (\boldsymbol{F0} - \sum_{i=1}^{k} Ap_i * Gp_i) > 0 \end{cases}$
        $\boldsymbol{F0} \leftarrow \boldsymbol{F0} - \sum_{i=1}^{k} Ap_i * Gp_i$
    end for
end procedure
```

The pseudocode of the implementation conducted is presented below. Algorithm 1 is the main procedure, fixing algorithm parameters and calling subroutines. We had to define a set of thresholds to determine the insertion, optimization and validity of the commands. PC and AC are used to denote the phrase and accent commands, respectively. Algorithm 2 implement the Fb estimation (step two). Algorithms 3, 4 and 5 implement the phrase commands estimation (step 2 and 3). Algorithms 6, 7 and 8 implement the phrase commands estimation (step 4 and 5).

The parameter optimization of each command can be made by any method that fulfills the restrictions contained in the pseudocode. In our implementation we conducted a grid search in the neighborhood of the parameter initial values.

In the next section we present the results obtained with this automatic implementation on three speech corpora.

III. Experiments Results and Discussion

To evaluate the performance of our method, we conducted experiments with the automatic implementation using three speech corpora. The raw F0 values were filtered, using phoneme labels to remove spurious values on voiceless segments. Afterward, we used a simple window mean filter to reduce micro-prosody. Compound words were split into simple words.

The estimated model, phrase and accent command amplitudes and positions, as well as F_b, α, β and γ, were used to resynthesize the F0's values by means of the Fujisaki's model. The semitone scale was used to evaluate the resulting contours versus the real F0 contour.

Algorithm 5 New phrase commands insertion. Step 3.

```
procedure InsertPCs($\boldsymbol{T0}$, $\boldsymbol{Ap}$, $\boldsymbol{F0}$, $\boldsymbol{F0}_{PCs}$)
    $t_{end} \leftarrow$ End of sentence.
    for each content words, $n = 2, 3, \ldots$ do
        $t_{sv} \leftarrow$ midpoint of stressed vowel.
        $t_{sv+} \leftarrow$ midpoint of next stressed vowel.
        $\{t_l; t_r\} \leftarrow$ position of the $\boldsymbol{F0}_{PCs}$ local minimums
                closest at $t_{sv}$, left and right, respectively.
        $F0_{min} \leftarrow min(\boldsymbol{F0}_{PCs}(t_l); \boldsymbol{F0}_{PCs}(t_r))$
        $t_{wrd_n} \leftarrow$ onset of $nth$ content words.
        $p \leftarrow$ index of previous PC.
        if $\{\neg$(first content word after a PC) $\}$
         $\wedge\{[t_{wrd_n} - T0_p] > T0_{hyp}\}$
         $\wedge\{\ \{F0_{min} > \theta_{F0_1}\}$
            $\vee\{\ \ \{\boldsymbol{F0}_{PCs}(t_r) > \theta_{F0_2}\}\ \wedge\ \{t_l < t_{wrd_{n-1}}\}\ \}$
            $\vee\{\ \ \{\boldsymbol{F0}_{PCs}(t_l) > \theta_{F0_2}\}$
               $\wedge\{\ [t_{end} - t_{wrd_n}] > \theta_{T0_1}\}$
               $\wedge\{t_l < t_{wrd_{n-1}}\}\}$
            $\vee\{\ \ \{$last content word$\}$
               $\wedge\{\ \{\ \{\boldsymbol{F0}_{PCs}(t_l) > \theta_{F0_1}\}$
                    $\wedge\{\ \{\ [t_{end} - t_l] > \theta_{T0_1}\}$
                       $\vee\{\ [t_{end} - t_{wrd_n}] > \theta_{T0_1}\}\}\}$
                 $\vee\{\ \{\boldsymbol{F0}_{PCs}(t_l) > 2 * \theta_{F0_1}\}$
                    $\wedge\{\ [t_{end} - t_l] > \theta_{T0_2}\}\}\}\}$
            $\vee\{\ \ \{\neg$ last content word$\}$
               $\wedge\{\boldsymbol{F0}_{PCs}(t_r) > \theta_{F0_2}\}$
               $\wedge\{\ \left[t_{sv_+} - t_r\right] > \theta_{T2T1_1}\}\}\}$ then
            $\boldsymbol{T0} \leftarrow sort\left([\boldsymbol{T0}; t_{wrd_n}]\right)$
            $\boldsymbol{F0}_{PCs} \leftarrow \boldsymbol{F0}$
            OptimizePCs($T0$, $\boldsymbol{Ap}$, $\boldsymbol{F0}_{PCs}$)
        end if
    end for
    $\boldsymbol{F0} \leftarrow \boldsymbol{F0}_{PCs}$
end procedure
```

Algorithm 6 Accent commands estimation. Step 4.

```
procedure EstimateACs($\boldsymbol{T0}$, $\boldsymbol{T1}$, $\boldsymbol{T2}$, $\boldsymbol{Aa}$, $\boldsymbol{F0}$)
    $\boldsymbol{T2} \leftarrow$ midpoint of stressed vowel on all content words.
    for from the 2nd to penultimate $T0_k$, $k = 2, 3, \ldots$ do
        $t_{wrd} \leftarrow$ End of previous word at $T0_k$.
        $\boldsymbol{T2} \leftarrow [\boldsymbol{T2}; t_{wrd}]$
    end for
    for all $T2_k$, $k = 1, 2, ..$ do
        $\{t_{min_l}; t_{min_r}\} \leftarrow$ position of the $F0$ local minimums closest at $T2_k$, left and right, respectively.
        $\{t_{max_l}; t_{max_r}\} \leftarrow$ position of the $F0$ local maximums closest at $T2_k$, left and right, respectively.
        if $T2_k \notin$ end of an intermediate phrase then
            if $\{\ \{k = 1\}$
                $\wedge\{\ \{\ \{|T2_k - t_{max_r}| < |T2_k - t_{max_l}|\}$
                    $\wedge\{t_{min_r} > t_{max_r}\}\}$
                  $\vee\{\boldsymbol{F0}(t_{max_l}) < \theta_{\boldsymbol{F0}_7}\}\}\}$
             $\vee\{\ \{k \neq 1\}$
                $\wedge\{\ \{T2_{k-1} > t_{max_l}\}$
                  $\vee\{\{t_{min_l} > t_{max_l}\} \wedge \{\boldsymbol{F0}(t_{min_l}) < \theta_{F0_8}\}\}$
                  $\vee\{\ \{\theta_{T2T1_2} * |T2_k - t_{max_r}| <$
                            $|T2_k - t_{max_l}|\}$
                    $\wedge\{\boldsymbol{F0}(t_{max_l}) < \theta_{F0_5} * \boldsymbol{F0}(t_{max_r})\}\}$
                  $\vee\{\boldsymbol{F0}(t_{max_l}) < \theta_{F0_3}\}$
                  $\vee\{\ \{(t_{wrd} + \theta_{T2T1_3}) > t_{max_l}\}$
                    $\wedge\{\boldsymbol{F0}(t_{max_r}) > \theta_{F0_3}\}$
                    $\wedge\{\boldsymbol{F0}(t_{max_l}) > \theta_{F0_6}\}\}\}\}$ then
                $T2_k \leftarrow t_{max_r}$
            else
                $T2_k \leftarrow t_{max_l}$
            end if
        else
            if $\{t_{max_l} > t_{min_l}\} \wedge \{\boldsymbol{F0}(t_{max_l}) > \boldsymbol{F0}(T2_k)\}$ then
                $T2_k \leftarrow t_{max_l}$
            end if
        end if
        $t_l \leftarrow$ position of the $F0$ minimum closest at left of $T2_k$.
        $T1_k \leftarrow t_l$
        OptimizeAC($T1_k$, $T2_k$, $Aa_k$, $\boldsymbol{F0}$)
        if $\neg\{\boldsymbol{F0}(T2_k) < \theta_{F0_3}\}\ \vee\ \neg\{\{T2_k - T1_k\} < \theta_{T2T1_min}\}$ then
            Erase the $kth$ command.
        end if
    end for
end procedure
```

A. Speech Material

We tested the automatic method in three different languages: German, English and Spanish.

1) German Database: For German we used the IMS Radio News Corpus [35] which consists of German news texts read by professional male speakers. The data selection comprises 73 news articles automatically segmented into phonemes according to the German SAMPA[2] inventory followed by manual corrections. The syllables with lexical stress also were manually labeled.

2) English Database: For English we used the CSTR US KED Timit database[3] which contains 453 phonetically balanced utterances spoken by a US male speaker. The database was hand labeled in phonemes, syllables and words, and carefully corrected. The syllables with lexical stress were also manually labeled.

3) Spanish Database: For Spanish we use the Emilia corpus, it was created to be used in a text-to-speech system [36]. The corpus sentences had natural inflections with different number of intonational phrases, recorded by a professional female announcer, native of Buenos Aires City. Its text corpus consists of 1591 declarative sentences and 235 interrogative sentences, extracted from Argentine newspapers published in Buenos Aires. The sentences contain up 97% of all Spanish syllables, in both stress and unstressed conditions, and all possible syllabic positions within the word [31], [37].

The files were manually labeled on different tiers: phonetic according to Argentinian SAMPA [38], orthographic,

[2] http://coral.lili.unibielefeld.de/Documents/sampa-d-vmlex.html

[3] http://festvox.org/dbs/dbs_kdt.html

Algorithm 7 Accent commands insertion. Step 5.

```
procedure InsertACs($\boldsymbol{T1}$, $\boldsymbol{T2}$, $\boldsymbol{F0}$ )
    for all function word, $k = 1, 2, \dots$ do
        $\{t_{ons}; t_{end}\} \leftarrow$ onset and end of $kth$ function words.
        if $\{ \ (T2_k > t_{ons}) \wedge (T2_k < t_{end})$
            $\wedge (T1_k > t_{ons}) \wedge (T1_k < t_{end})\} = \oslash$ then
            $F0_{max} \leftarrow max(\boldsymbol{F0}(t_{ons} : t_{end}))$
            if $F0_{max} > \theta_{F0_4}$ then
                $t_{new} \leftarrow t|\boldsymbol{F0}(t) = max(\boldsymbol{F0}(t_{ons} : t_{end}))$
                $\boldsymbol{T2} \leftarrow [\boldsymbol{T2}; t_{new}]$
                OptimizeAC($T1_{end}$, $T2_{end}$, $Aa_{end}$, $\boldsymbol{F0}$)
            end if
        end if
    end for
end procedure
```

Algorithm 8 Accent command optimization

```
procedure OptimizeAC($T1_k$, $T2_k$, $Aa_k$, $\boldsymbol{F0}$)
    $AC_k \leftarrow Aa_k * \{Ga_k(t - T1_k) - Ga_k(t - T2_k)\}$
    Optimize $\{T1_k, T2_k, \ Aa_k\}$, by:
        $\begin{cases} min(\boldsymbol{F0} - AC_k) \\ (\boldsymbol{F0} - AC_k) > 0 \end{cases}$
    $\boldsymbol{F0} \leftarrow \boldsymbol{F0} - AC_k$
end procedure
```

break levels between words, and tonal marks according to an extended ToBI method for Argentine Spanish [31]. Part-of-speech and syntactic layers were also indicated.

B. Setup

In Section II-D we present the pseudocode of our algorithm to estimate the model's parameters. In our implementation there is a set of parameters to be fix, which are listed in the first part of Algorithm 1. For our experiments we have fixed empirically the values of these parameters. We have used the same values in all experiments for the three corpora, and we have listed them in Table I.

TABLE I
VALUES OF ALGORITHM PARAMETERS.

Parameter	Value	Description
α	2	Model's parameters
β	20	
γ	0.9	
$F0_{offset}$	15 Hz	Fb offset
θ_{F0_1}	0.07 $ln\,Hz$	$ln(F0)$ thresholds to insert a PC
θ_{F0_2}	0.25 $ln\,Hz$	
θ_{F0_3}	0.05 $ln\,Hz$	$ln(F0)$ thresholds to allow a AC
θ_{F0_4}	0.14 $ln\,Hz$	
θ_{F0_5}	0.75	
θ_{F0_6}	0.25 $ln\,Hz$	
θ_{F0_7}	0.10 $ln\,Hz$	
θ_{F0_8}	1.00 $ln\,Hz$	
$\theta_{T0_{hyp}}$	0.375 s	Minimum allowed hyphenation of PCs
$\theta_{T2T1_{min}}$	0.02 s	Minimum allowed AC duration
θ_{T0_1}	0.75 s	Thresholds for distance from end of sentence to last $T0$
θ_{T0_2}	0.50 s	
θ_{T2T1_1}	1.00 s	Thresholds for insert a PC
θ_{T2T1_2}	0.8	
θ_{T2T1_3}	0.02 s	

C. Results and Discussion

The root mean square error (RMSE) in semitones (ST) and the average command density per second of the different experiments are shown in Fig. 4. We have also included standard deviations as a measured of dispersion values. The results obtained with the standard Mixdorff algorithm and our previous approach have been included as references.

The new automatic method outperform previous results. For the Spanish corpus, results improved 34% and 49% better than A-ME and L-ME, respectively. We can assume that the significant improvement in results is due to an increase in Ap rates: 28% and 111%, in A-ME and L-ME algorithms, respectively. The huge increase in the rates of Ap with respect to the L-EM can be explained by insertion of phrase commands at the beginning of the intermediate intonational phrases, unlike the L-EM algorithm where we only considered phrase commands associated with pauses. Slight differences in the Aa rates are also observed: 8% higher compared to the L-EM and 7% lower compared with A-ME.

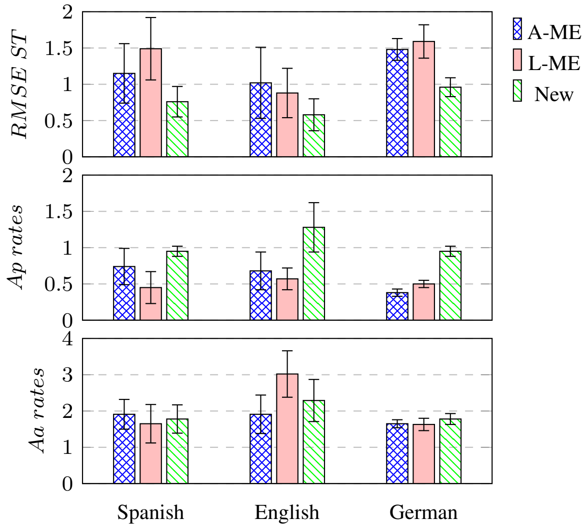

Fig. 4. Results for three databases and three algorithms. Algorithms: standard Mixdorff (A-ME); linguistically initialized (L-ME); and our automatic approach (New). The RMSE is given in ST and the rates in commands per second. Standard deviation is included as scattering measure.

For the English corpus, results improved in 43% and 34% for the A-ME and L-EM, respectively. Both Ap rates are increased by 88% and 125% compared with A-ME and L-EM methods, respectively. We also found variations in the Aa rates: 20% plus for the A-ME and a 25% minus compared to L-EM. This sharp drop in Aa rates with respect to L-EM can be a result of increased Ap rates: many accent commands were replaced by phrase commands associated with intermediate phrases.

For the German corpus, results improved in 35% and 40%

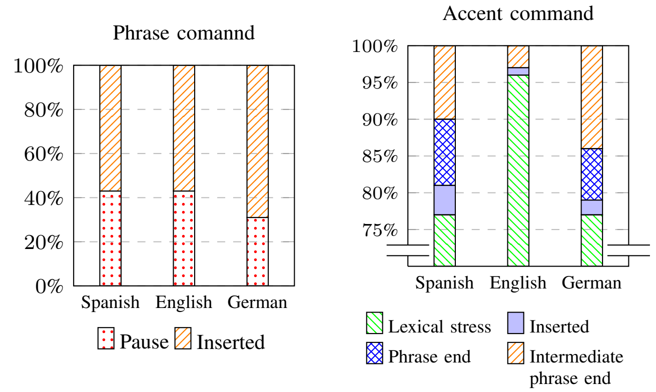

Fig. 5. Statistics of command origin, expressed in percent.

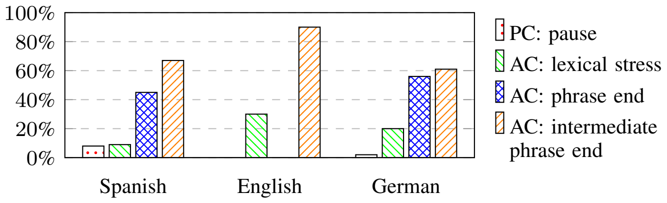

Fig. 6. Statistics of command deleted, discriminated by origin, expressed in percent. PC: phrase command. AC: accent command.

for the A-ME and L-EM, respectively. Both Ap rates are increased by 150% and 90% compared to A-ME and L-EM methods, respectively. It is difficult to find an explanation for this abrupt increase in rates Ap with respect to A-ME. We can mention that utterances in German corpus are two or three times longer than utterances in the Spanish and English corpus, so we can expect greater number of intermediate intonative phrases. We also found variations in the Aa rates: 8% and 9% minus for the A-ME and L-EM, respectively.

The origin of each command, expressed in percent, are shown in Fig. 5. In Fig. 6 we present the percentage of first-order model's commands which were dismissed.

For the three corpora, the number of phrase commands inserted are superior to that originated in a pause. As we had assumed in a previous paragraph, the number of generated intermediate intonation phrases, defined by inserting of phrase command, is higher in the German corpus. The number of phrase commands originated in a pause that are deleted is low. In particular, the corpus in English no has pauses within utterance.

The vast majority of accent commands are rooted in the stressed syllables of content words. This is most evident in the English corpus. The number of accent commands deleted associated with stressed syllables in content words are disperse in the three corpora. For the corpus in English, the non-use of a command accent at the end of the intermediate intonation phrases is remarkable.

Fb mean and deviation values estimated in the three corpora are shown in Fig. 7. As we had postulated, the dispersion of Fb estimated values is small: 10.18% for Spanish, 5.60% for English and 2.46% for German. Remember that the Spanish corpus corresponds to a female announcer, unlike the English and German speakers, who were male.

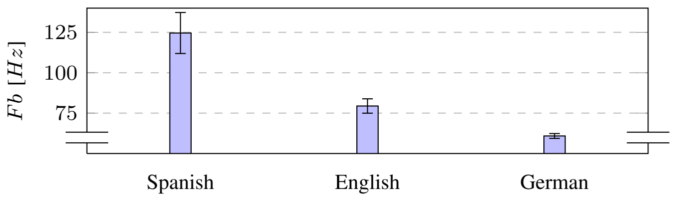

Fig. 7. Fb mean values, in Hz, estimated in the three corpora. Standard deviation is included as scattering measure.

The mean values and deviations of the parameters of phrase commands are shown in Fig. 8. $T0r$ is the relative position of phrase commands, defined as the distance to the event that gave its origin, as explained in Section II. For the three corpora, the new method has smaller amplitudes of commands: 58% in Spanish, 68% in English and 52% in German, on average. Furthermore, the amplitudes of phrase commands associated with a pause are much larger than the amplitudes of the inserted commands, i.e., the movements in the F0 contour of the intonation phrases that start after a break are more prominent than those without associated paused. The amplitudes of the inserted commands are 47%, 29% and 65% of the values obtained for the command amplitudes associated with pauses, for Spanish, German and English, respectively.

The mean values and deviations of the parameters of accent commands are shown in Fig. 9. $T1r$ is the relative position of accent commands, defined as the distance to the event that gave its origin, as explained in Section II, and $T2 - T1$ is the length of accent commands. In general, the amplitudes of accent commands are smaller with the new method compared with those obtained with A-ME and L-ME, but in different proportions for the three corpora and depending on the origin of the commands.

For Spanish, amplitudes of accent commands are approximately one-third small if they are associated with a stressed syllable, for both function or content words. Instead, accent commands associated with the end of an intonation phrase have an amplitude of approximately 50% with respect to A-ME and L-EM.

For English, the amplitudes of accent commands have a more erratic behavior depending on what gave rise to them. This can be explained in terms of the number of occurrences of each type of accent, see Fig. 5. Over all accent commands, the amplitude has an average value of 0.17, which is approximately 50% of the amplitudes obtained for A-ME and L-EM.

For German, average values of the amplitudes take more uniform values over all different types of accent commands. Only the commands associated with the end of intonation phrases, those that do not end in a pause, have amplitudes 29% lower than the other types of commands

Across all accent commands, the amplitude has an average value of 0.26, which is approximately 60% of the amplitudes obtained for A-ME and L-EM.

Taking into account the origin of each accent command, their lengths are similar in the three corpora and are shorter on average than those obtained with A-ME and L-ME. The

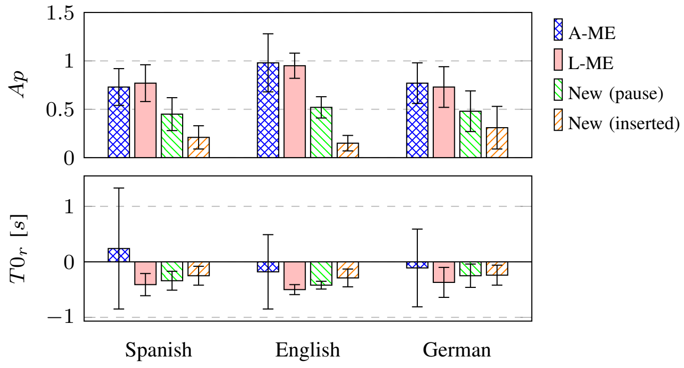

Fig. 8. Mean values and standard deviation of the estimated parameters for phrase commands. Algorithms: standard Mixdorff (A-ME); linguistically initialized (L-ME); and our automatic approach (New).

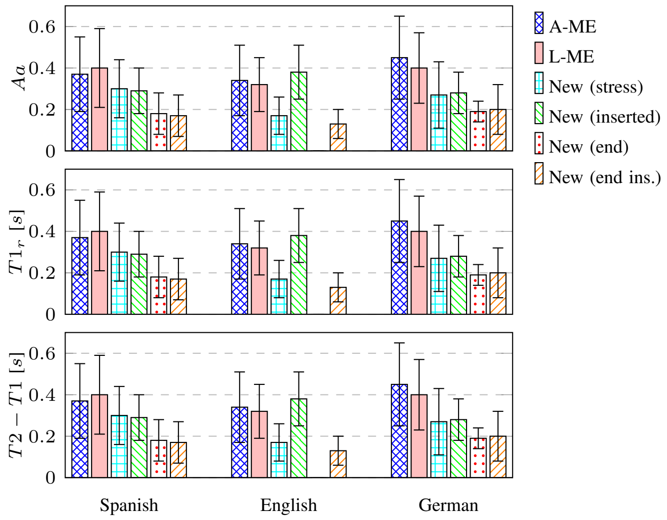

Fig. 9. Mean values and standard deviation of the estimated parameters for accent commands. Algorithms: standard Mixdorff (A-ME); linguistically initialized (L-ME); and our automatic approach (New).

commands associated with stressed syllables in the content words are the longest, and those that have origin at the end of intermediate sentences are the shortest.

In order to determine whether automatic implementation responds to the proposed algorithm, we compared the commands extracted by our manual and automatic approach, using an accuracy measure introduced in [25]. In a subset excerpt from Spanish corpus the command insertions present a full correspondence between both implementations.

IV. Conclusion and Future Works

In this paper, we have proposed a novel method for parameters extraction of Fujisaki's model. The new method can be performed manually, by a human labeler, or by an automatic implementation. The algorithm uses the structure of the sentence to obtain a first approximation of the command locations and a set of rules for insertion and local optimization of the commands.

Even though it was not the main objective, experiment results amply outperform other methods, confirming the assumptions raised about the relationship between text and commands. Automatic implementation results show improvements from 34% up 49% compared to other methods, in the analysis of three corpora: Spanish, English and German.

We believe that one of the great strengths of the present method is the correct insertion of phrase commands. Phrase command not only are boundaries of intermediate intonation phrases, but also influence the realization of accent commands. Furthermore, the proper extraction of accent commands tends to generate commands of lower amplitude and duration, in accordance of the energy-saving principle.

The new method performs a very simple preprocessing of F0 contour without either stylisation or filtering, avoiding the problems associated with this processing.

The A-ME method only requires the F0 contour to extract Fujisaki model parameters. In addition the L-ME method requires the position of pauses and the positions of the stressed syllables in function words. The approach presented in this paper also adds the need for labeling all words and their stressed syllables. Undoubtedly the manual labeling of words and stressed syllables is too expensive. But if the final aim is to analyze the relationship between Fujisaki model commands vs text and/or other suprasegmental information, labeling corpus will be needed anyway. A future work will be apply our method using the labels made with an automatic aligner, which would make our method fully automatic.

We are planning to analyze the relationship between Fujisaki's model parameters estimated with the method introduced here and the function of intonation.

Acknowledgment

The authors would like to thank the anonymous reviewers for their insightful feedback. This research was supported by Ministerio de Ciencia y Tecnología and Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina.

References

[1] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *Journal of Acoustic Society*, vol. 5, no. 4, pp. 233–242, 1984.

[2] P. S. Rossi, F. Palmieri, and F. Cutugno, "A method for automatic extraction of Fujisaki-model parameters," in *Proc. of Speech Prosody 2002*, B. Bel and I. Marlien, Eds., Aix-en-Provence, April 2002, pp. 615–618.

[3] B. Uslu and H. Ilk, "Fujisaki intonation model in Turkish Text-to-Speech Synthesis," in *Signal Processing and Communications Applications Conference, 2009. SIU 2009. IEEE 17th*, 2009, pp. 844–847.

[4] E. Navas and I. Hernadez., "Modelado de la entonación en Euskera utilizando el modelo de Fujisaki y árboles de regresión binarios," in *Resumenes de las I Jornadas de Tecnologías del Habla*, Sevilla, Spain, November 2000.

[5] H. Fujisaki, S. Narusawa, S. Ohno, and D. Freitas, "Analysis and modeling of f0 contours of Portuguese utterances based on the command-response model," in *Proc. of Eurospeech 2003*, vol. 3, Geneva, Italy, September 2003, pp. 2317–2320.

[6] H. Fujisaki, "Modeling in the study of tonal features of speech with application to multilingual speech synthesis," in *Proc. of Joint International Conference of SNLP and Oriental COCOSDA*, Hua-Hin, Thailand, May 2002, pp. D1–D10.

[7] H. Fujisaki, C. Wang, S. Ohno, and W. Gu, "Analysis and synthesis of fundamental frequency contours of Standard Chinese using the command-response model," *Speech Communication*, vol. 47, no. 1-2, pp. 59–70, September 2005.

[8] H. Fujisaki, S. Ohono, K. ichi Nakamura, M. Guirao, and J. Gurlekian, "Computational modeling of accent and intonation in declarative sentences in Spanish," *J. Acoust. Soc. Am.*, vol. 95, no. 5, p. 2949, 1994.

[9] K. Hirose, H. Fujisaki, and H.Kawai, "Generation of prosodic symbols for rule-synthesis of connected speech of Japanesel," in *Proc of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP86)*, vol. 4, Tokio, Japan, April 1986, pp. 2415–2418.

[10] H. M. Torres, "Generación automática de la prosodia para un sistema de conversión de texto a habla." Ph.D. dissertation, Universidad de Buenos Aires, Buenos Aires, Argentina, Agosto 2008.

[11] H. M. Torres and J. A. Gurlekian, "Parameter estimation and prediction from text for a superpositional intonation model," in *Proc of the 20 Konferenz Elektronische Sprachsignalverarbeitung*. TUDpress Verlag der Wissenschaften: TUDpress Verlag der Wissenschaften, September 2009, pp. 238–247.

[12] M. O'Reilly and A. N. Chasaide, "Analysis of intonation contours in portrayed emotions using the Fujisaki model," in *Proc. of the The Second International Conference on Affective Computing and Intelligent Interaction*, R. Cowie and F. de Rosis, Eds., Lisbon, 2007, pp. 102–109.

[13] P. Zervas, I. M. N. Fakotakis, and G. Kokkinakis, "Employing Fujisaki's intonation model parameters for emotion recognition," in *Advances in Artificial Intelligence*, ser. Lecture Notes in Computer Science, A. Grigoris, G. Potamias, C. Spyropoulos, and D. Plexousakis, Eds. Berlin: Springer, 2006, vol. 3955, pp. 443–453.

[14] H. Pfitzinger and H.Mixdorff, "Evaluation of F0 stylisation methods and Fujisaki-model extractors," in *Proc. of the 20 Konferenz Elektronische Sprachsignalverarbeitung*, Dresden, September 2009, pp. 228–237.

[15] H. Fujisaki, "Prosody, Information and Modelling with emphasis on Tonal Features of Speech," in *Proc. of Workshop on SLP*, Mumbai, India, January 2003, pp. 5–14.

[16] H. Mixdorff, "A novel approach to the fully automatic extraction of Fujisaki model parameters," in *Proc. of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00)*, vol. 3, Istanbul, June 2000, pp. 1281–1284.

[17] ——, "An integrated approach to modeling German prosody," PhD Thesis, Universitatsverlag, Dresden, 2002.

[18] S. Narusawa, N. Minematsu, K. Hirose, and H. Fujisaki, "A method for automatic extraction of model parameters from fundamental frequency contours of speech," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02).*, vol. 3, Istanbu, Turkey, May 2002, pp. 1281–1284.

[19] H. Fujisaki, S. Narusawa, and M. Maruno, "Preprocessing of fundamental frecuency countours of speech for automatic parameter extracction," in *Proc. of the 5th International Conference on Signal Processing (ICSP 2000)*, vol. 2, Beijing, China, August 2000, pp. 722–725.

[20] D. Hirst and R. Espesser, "Automatic modelling of fundamental frequency using a quadratic spline function," *Travaux de lInstitut de Phontique dAix*, vol. 15, pp. 75–85, 1993.

[21] J. Gutierrez-Arriola, J. Montero, D. Saiz, and J. Pardo, "New rule-based and data-driven strategy to incorporate Fujisaki's F0 model to a text to speech system in Castillian Spanish," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2001)*, vol. 2, Salt Lake City,USA, May 2001, pp. 821–824.

[22] V. Strom, "Detection of accents, phrase boundaries and sentence modality in German with prosodic features," in *In Proc.of European Conf. on Speech Communication and Technology*, 1995, pp. 2039–2041.

[23] H. Kruschke and M. Lenz, "Estimation of the parameters of the quantitative intonation model with continuous wavelet analysis." in *INTERSPEECH*, Geneva, Switzerland, September 1-4 2003.

[24] S. Silva and S. Netto, "Closed-form estimation of the amplitude commands in the automatic extraction of the Fujisaki's model," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '04*, vol. 1, May 2004, pp. 1621–1624.

[25] H. Kameoka, K. Yoshizato, T. Ishihara, K. Kadowaki, Y. Ohishi, and K. Kashino, "Generative modeling of voice fundamental frequency contours," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 6, pp. 1042–1053, June 2015.

[26] K. Hirose, Y. Furuyama, S. Narusawa, and N. M. H. Fujisaki, "Use of linguistic information for automatic extraction of F0 contour generation process model parameters," in *Proc. of the 8th European Conference on Speech Communication and Technology (EUROSPEECH 2003)*, Geneva, Italy, Sptember 2003, pp. 141–144.

[27] K. Hirose, Y. Furuyama, and N. Minematsu, "Corpus-based extraction of f0 contour generation process model parameters," in *Proc. of the INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal, September 2005, pp. 3257–3260.

[28] P. D. Agüero and A. Bonafonte, "Consistent estimation of Fujisaki's intonation model parameters," in *Proc. of 10th International Conference Speech and Computer (SPECOM 2005)*, Patras, Grecia, October 2005.

[29] H. M. Torres and J. A. Gurlekian, "Linguistically motivated parameter estimation methods for a superpositional intonation model," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, pp. 1–13, 2014.

[30] P. Prieto, *Las teorías lingüísticas de la entonación*. Barcelona: Ariel, 2003.

[31] J. A. Gurlekian, H. Rodriguez, L. Colantoni, and H. M. Torres, "Development of a prosodic database for an Argentine Spanish text to speech system," in *Proc. of the IRCS Workshop on Linguistic Databases*, B. Bird and Liberman, Eds. University of Pennsylvania, Philadelphia, USA: SIAM, December 2001, pp. 99–104.

[32] H. Fujisaki and S. O. C. Wang, "A command-response model for f0 contour generation in multilingual speech synthesis," in *Proc. of the 3rd ESCA/COCOSDA Intern. Workshop on Speech Synthesis*, G. Kokkinakis, N. Fakotakis, and E. Dermatas, Eds., Blue Mountains, Australia, November 1998, pp. 299–304.

[33] P.-E. Honnet, B. Gerazov, and P. N. Garner, "Atom decomposition-based intonation modelling," in *IEEE 40th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Australia, April 2015, pp. 4744–4748.

[34] R. Schubert, O. Jokisch, and D. Hirschfeld, "A modified parameterization of the fujisaki model," in *Proc. of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH-2010*, Makuhari, Chiba, Japan, September 2010, pp. 1796–1799.

[35] S. Rapp, "Automatisierte erstellung von korpora für die prosodieforschung," Ph.D. dissertation, Univ. Stuttgart, 1998, phD Dissertation.

[36] H. M. Torres, J. A. Gurlekian, and C. C. Mercado, "Aromo: Argentine Spanish TTS system," in *Proc. of. VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop, IberSPEECH 2012*, Madrid, Spain, November 2012, pp. 416–421.

[37] H. M. Torres, "Creación de un corpus de texto para la construcción de un sistema TTS," Laboratorio de Investigaciones Sensoriales, UBA-CONICET, Buenos Aires, Argentina, Informe técnico ISSN 0325-2043, Diciembre 2012.

[38] J. A. Gurlekian, L. Colantoni, and H. M. Torres, "El alfabeto fonético SAMPA y el diseño de córpora fonéticamente balanceados," *Fonoaudiológica*, vol. 47, no. 3, pp. 58–70, 2001.

Humberto M. Torres received the Bioengineer degree from Entre Ríos University, Oro Verde, Entre Ríos, Argentina in 1999, and a Ph.D. from Buenos Aires University, Buenos Aires, Argentina, in 2008.

He has been a researcher at the Laboratorio de Investigaciones Sensoriales, Consejo Nacional de Investigaciones Científicas y Técnicas, Buenos Aires, since 2009. He became an Assistant Professor in the Department of Biomedical at Buenos Aires University, in 2011. Since 1998, he has been working on speech technology.

Dr. Torres received the accolade of Best Paper at the 1998 Latin American Biomedical Engineering on MP4 Artificial Intelligence, and he was a finalist in 2014 Sadosky Awards.

Jorge A. Gurlekian received the Electronic Engineer degree in 1974 from Universidad Tecnológica Nacional, Buenos Aires, Argentina, and a Ph.D. from Buenos Aires University, Buenos Aires, Argentina, in 2012.

He was a visiting scientist at MIT-RLE during 1979-80 with a fellowship of the Scientific and Technological National Research Council of Argentina. Invited in 1985 by JICA at Tokyo, Osaka and Sapporo University for his specialization in Speech Communication and hired by contract at CRL in 1988 Osaka, ATR in 1989 and Doshisha University in 2002. He has been the director of the Laboratorio de Investigaciones Sensoriales, Consejo Nacional de Investigaciones Científicas y Técnicas, Buenos Aires, since 2009.

The Funprecit Foundation recognized his contributions in the field with the 2007 technological leader award.