EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Linguistically motivated parameter estimation methods for a superpositional intonation model

Humberto M Torres[1*], Jorge A Gurlekian[1], Hansjörg Mixdorff[2] and Hartmut Pfitzinger[3]

Abstract

This paper proposes two novel approaches for parameter estimation of a superpositional intonation model. These approaches present linguistic and paralinguistic assumptions for initializing a pre-existing standard method. In addition, all restrictions on the configuration of commands were eliminated. The proposed linguistic hypotheses can be based on either pitch accents or lexical stress, which give rise to two different estimation methods. These two hypotheses were validated by comparison of the estimation performance relative to two standard methods, one manual and one automatic. The results of the experiments for German, English and Spanish corpora show that the proposed methods outperform the standard ones.

Keywords: Superpositional F0 modeling; Automatic F0 contour estimation; Fujisaki model

1 Introduction

The Fujisaki model of intonation [1] has been tested for different languages [2-8], standing out for its simplicity and strong physiological basis. Currently, it is widely used in different application areas [9-13]. The model parameterizes F0 contours in an efficient way. With a small number of parameters, we can achieve a desired level of fitting accuracy. A task that has not been satisfactorily solved is the automatic model parameter extraction, that is, parameter estimation from F0 contours, since it is not directly reversable and hence there is no unique representation. As a general rule, if we desire a higher accuracy, we need to include more parameters in the model. However, if we aim at a set of parameters which will be linguistically interpretable, we invariably sacrifice accuracy in the F0 contour fitting, and estimation of parameter values must be done manually.

Several approaches have been proposed to estimate the model parameters [11,14,15]. One of the currently popular methods for parameter extraction, successfully tested for different languages, is the one proposed by Mixdorff [15] to which we will refer as the 'standard method'. Although this method is completely automatic, the author proposes a *post-hoc* manual correction to eliminate spurious commands which cannot be justified linguistically [16].

Pfitzinger and Mixdorff [17] discuss the accuracy of the current methods to estimate the model 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. The algorithm initialization is a critical issue, given that different sets of initial parameter values produce different model estimations with varying accuracy in fitting of the F0 contour.

Hirose et al. [18] have proposed to introduce linguistic information in the estimation process for a Japanese corpus. The model estimation is performed in two stages: First, an 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 [19], 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. These regression trees were created automatically from a portion of the corpus that was manually analyzed beforehand.

For Spanish, Torres y Gurlekian [11] proposed to consider linguistic aspects, such as the positions of pauses and syllables with lexical stress. The parameter values were

*Correspondence: hmtorres@conicet.gov.ar
[1]Laboratorio de Investigaciones Sensoriales, INIGEM, CONICET-UBA, Av. Córdoba 2351, 9 Piso Sala 2, 1120 Ciudad Autonoma de Buenos Aires, Argentina
Full list of author information is available at the end of the article

estimated by genetic algorithms (GA), achieving a performance similar to the standard method, but with too high algorithmic complexity. The ability to perform a global search in an n-dimensional solution space is the main justification of using GA, even more when the error surface has many local minima.

The purpose of this paper is to introduce linguistic information in the model estimation and also reduce the source of variability on the initialization of the parameter extraction method. Specifically, we introduce two new methods for parameter extraction, in which linguistic information is introduced to the method proposed by Mixdorff [15].

Our main idea is to initialize the estimation method with *a priori* information about the values of the model parameters. In the first case, this information is extracted from labels of tones and break indices, as proposed in the Tones and Break Index (ToBI) system [20]. In the second case, we take into account information on the locations of pauses and syllables with lexical stress. In this paper, we present the results of applying our estimation methods to German, English and Spanish speech databases.

This paper is organized as follows: In section 2, we present a brief introduction to the Fujisaki intonation model and the Mixdorff model estimation method. Our approaches to model estimation are presented in section 3. The speech databases used in this paper are introduced in section 4. The experiments and the results obtained are presented and discussed in section 5. Finally, the conclusions are discussed in section 6.

2 The Fujisaki model

This model - called superpositional - 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 [21]: a base frequency, accents and phrase components, as shown in Figure 1.

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 contour can be expressed by

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

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

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

where

- Fb: baseline value of fundamental frequency,
- 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

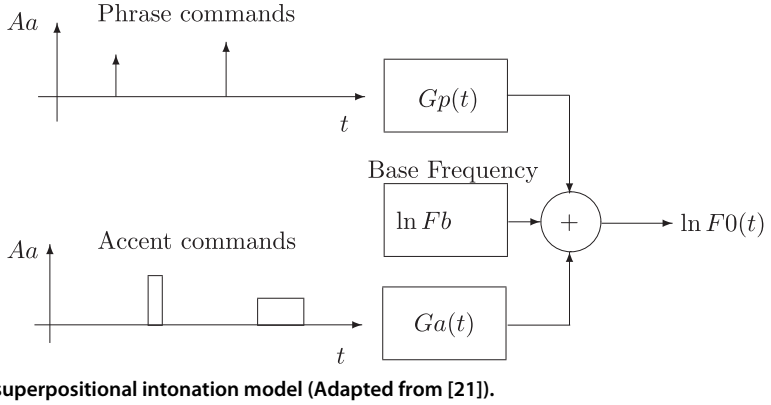

Figure 1 A scheme of the superpositional intonation model (Adapted from [21]).

control. Together with γ, they can be considered practically constant for all speakers. *Fb* must be estimated for each utterance, but is assumed to be constant for each speaker [21].

An example of Fujisaki model-based F0 contour parameterization is shown in Figure 2. We can see the slow F0 declination due to the phrase command, and local F0 movements triggered by the accent commands.

2.1 Mixdorff estimation method

Mixdorff presented an automated parameter extraction approach based on F0 measurements [15]. We will call this method *A-ME*, and it represents our baseline system. Below, we briefly present its main features.

After F0 contour interpolation and smoothing using Momel [22], the resulting spline contour is passed through a high-pass filter with a stop frequency at 0.5 Hz, similar to [23]. The output of the high-pass filter (henceforth referred to as 'high-frequency contour' or HFC) is subtracted from the spline contour yielding a 'low-frequency contour' (LFC), containing the sum of phrase components and *Fb*. The latter is initially set to the overall minimum of the LFC. Consecutive minima are detected in the HFC delimiting potential accent commands whose *Aa* is initialized to reach the maximum of *F*0 between the two minima. Since the onset of a new phrase command is characterized by a local minimum in the phrase component the LFC searches for local minima, applying a minimum distance threshold of 1 s between consecutive phrase commands. In order to initialize the magnitude value *Ap* assigned to each phrase command, the next local maximum is detected in the part of the LFC after the potential onset time *T*0. *Ap* is then calculated in proportion to F0 at that relative point while also considering contributions of preceding commands. The analysis-by-synthesis procedure is performed in three steps that are designed to optimize the initial parameter set iteratively by applying a hill-climb search to reduce the overall mean square error in the log frequency domain. At the first step, phrase and accent components are optimized separately, using respectively LFC and HFC as the targets. Next, phrase component, accent component and *Fb* are optimized together, with the spline contour as the target. In the final step, the parameters are fine-tuned by making use of a weighted representation of the extracted original *F*0 contour. The weighting factor applied is the product of degree of voicing and frame energy for every *F*0 value, hence favoring 'reliable' portions of the contour in vowel nuclei, for instance.

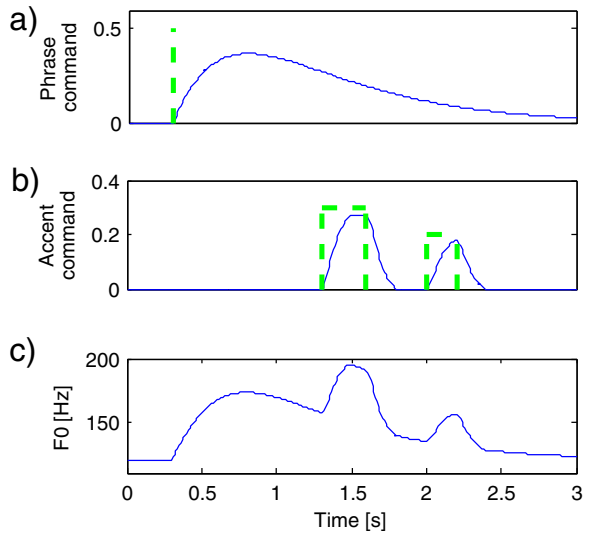

Figure 2 Example of Fujisaki model. (a) In green dashed lines, phrase command with $T0 = 0.3$, $Ap = 0.5$ and $\alpha = 2$, and its respective response in blue; **(b)** In green dashed lines, two accent command with $Aa = [0.3\ 0.2]$, $T1 = [1.3\ 2]$, $T2 = [1.6\ 2.2]$, $\beta = 20$, and $\gamma = 0.9$, and their response superimposed in blue; and **(c)** *F*0 model output, with $Fb = 120$ Hz.

In A-ME, the parameters β and γ are constants for each speaker, and their values are fixed *a priori*. *Fb* and α can be varied initially but will eventually be kept constant for a given speaker, for instance, by finding the median value of these parameters for the entire corpus. It should be stressed, however, that the Fujisaki model does not preclude optimizing these parameters for each command and each utterance in the corpus. We will call this method *A-ME+*.

Later, the automatic parameters can be manually corrected [16], in order to reduce inevitable misdetections and enforce linguistic criteria. We will call *M-ME* at this method with the hand-corrected parameters.

3 Linguistically motivated parameter estimation

In German, English and Spanish, as in many other languages, we can find two linguistic features: pitch accents and lexical stress. The pitch accents are associated with movements in the intonation contours and used to mark contrasts between different parts of a sentence [20,24]. Lexical stress are in contrast an intrinsic property of words. In general, it holds that the high tones in pitch accents are associated with lexical stress, but not vice versa.

In this work, we propose two modifications of Mixdorff's automatic method of parameter model extraction. One modification is related to the construction of a prototype for initialization, and the second modification is related to eliminate restrictions in model parameter values. The prototypes will be built from lexical stress or pitch accent information. They will be herein called lexically motivated or *L-ME* method, and tonally motivated or *T-ME* method.

The idea is to reduce parameters which are already considered speaker-dependent as much as possible and then to put the effort to estimate the remaining parameters. Those can be fixed in advance or limited to a certain range according to our linguistic hypotheses. Others will depend on higher level information, such as phrase type, intention, speaker mood, etc. Since this information is not available in advance, we will suppose that the values are only influenced by the text.

In addition, we removed some restrictions from A-ME, such as minimum or maximum values of command amplitudes, as well as minimum durations and distances between commands.

In summary, our hypothesis is that the model parameters will only depend on the text and that the speaker characteristics will remain invariant.

An outline of the proposed estimation methods are shown in Figure 3.

3.1 Lexically motivated estimation method

For the L-ME method, we made the following assumptions: First, the position of accent commands will be close to the location of syllable with lexical stress. We only considered stressed syllables in content words. Second, 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 [25]. Third, phrase commands will be near intonational phrase beginnings, as has been reported in previous studies [11,24].

To introduce these assumptions, we propose a *prototype* of initial model parameters, as follows:

- *Fb*, α, β and γ parameters are fixed for each corpus. These values are obtained from A-ME or M-ME.
- One phrase command per intonational phrase. The phrase command position is measured from the beginning of each intonational phrase, and we called it $T0_r$.
- One accent command for each syllable with lexical stress in content words. The position of this accent command is measured relative to location of the syllables with lexical stress. We called it $T1_r$.
- One accent command for each intermediate intonational phrase. The position of this accent command is measured relative to ending of intermediate intonational phrase. We called it $T1_r$.

3.2 Tonally motivated estimation method

For the T-ME method, we made the following assumptions: First, the position of accent commands will be close to the location of syllable with high-tone pitch accent. One way of labelling these events is through the ToBI system [20,24]. Second, it is reasonable to expect accent commands occurring at or near the end of intermediate intonational phrases, just as in the L-ME method. Third, phrase commands will be near intonational phrase beginnings, just as in the L-ME method.

For the T-ME method, we propose a *prototype* of initial model parameters, as follows:

- *Fb*, α, β and γ parameters are fixed for each corpus. These values are obtained from A-ME or M-ME.
- One phrase command per intonational phrase. The phase command position is measured from the beginning of each intonational phrase, and we called it $T0_r$.
- One accent command for each high tone in pitch accent. The position of this accent command is measured relative to location of syllables with high tone in pitch accent. We called it $T1_r$.
- One accent command for each intermediate intonational phrase. The position of this accent command is measured relative to ending of intermediate intonational phrase. We called it $T1_r$.

4 Speech material

The model prototypes were used as initial parameters of linguistic motivated methods L-ME and T-ME. The

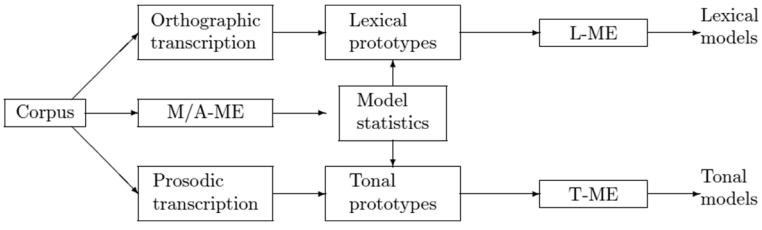

Figure 3 Outline of the two proposed methods. The *corpus* comprises wave files, transcription and phonetic F0 contours. A/M-ME means *Automatic or Manual Mixdorff Estimation method for parameter extraction*.The *orthographic transcription* (temporal alignment of pauses, part-of-speech and stressed syllables) is combined with the information extracted of models estimated by A/M-ME (mean value of *Ap*, $T0_r$, *Aa*, $T1_r$, $T2-T1$, α, β, γ and *Fb*), to obtain the *prototypes* for model initialization. With these prototypes we initialize the L-ME and we re-estimated the parameters of *lexical model*. In the same way, *prosodic transcription* (temporal alignment of pauses and syllables with tonal accents) is used to estimate the *tonal model*.

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

We tested our method in three different languages: German, English and Spanish. Below is a brief description of each corpus used in our experiments.

4.1 German database

For German, we used the IMS Radio News Corpus [26] which consists of German news texts read by professional speakers. The data were automatically segmented into phonemes according to the German SAMPA [27] inventory followed by manual corrections. Prosody of the data was manually labelled following the G-ToBI [28] conventions. The syllables with lexical stress also were manually labelled.

The reference data for the Fujisaki model were extracted automatically [15] and manually corrected following linguistic criteria [16] and using the interactive Fuji-ParaEditor [29]. Although raw F0 data are provided with the corpus extracted in 10 ms steps *via* get F0 of ESPS waves, a substantial correction was necessary. Our data selection comprises 73 news articles read by one male speaker.

4.2 English database

For English, we used the CSTR US KED Timit database [30] which contains 453 phonetically balanced utterances spoken by a US male speaker. The database was hand-labelled in phonemes, syllables and words, and carefully corrected. The syllables with lexical stress were also manually labelled.

For this corpus, we have neither labelling of prosody nor manually extracted intonation model parameters. Therefore, we only used information on lexical stress for estimating the parameter values.

4.3 Spanish databases

We use two databases in Spanish to test our models. Both were recorded by two professional female announcers, natives of Buenos Aires.

The first one, which we call DB1, was created with the aim to study prosody [24]. The corpus sentences had marked natural inflections and with different number of intonational phrases. Its text corpus consists of 741 declarative sentences extracted from Argentine newspapers published in Buenos Aires. The sentences contain 97% of all Spanish syllables, in both stress and unstressed conditions, and all possible syllabic positions within the word.

The second database was created to be used in a text-to-speech system, and we call it DB2. DB2 text corpus was based on DB1 supplemented with new sentences, in order to reach a broad coverage of diphones [31]. Also, we included 235 interrogative sentences. The corpus contains 1,826 sentences.

Recordings were made in a sound-proof chamber, with an AKG dynamic microphone and 16 Khz/16 bit sampling rate conversion. The speakers were instructed to read the sentences with natural tonal variations. The speech material collected was of approximately 40 min for DB1 and 140 min for DB2.

For the two databases, each sound file was manually labelled twice, by musically trained speech therapists who distinguished prosodic occurrences as intonational groups and accents [32]. The files were labelled on different tiers: phonetic according to Argentinian SAMPA [33], orthographic, break levels between words, and tonal marks according to an extended ToBI method for Argentine Spanish [24]. Part-of-speech and syntactic layers were also indicated.

Table 1 Parameter values used to build the prototypes for all models

Language	**German**	**German**	**German**	**English**	**Spanish**	**Spanish**	**Spanish**
Method	**T-ME**	**T-ME**	**L-ME**	**L-ME**	**T-ME**	**L-ME**	**L-ME**
Reference	**M-ME**	**A-ME**	**A-ME**	**A-ME**	**A-ME**	**A-ME**	**A-ME**
Corpus	**IMS**	**IMS**	**IMS**	**KDE**	**DB1**	**DB1**	**DB2**
α	0.95	2	2	2	2	2	2
β	20.3	20	20	20	20	20	20
Fb	50	50	50	60	100	100	100
Ap	1.13	0.75	0.75	1.0	0.8	0.8	0.75
$T0_r$	−0.1	−0.5	−0.5	−0.5	−0.5	−0.5	−0.5
Aa	0.3	0.4	0.4	0.34	0.4	0.4	0.33
$T1_r$	−0.045	−0.1	−0.1	−0.1	−0.15	−0.15	−0.1
$T2$-$T1$	0.24	0.3	0.3	0.3	0.25	0.3	0.25

Table 2 Results for manual and two automatic estimation approaches for German

	RMSE	*Ap* rates	*Aa* rates
M-ME	1.48 ± 0.19	0.44 ± 0.05	1.11 ± 0.09
A-ME	1.33 ± 0.14	0.44 ± 0.06	1.69 ± 0.1
A-ME+	1.33 ± 0.21	0.44 ± 0.05	1.10 ± 0.08

The RMSE is given in ST and the rates in commands per second. Standard deviation is included as scattering measure.

5 Results

5.1 Prototypes

Prototypes are built from the analysis of the reference models and linguistic restrictions explained in section 3. First, we have to obtain the reference models, M-ME or A-ME. Second, the mean values of the parameters are employed to build the prototype. Mean values are estimated over all parameter instances and for each analyzed database.

In Table 1, values used to build all the models analyzed in this work are shown. $T2$-$T1$ is the duration of accent commands. For both methods, the positions of commands, $T0_r$ and $T1_r$, were extracted from manual labelling, but there are alternatives for automatic labelling [34,35].

Finally, with these benchmarks, the prototypes for each sentence to be processed are built.

5.2 German

5.2.1 Estimation approaches for M-ME, A-ME and A-ME+

For German, we have M-ME- and A-ME- estimated models. The root mean square error (RMSE) in semitones (ST) and the average command density per second of the different experiments are shown in Table 2. We have also included the standard deviation as measured dispersion values. The two automatic models have the same performance, with an RMSE slightly lower than that of the manual. But A-ME+ models have meaningful lower numbers of accent commands than A-ME. We can assume that a higher number of free parameters in A-ME+ reduces accent command density.

The histograms of parameter values are shown in Figure 4. The relative position of the accent commands, $T0_r$ and $T1_r$, are measured as described in section 5.1. The histograms of M-ME and A-ME+ parameters are similar, with the exception of the Ap parameter, where the manual models presents higher value amplitudes. The A-ME models have accent commands with greater amplitude and duration than the others. Moreover, the phrase command amplitudes in the A-ME have less dispersion on their values. The temporal location of the phrase commands is similar for the three models, but the A-ME models have an extra peak in the histogram of the accent command positions, approximately at 0.25 s. This indicates the presence of accent commands that appear after the pitch accent.

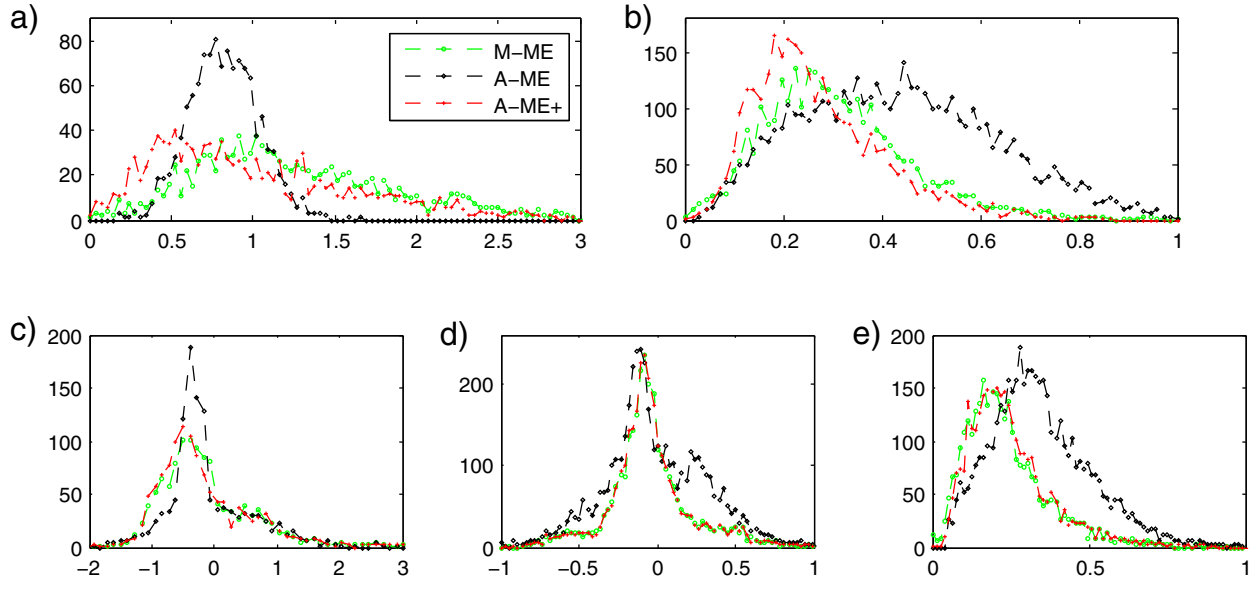

Figure 4 Histogram of amplitudes, relative positions and duration values of accent and phrase commands for German models. The methods were Mixdorff's manual (M-ME), automatic (A-ME) with global parameter optimization, and automatic+ (A-ME+) with local parameter optimization. **(a)** Phrase command amplitude, **(b)** accent command amplitude, **(c)** phrase command relative position, **(d)** accent command relative position, and **(e)** accent command duration.

Table 3 Results for the tonal method with two initialization methods for German

	RMSE	*Ap* rates	*Aa* rates
M-ME	1.48 ± 0.19	0.44 ± 0.05	1.11 ± 0.09
A-ME	1.33 ± 0.14	0.44 ± 0.06	1.69 ± 0.1
T-ME (by M-ME)	1.46 ± 0.27	0.49 ± 0.06	1.27 ± 0.13
T-ME (by A-ME)	1.50 ± 0.21	0.49 ± 0.06	1.35 ± 0.14
T-ME Prototype		0.49 ± 0.06	1.53 ± 0.16

The RMSE is given in ST and the rates in commands per second. Standard deviation is included as scattering measure.

In what follows, we will only use the A-ME, where the values of Fb, α, β and γ shall be considered constants for all sentences of each speaker. Keeping these values constant enables us to compare the distributions of the amplitudes, location and duration of the model commands.

5.2.2 T-ME

As mentioned above, there are two possible ways of creating prototypes to initialize the automatic method: (1) from parameters estimated automatically (A-ME) or (2) corrected manually (M-ME). We explore the two alternatives for the tonally based method, without appreciable differences in results.

Results are shown in Table 3. Information from the prototypes generated to initialize the methods are also added. Results for two initialization approaches are comparable to the manual method, and with a performance slightly lower than the automatic estimation method.

As can be seen, the density of phrase commands remains unchanged. Moreover, the accent command density is similar for both alternatives of initialization, but they are lower than that of the prototypes and the automatic approach, and are higher than those obtained from the manual. Since the models obtained with the proposed method have a lower density of accent commands than the prototypes, the optimization algorithm removes those commands which are considered unnecessary. This is a very interesting property of the method, as it reduces model redundancy.

The histograms of parameter values are shown in Figure 5. Something interesting to note is that the two new tonal model histograms apparently match the histogram of its prototype. Moreover, both peaks correspond to the values assigned by the respective prototypes. We can assume that the method of parameter estimation determined that these values are a good approximation, and does not modify their values in the process of parameter setting.

In Figure 5e, we can note the presence of accent commands of very short duration. In the original version, there is a restriction of minimum accent duration which is allowed, that we have eliminated here.

In view of the results obtained, we create initialization prototypes taking as references the models from the automatic estimation.

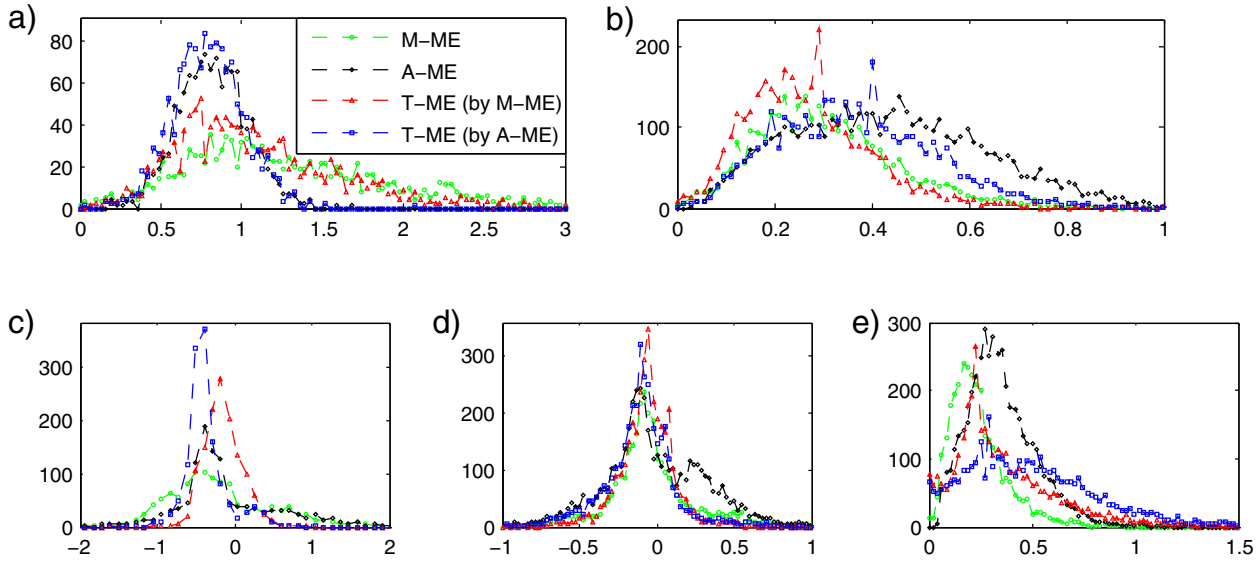

Figure 5 Histogram of amplitudes, relative positions and duration values of accent and phrase commands for tonal German models. The methods were Mixdorff's manual (M-ME), automatic (A-ME) and tonal with Mixdorff's manual (T-ME by M-ME) and automatic (T-ME by A-ME) initialization. **(a)** Phrase command amplitude, **(b)** accent command amplitude, **(c)** phrase command relative position, **(d)** accent command relative position, and **(e)** accent command duration.

Table 4 Results for the lexical method on a reduced data set of German corpus

	RMSE	*Ap* rates	*Aa* rates
M-ME	1.81 ± 0.12	0.48 ± 0.05	1.12 ± 0.06
A-ME	1.48 ± 0.15	0.38 ± 0.05	1.65 ± 0.11
T-ME	1.75 ± 0.24	0.50 ± 0.05	1.42 ± 0.13
L-ME	1.59 ± 0.23	0.50 ± 0.05	1.63 ± 0.17
L-ME Prototype		0.50 ± 0.05	1.72 ± 0.16

The RMSE is given in ST and the rates in commands per second. Standard deviation is included as scattering measure.

5.2.3 L-ME

The use of lexical stress was another alternative to create the prototype initialization of the parameter estimation method. The occurrence of an accent command is associated with a lexical stress, as described in section 5.1.

The results of the experiments with this approach are shown in Table 4. When we gathered these results, only a small number of sentences labelled with lexical stress was available. Therefore, our results are restricted to this subset. For this reason, we include the results of the other estimation methods on the same subset of sentences in this table. The lexical method gives better results than the manual and the tonal, but worse than the automatic. The density of the accent commands is similar to the automatic method, but much higher than the manual. Again, we can also observe how the proposed method significantly reduces the number of original accent commands, by removing the unnecessary commands from the prototypes.

In Figure 6, the histograms of command parameter values for this approach are shown. Its behavior is similar to that observed in Figure 5 for tonal models.

Figure 7a shows a sentence example of fundamental frequency generated from the Mixdorff methods, manual and automatic, and the methods presented in this paper, tonal and lexical. We have also included the syllabic phonetic labelling and the original F0 contour for comparison. In Figure 7b,c,d,e, we can see the accent and phrase commands inserted by the four methods. Also, the RMSE in ST of the whole sentence for each method is shown. It is interesting to note that different settings of commands generate an acceptable parametrization of the original F0 contour. We can also observe that the different models take into account some movements of the F0 contour, while smoothing out others.

5.3 English

For the English corpus, we explored the lexical method. In this case, we fixed *a priori* the *Fb*, α, β and γ parameter values for automatic model estimation. The manual labelling of phonemes in content words and syllables with lexical stress were available in the database to build the prototype of intonation model, in a similar fashion to the previous case of the lexical German models. Here, we do not have index pauses between words, so we assigned a phrase command after each pause. This is a very rough first approximation, given that many sentences of this

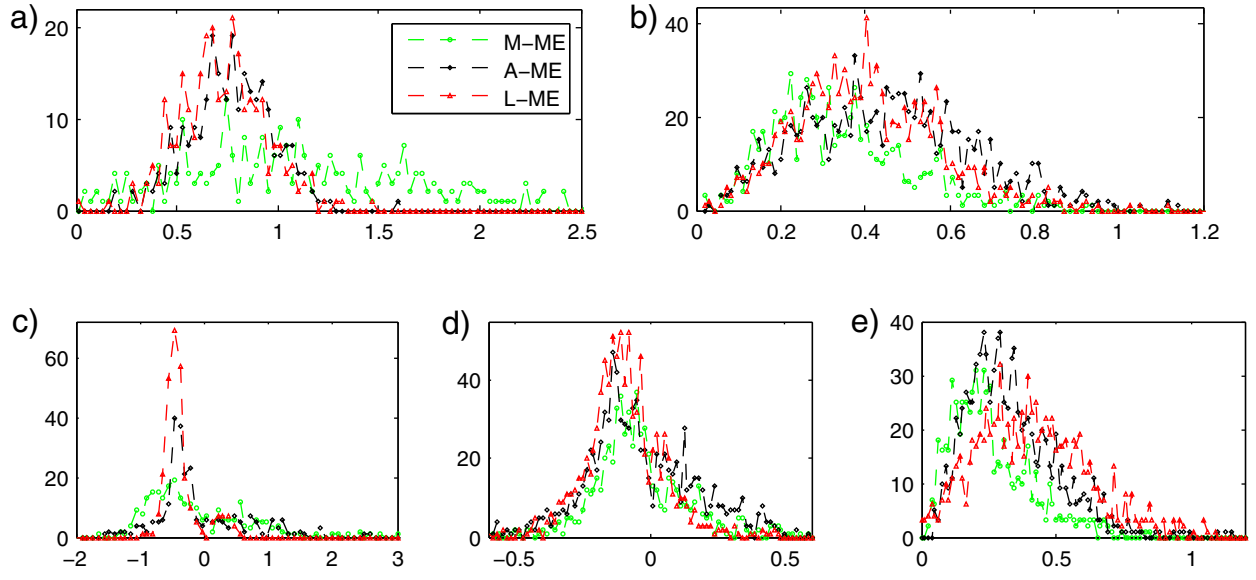

Figure 6 Histogram of amplitudes, relative positions and duration values of accent and phrase commands for lexical German models. Mixdorff's manual (M-ME), automatic (A-ME) and lexical (L-ME) methods with Mixdorff's automatic initialization. **(a)** Phrase command amplitude, **(b)** accent command amplitude, **(c)** phrase command relative position, **(d)** accent command relative position, and **(e)** accent command duration.

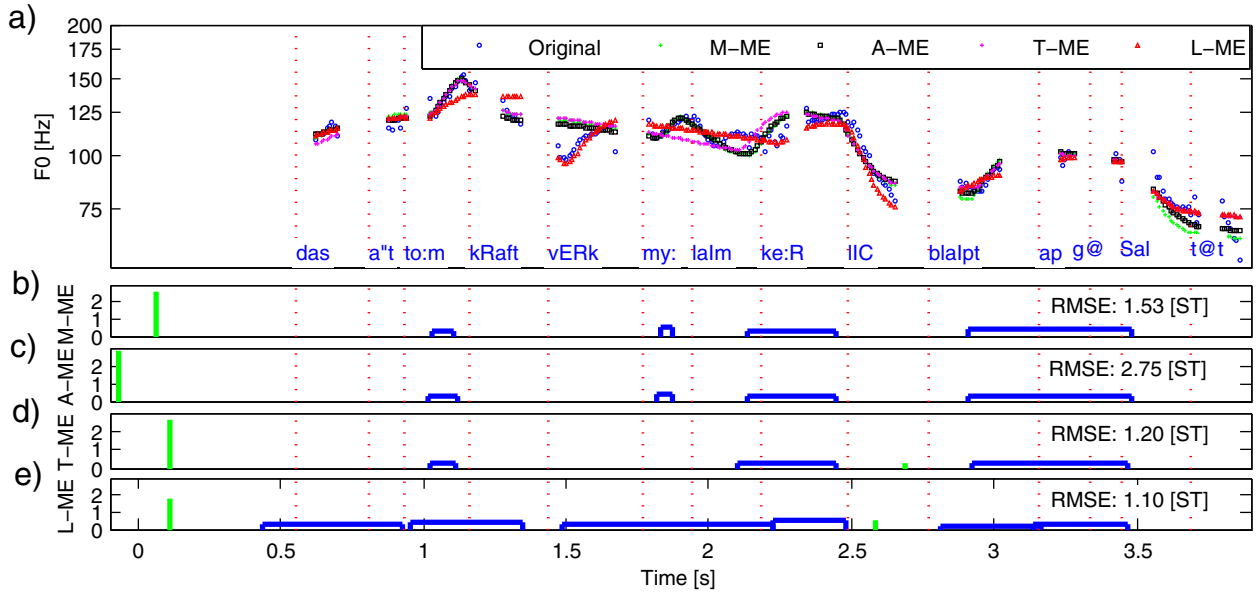

Figure 7 Example of F0 contour estimation of the four methods for German. (a) F0's and phonetic syllable transcription. **(b** and **c)** Commands for Mixdorff's manual (M-ME) and automatic (A-ME) methods. **(d** and **e)** Commands for tonal (T-ME) and lexical (L-ME) methods of estimation. RMS error in ST for a whole sentence were included. The sentence corresponds to a portion of a long sentence, and for the text *'Das Atomkraftwerk Mühlheim-Kehrlich bleibt abgeschaltet'* (*'The nuclear power plant Mühlheim-Kehrlich stays shut down'*).

corpus, between pause and pause have at least two intonation phrases.

The results are shown in Table 5. The results on the English corpus are better than for German. The lexical approach has an error slightly lower than the automatic model at the expense of a higher density of accent commands. As stated above, a high density of both phrase and accent commands improves the fitting of the F0 contour. Moreover, in most cases, English content words are mainly shorter, monosyllabic and disyllabic, than German and Spanish, where content words have two or three syllables on average [36,37]. Unlike the German case, the estimation method does not eliminate many accent commands from the English prototypes. Five percent of the accent commands from the prototypes are removed in lexical German models, and only 1% are removed in the English models.

In Figure 8, the histograms of command parameter values for this approach are shown. Histograms of the automatic and lexical parameters are similar. We can see the low dispersion of the positions of the command phrase which suggests taking a fixed value for this parameter.

Figure 9 shows an example of fundamental frequency generation. We have included the results of command prototypes and A-ME commands for comparison. In this figure, we can see how the search algorithm optimizes the initial command prototypes to obtain a better parametrization of the F0 contour.

Table 5 Results for the lexical method for English

	RMSE	*Ap* rates	*Aa* rates
A-ME	1.02 ± 0.49	0.68 ± 0.26	1.91 ± 0.53
L-ME	0.88 ± 0.34	0.57 ± 0.15	3.02 ± 0.64
L-ME Prototype		0.57 ± 0.15	3.05 ± 0.61

The RMSE is given in ST and the rates in commands per second. Standard deviation is included as scattering measure.

5.4 Spanish

For Spanish, we built the tonal models with the DB1 database and lexical models with the DB1 and DB2 databases. The prototypes for the DB1 corpus were constructed similarly to the German corpus, and prototypes for DB2 were built in a similar way as the English corpus. For both cases, we also built the reference A-ME models.

The experiment results are shown in Table 6. For DB1, the tonal model has a similar performance to the automatic model and the lexical approach has a slightly better performance of about 6% than T-ME and A-ME. The phrase command density is the lowest for the automatic model, approximately 40% lower than the proposed models. The tonal model has the lowest density of accent commands, and the lexical model has the highest density. A high number of commands improve the approximation of the models to the original F0 contours. We must also emphasize that the methods proposed removes approximately 15% of accent command prototypes.

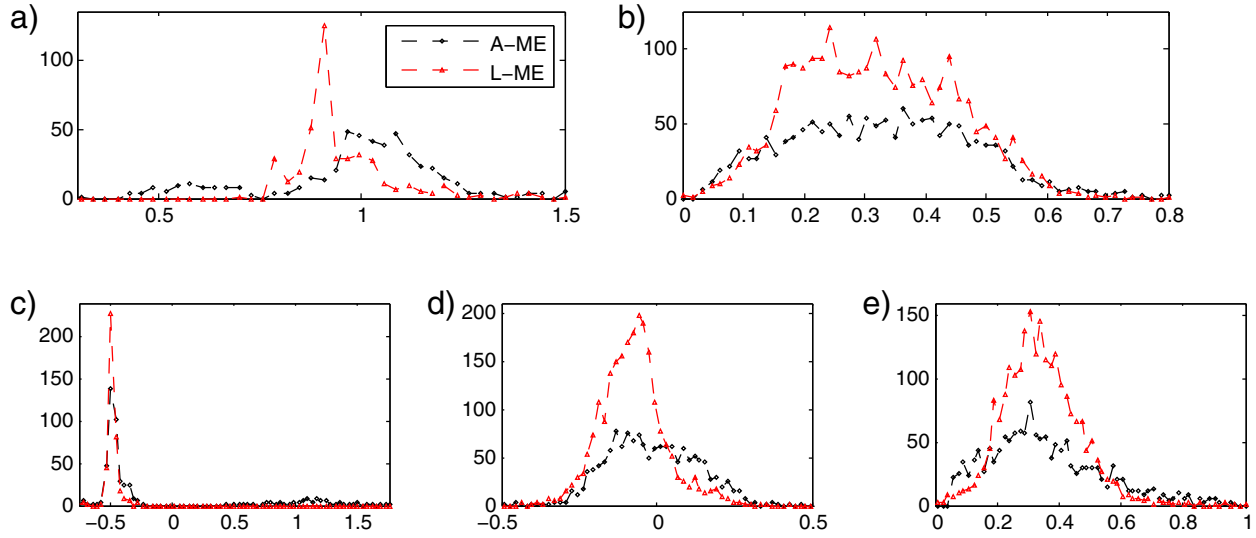

Figure 8 Histogram of amplitudes, relative positions and duration values of accent and phrase commands for English models. Mixdorff's automatic (A-ME) and lexical (L-ME) methods. **(a)** Phrase command amplitude, **(b)** accent command amplitude, **(c)** phrase command relative position, **(d)** accent command relative position, and **(e)** accent command duration.

For the DB2 database, lexical models have a performance lower than automatic approach. Unlike other corpora analyzed in this paper, however, the lexical models have a density of commands lower than the automatic models: 40% fewer phrase commands and 14% fewer accent commands.

The large difference in the density of phrase commands between the two corpora is explained by the fact that, in DB1, we assign one command for each intonation phrase, and in DB2, we assign one command for each pause.

In Figure 10, the histograms of command parameter values are shown. For DB2, the phrase command

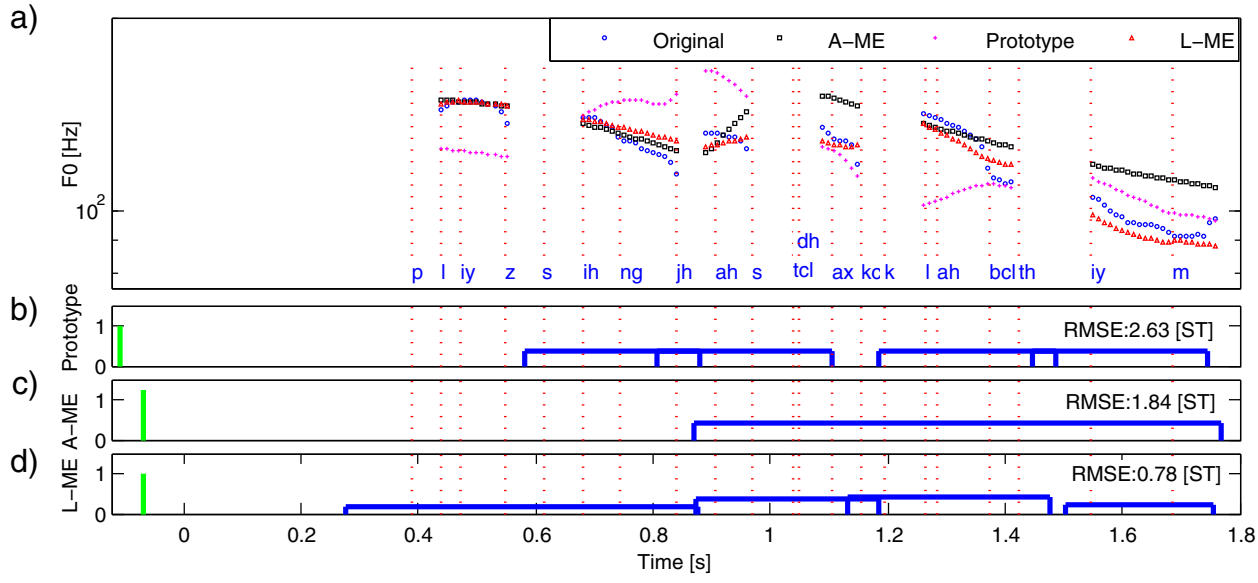

Figure 9 Example of F0 contour estimation for English. (a) F0's and phonetic syllable transcription. **(b)** Prototype commands, **(c)** commands for Mixdorff's automatic (A-ME) method, and **(d)** commands for the lexical (L-ME) method of estimation. RMS error in ST for a whole sentence was included. The text sentence is*'Please sing just the club theme'*.

Table 6 Results for DB1 and DB2 Spanish databases

		RMSE	*Ap* rates	*Aa* rates
DB1	A-ME	1.36 ± 0.47	0.61 ± 0.21	1.84 ± 0.35
	T-ME	1.32 ± 0.45	1.04 ± 0.20	1.57 ± 0.40
	L-ME	1.28 ± 0.47	1.04 ± 0.20	1.99 ± 0.51
	T-ME Prototype		1.04 ± 0.20	1.89 ± 0.36
	L-ME Prototype		1.04 ± 0.20	2.35 ± 0.43
DB2	A-ME	1.15 ± 0.41	0.74 ± 0.25	1.91 ± 0.41
	L-ME	1.49 ± 0.43	0.45 ± 0.22	1.65 ± 0.53
	L-ME Prototype		0.45 ± 0.22	1.76 ± 0.53

The RMSE is given in ST and the rates in commands per second. Standard deviation is included as scattering measure.

amplitudes are slightly higher than in the DB1. Furthermore, for the DB2 database, the accent command durations of lexical models is slightly higher than those of automatic models. The distributions for other parameters are similar for all models. Again, Figure 6e reveals accent commands of short duration.

Figure 11 shows an example of fundamental frequency generated. This sentence was taken from DB2. In this figure, we can see how the automatic method inserts greater amount of phrase commands than the lexical method. We can also see that in the second part of the sentence, approximately from 2.7 to 3.8 s, the automatic method introduces an accent command of long duration, which simplifies two tonal movements in the F0 contour.

6 Discussion and conclusions

The main contribution of the methods proposed in this paper is a successful link between *tonal movements* and both *pitch accent* and *lexical stress* through *accent commands*. This mutual relationship could be used in the design of an automatic tonal tagger.

The results from this study confirm our hypotheses and our assumptions about the location of the commands. Furthermore, this approach has proved to work in three different languages, with different word lengths and different syntactic structures.

The use of initial prototypes permits a better estimation of F0 contours than that obtained in a fully free approach. Prototypes can be seen as initial conditions that delimit the number of commands. Our methods assign a command for each linguistic event and also take care of removing those commands that are not necessary. Preliminary studies show that this is due to a large percentage of low boundary tones for which there is no need to insert an accent command. Another reason is the proximity between prototype accents that may lead to fusion or elimination of one of them [38]. More work should be carried out to elucidate which command accent prototypes are erased, in order to omit them in the initial prototype of our algorithm.

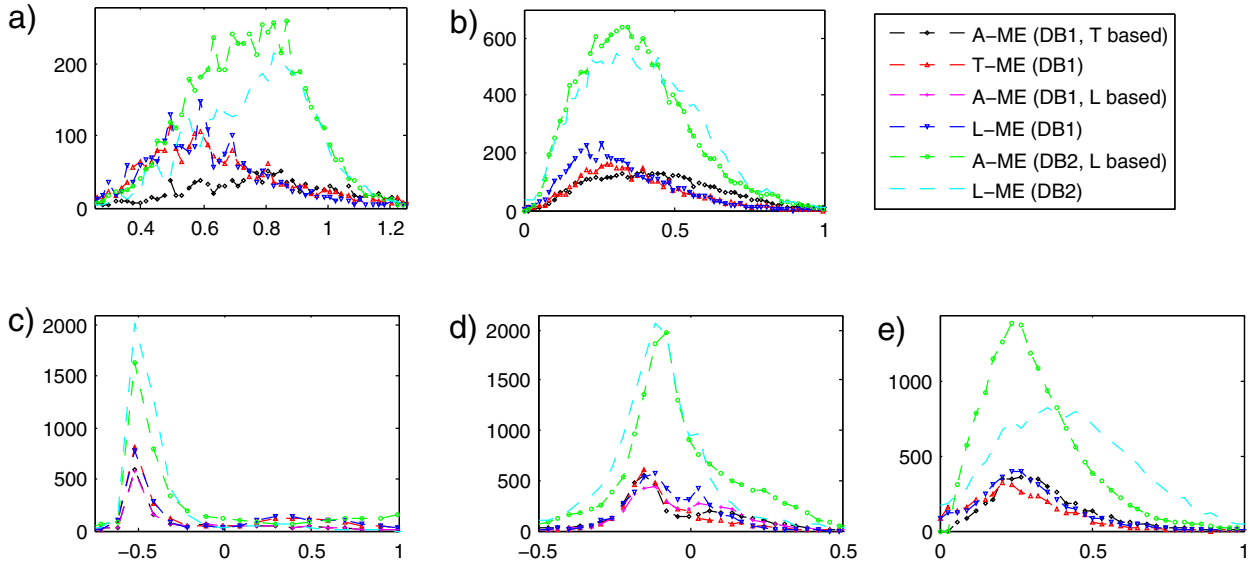

Figure 10 Histogram of amplitudes, relative positions and duration values of accent and phrase commands for Spanish models. Mixdorff's automatic method based on pitch accent over DB1 corpus (A-ME, DB1, T based), tonal method over DB1 corpus (T-ME, DB1), Mixdorff's automatic method based on lexical stress over DB1 corpus (A-ME, DB1, L based), lexical method over DB1 corpus (L-ME, DB1), Mixdorff's automatic method based on lexical stress over DB2 corpus (A-ME, DB2, L based), lexical method over DB2 corpus (L-ME, DB2). **(a)** Phrase command amplitude, **(b)** accent command amplitude, **(c)** phrase command relative position, **(d)** accent command relative position, and **(e)** accent command duration.

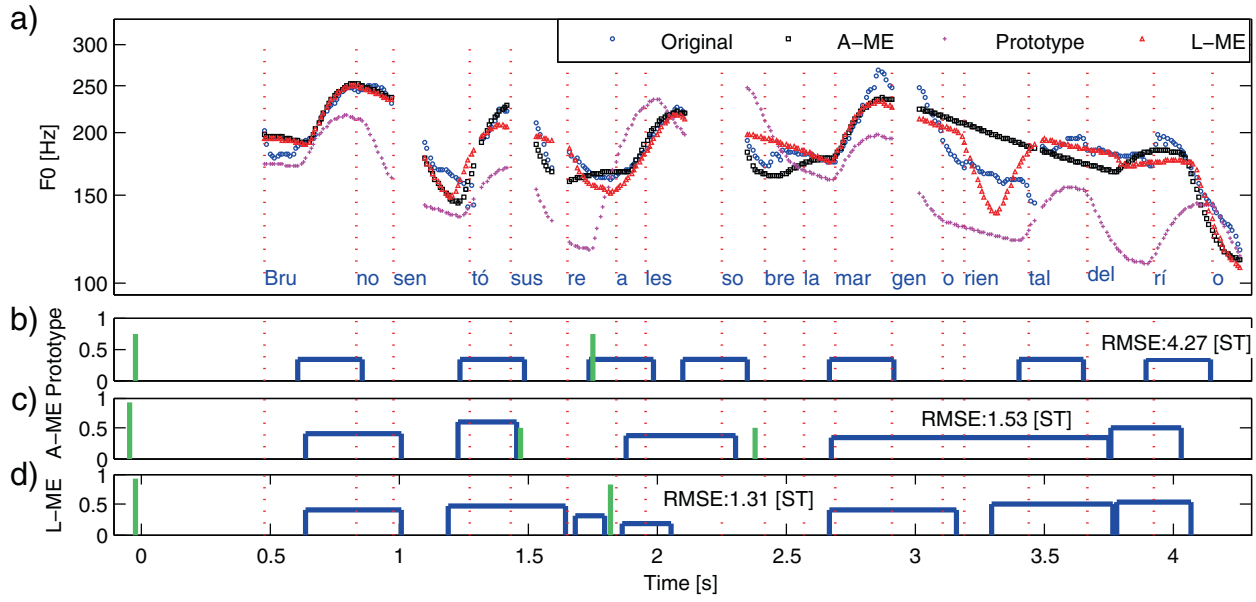

Figure 11 Example of F0 estimation for Spanish. (a) F0's and phonetic syllable transcription. **(b)** Prototype commands, **(c)** commands for Mixdorff's automatic method, and **(d)** commands for lexical method of estimation. RMS error in ST for a whole sentence were included. The text sentence is *'Bruno sentó sus reales sobre la margen oriental del río'* (*'Bruno sat on his royals over the eastern bank of the river'*).

When we compare the histograms of the accent command durations, our approaches versus the standard methods, we can see a significantly increased amount of accent commands of very short duration. These short accents are manifested as tiny movements of F0 contour, with an amplitude lower than 1.5 semitones, a value taken as minimum difference in order to be perceived as a prominence [39,40]. In the future, we should consider the role that these small commands play in the Fujisaki model, and eventually modify our algorithm to suppress accent commands that are modeling micro-prosody.

We have lifted restrictions on the amplitudes, overlaps, minimum/maximum parameter values of commands that were imposed by the original Mixdorff algorithm. We believe that physiological studies should be conducted in order to determine the existence of such restrictions and what values they take in case they do occur.

The histograms of the phrase command relative positions have a very sharp peak at −0.5 s, for both methods and for all corpora analyzed. This T0 value generates a peak in the F0 contour at the intonation phrase onset. In future studies, we will evaluate the possibility of maintaining this parameter as a constant.

The performance of the methods to fit the F0 contour in each of the languages analyzed cannot be compared, due to differences in the sentence structures of each corpus. In all cases, however, their performances are equal to or better than those obtained with the reference models.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This research has been carried out with the support of Ministerio de Ciencia y Tecnología and Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina.

Author details

[1]Laboratorio de Investigaciones Sensoriales, INIGEM, CONICET-UBA, Av. Córdoba 2351, 9 Piso Sala 2, 1120 Ciudad Autonoma de Buenos Aires, Argentina. [2]Department of Computer Sciences and Media, Beuth University Berlin, Berlin, Germany. [3]Pfitzinger Voice Design, Lübeck, Germany.

Received: 11 December 2013 Accepted: 17 June 2014
Published: 15 July 2014

References

1. H Fujisaki, K Hirose, Analysis of voice fundamental frequency contours for declarative sentences of Japanese. J. Acoust. Soc. **5**(4), 233–242 (1984)
2. PS Rossi, F Palmieri, F Cutugno, ed. by B Bel, I Marlien, A method for automatic extraction of Fujisaki-model parameters, in *Proceedings of Speech Prosody 2002* (Aix-en-Provence, 11–13 April 2002), pp. 615–618
3. B Uslu, HG Ilk, Fujisaki intonation model in Turkish text-to-speech synthesis, in *The IEEE 17th Signal Processing and Communications Applications Conference, 2009. SIU 2009* (Antalya, Turkey, 844). doi:10.1109/SIU.2009.5136528
4. E Navas, 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, 2000)
5. H Fujisaki, S Narusawa, S Ohno, D Freitas, Analysis and modeling of F0 contours of Portuguese utterances based on the command-response model, in *Proceedings of Eurospeech 2003*, vol. 3 (Geneva, Switzerland, 1–4 September 2003), pp. 2317–2320
6. H Fujisaki, Proceedings of Joint International Conference of SNLP and Oriental COCOSDA (Hua-Hin, Thailand, 9–11 May 2002), pp. 1–10

7. H Fujisaki, C Wang, S Ohno, W Gu, Analysis and synthesis of fundamental frequency contours of standard Chinese using the command-response model. Speech Commun. **47**(1–2), 59–70 (2005)
8. H Fujisaki, S Ohono, K-i Nakamura, M Guirao, J Gurlekian, Computational modeling of accent and intonation in declarative sentences in Spanish. J. Acoust. Soc. Am. **95**(5), 2949 (1994)
9. K Hirose, H Fujisaki, H Kawai, Generation of prosodic symbols for rule-synthesis of connected speech of Japanese, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP86)*, vol. 4 (Tokyo, Japan, 7–11 April 1986), pp. 2415–2418
10. HM Torres, *Generación automática de la prosodia para un sistema de conversión de texto a habla,* PhD thesis, (Universidad de Buenos Aires, Buenos Aires, Argentina, Agosto 2008)
11. HM Torres, JA Gurlekian, Parameter estimation and prediction from text for a superpositional intonation model, in *Proceedings of the 20 Konferenz Elektronische Sprachsignalverarbeitung* (TUDpress Verlag der Wissenschaften, 2009), pp. 238–247
12. M O'Reilly, AN Chasaide, ed. by R Cowie, F de Rosis, Analysis of intonation contours in portrayed emotions using the Fujisaki model, in *Proceedings of the The Second International Conference on Affective Computing and Intelligent Interaction* (Lisbon, Portugal, 12–14 September 2007), pp. 102–109
13. P Zervas, IMN Fakotakis, G Kokkinakis, ed. by A Grigoris, G Potamias, C Spyropoulos, and D Plexousakis, Employing Fujisaki's intonation model parameters for emotion recognition, in *Advances in Artificial Intelligence.* Lecture Notes in Computer Science, vol. 3955 (Springer Berlin, 2006), pp. 443–453
14. S Narusawa, N Minematsu, K Hirose, H Fujisaki, A method for automatic extraction of model parameters from fundamental frequency contours of speech, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, vol. 1 (Orlando, Florida, 13–17 May 2002), pp. I-509–I-512
15. H Mixdorff, A novel approach to the fully automatic extraction of Fujisaki model parameters, in *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00)*, vol. 3 (Istanbul, Turkey, 5–9 June 2000), pp. 1281–1284
16. H Mixdorff, *An integrated approach to modeling German prosody.* Phd thesis, (Universitatsverlag, Dresden, Germany, 2002)
17. H Pfitzinger, H Mixdorff, Valuation of F0 stylisation methods and Fujisaki-model extractors, in *Proceedings of the 20 Konferenz Elektronische Sprachsignalverarbeitung* (Dresden, Germany, 21–23 September 2009), pp. 228–237
18. K Hirose, Y Furuyama, S Narusawa, NMH Fujisaki, Use of linguistic information for automatic extraction of F0 contour generation process model parameters, in *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH 2003)* (Geneva, Switzerland, 1–4 September 2003), pp. 141–144
19. K Hirose, Y Furuyama, N Minematsu, Corpus-based extraction of F0 contour generation process model parameters, in *Proceedings of the INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology* (Lisbon, Portugal, 4–8 September 2005), pp. 3257–3260
20. M Beckman, G Ayers, Guidelines for ToBI labelling, (Ohio State University, Version 3.0, 1997). http://www.ling.ohio-state.edu/~tobi/ame_tobi/labelling_guide_v3.pdf. Accessed 1 July 2013
21. H Fujisaki, Prosody, information and modelling with emphasis on tonal features of speech, in *Proceedings of Workshop on SLP* (Mumbai, India, 9–13 December 2003), pp. 5–14
22. D Hirst, R Espesser, Automatic modelling of fundamental frequency using a quadratic spline function. Travaux de l'Institut de Phonétique d'Aix. **15**, 75–85 (1993)
23. V Strom, Detection of accents, phrase boundaries and sentence modality in German with prosodic features, in *Proceedings of the European Conference on Speech Communication and Technology* (Madrid, Spain, 18–21 September 1995), pp. 2039–2041
24. JA Gurlekian, H Rodriguez, L Colantoni, HM Torres, ed. by BBird Liberman, Development of a prosodic database for an Argentine Spanish text to speech system, in *Proceedings of the IRCS Workshop on Linguistic Databases* (SIAM, University of Pennsylvania Philadelphia, USA, 2001), pp. 99–104
25. P Prieto, *Las Teorías Lingüísticas de La entonación*, (Ariel, Barcelona, 2003)
26. S Rapp, *Automatisierte erstellung von korpora für die prosodieforschung,* PhD thesis, (Univ. Stuttgart, 1998). PhD Dissertation
27. D Gibbon, German SAMPA. (Dokumentation V1.0. Bielefeld, 1995). http://coral.lili.uni-bielefeld.de/Documents/sampa-d-vmlex.html. Accessed 1 April 2013
28. M Grice, S Baumann, Deutsche intonation und gtobi. Linguistische Ber. **191**, 267–298 (2002)
29. H Mixdorff, FujiParaEditor (2009). http://public.beuth-hochschule.de/~mixdorff/thesis/fujisaki.html. Accessed 1 March 2013
30. Cstr us ked timit database (2001). http://www.festvox.org/dbs/dbs_kdt.html. Accessed 1 December 2013
31. H Torres, Creación de un corpus de texto para la construcción de un sistema tts. Informe técnico ISSN 0325-2043, Laboratorio de Investigaciones Sensoriales, UBA-CONICET, (Buenos Aires, Argentina, Diciembre 2012)
32. L Colantoni, J Gurlekian, Convergence and intonation: historical evidence from Buenos Aires Spanish. Bilingualism: Lang. Cogn. **7**(2), 107–119 (2004)
33. JA Gurlekian, L Colantoni, HM Torres, El alfabeto fonético SAMPA y el diseño de córpora fonéticamente balanceados. Fonoaudiológica. **47**(3), 58–70 (2001)
34. A Rosenberg, J Hirschberg, Detecting pitch accent using pitch-corrected energy-based predictors, in *Proceedings of Interspeech* (Antwerp, Belgium, 27–31 August 2007), pp. 2777–2800
35. H Torres, J Gurlekian, ed. by B Bel, I Marlien, Automatic determination of phrase breaks for Argentine Spanish, in *Proceedings of the Speech Prosody 2004 (SP-2004),* (Nara, Japan, 23–26 March 2004), pp. 553–556
36. M Guirao, MAG Jurado, *Estudio Estadístico del Español*. (CONICET, Buenos Aires, 1993)
37. KJ Kohler, *Einführung in die Phonetik des Deutschen*. (Erich Schmidt Verlag, Berlin, 1977)
38. G Toledo, J Gurlekian, Choque de acentos tonales frente al fraseo. Revista Philologica Romanica. **11**, 43–66 (2011)
39. ACM Rietveld, C Gussenhoven, On the relation between pitch excursion size and prominence. J. Phon. **13**, 299–308 (1985)
40. AP Bertrán, AMF Planas, EM Celdrán, AO Escandell, MCA Céspedes, ed. by JD García, Umbrales tonales en español peninsular, in *Actas del II Congreso de Fonética Experimental* (Sevilla, 2002), pp. 272–278

doi:10.1186/s13636-014-0028-3
Cite this article as: Torres *et al.*: **Linguistically motivated parameter estimation methods for a superpositional intonation model.** *EURASIP Journal on Audio, Speech, and Music Processing* 2014 **2014**:28.