Two new estimation methods for a superpositional intonation model

Humberto M. Torres , Hansjörg Mixdorff , Jorge A. Gurlekian , Hartmut R. Pfitzinger

Laboratorio de Investigaciones Sensoriales, Hospital de Clínicas, UBA, Argentina
Deparment of Computer, BHT Berlin University of Applied Sciences, Germany
Inst. of Phonetics and Digital Speech Processing, Christian-Albrechts-University, Germany
hmtorres@conicet.gov.ar, jag@fmed.uba.ar, mixdorff@tfh-berlin.de, hpt@ipds.uni-kiel.de

10.21437/Interspeech.2010-9

Abstract

This work presents two new approaches for parameter estimation of the superpositional intonation model for German. These approaches introduce linguistic and paralinguistic assumptions allowing the initialization of a previous standard method. Additionally, all restrictions on the configuration of accents were eliminated. The proposed linguistic hypotheses can be based on either tonal or lexical accent, which gives rise to two different estimation methods. These two kind of hypotheses were validated by comparison of the estimation performance relative to two standard methods, one manual and one automatic. The results show that the proposed methods far exceed the performance of the automatic method and are slightly beyond the manual method of reference.

Index Terms: Text-To-Speech, Superpositional F0 modeling, Automatic F0 contour estimation, Fujisaki model.

1. Introduction

The Fujisaki model of intonation [1] has been tested for different languages, standing out for its simplicity and strong physiological basis. Currently, it is widely used in different application areas [2][3][4]. A task that has not been satisfactorily resolved is the model's parameter extraction, that is the parameter estimation from intonation contours, since it is not directly invertible. One of the currently popular methods, successfully tested for different languages, is the one proposed by Mixdorff [5]. While this method is completely automated, the author proposes an additional manual correction to eliminate spurious and linguistic nonsense accents [6].

Pfitzinger [7] discusses the accuracy of such different methods that estimate model parameters solely on the basis of the original F0 contours. In previous work [2] it was proposed to consider linguistic aspects, such as the positions of pauses and lexical accents. In this paper, we introduce two new methods for parameter extraction, in which linguistic information is implemented into the method proposed by Mixdorff [5].

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: a base frequency ( ), tonal accents, and phrase accents. Phrase accents are calculated as the response to a second order linear filter critically excited with a delta function called phrase command. Tonal accents result from the response to the same filter, excited with a step function called accent command.

The parameters and in the equations of the Fujisaki model characterize the dynamic properties of the laryngeal mechanisms of phrase and accent control. Together with they can be considered practically constant for all speakers. must be estimated for each emission. The final parameters to be calculated are the occurrence of phrase commands, amplitude and position values of the phrase accents (and), amplitude and position values of tonal accents (, and).

2.1. Mixdorff model estimation method

Mixdorff presented an automated parameter extraction approach based on F0 measurements [5]. We will call this method *automatic*, and it represents our baseline system:

After F0 contour interpolation and smoothing using Momel, the resulting spline contour is passed through a high-pass filter with a stop frequency at 0.5 Hz, similar to [8]. 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 . The latter is initially set to the overall minimum of the LFC. Consecutive minima are detected in the HFC delimiting potential accent commands whose is initialized to reach the maximum of 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 assigned to each phrase command, the next local maximum is detected in the part of the LFC after the potential onset time . 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 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 contour. The weighting factor applied is the product of degree of voicing and frame energy for every value, favoring 'reliable' portions of the contour.

Later, the *automatic* parameters were manually corrected [6], in order to reduce inevitable misdetections. We will call this method *manual* during the reminder of the paper.

3. Linguistically motivated parameter estimation

In this work we propose two modifications of Mixdorff's automatic method. The idea is to reduce the speaker dependent parameters as much as possible and to then estimate the remaining parameters, presumably associated with the text structure. Those can be fixed in advance or limited in range according to our linguistic hypothesis. Others will depend on upper level information, such as phrase type, intentionality, speaker mood, etc. Since this information is not available in advance, we will suppose that the values are only influenced by the text. In summary, our hypothesis is that model parameters will only depend on the text and that the speaker characteristics will remain invariable.

With this in mind, the following assumptions can be made. First, the position of accent commands will be close to the location of a accents. Second, it is reasonable to expect accent commands occurring near the end of intermediate phrases; these are called "boundary tones" in the ToBI system. Third, phrase accents will be near intonational phrase beginnings, as has also been reported in other studies [2].

To enter these assumptions, we created a prototype of initial model parameters, as follows:

, , and parameters are fixed and obtained from a manual or automatic method.

One phrase command per intonational phrase: Its position is taken as the average relative distance to the beginning of each intonational phrase; the average is obtained over all sentences, i.e. the entire database.

One accent command for each accent, and one at the end of each intonational phrase. Its position is associated with an accent, and is taken as the average relative distance to each accent. Commands related to a juncture accent are taken as the average relative to the end of each intonational phrase; the average is obtained over all sentences, i.e. the entire database.

In addition, we removed some restrictions from the original method, such as minimum or maximum values of amplitudes, as well as minimum durations and distances between commands.

3.1. Tonal and lexical accents

In German, as in many other other languages, we can find two different types of accents: tonal and lexical. The tonal accents are associated with movements in the intonation contours and used to mark contrasts between different parts of a sentence. One way of labeling these events is through the Tone and Break Indices (ToBI) system [9]. The lexical accents are in contrast an intrinsic property of words. In general, it holds that the tonal accents are associated with lexical stress, but not vice versa.

As above, we will have two possible sets of accent commands for each sentence. In this paper we will explore these two alternatives.

The positions of both types of accents were extracted from manual labeling, but there are alternatives for automatic labeling [10] [11].

4. Speech material and experiments

The experiments are based on the IMS Radio News Corpus [12] which consists of German news texts read by professional speakers. The reference data for the Fujisaki model was extracted automatically [5] and manually corrected following linguistic criteria [6] and using the interactive Fuji-ParaEditor [13]. 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 adding up to 48 minutes of speech, of which 1,670 seconds or 167,039 F0 frames were voiced.

The two new methods were both developed on the basis of this corpus. The model prototypes were used as initial parameters of Mixdorff's automatic method. The phrase and accent command amplitudes and positions produced by the four Fujisaki model extractors as well as , and were used to resynthesise the F0 values by means of the Fujisaki model, which is defined in the log F0 domain. Thus, our evaluation uses a semitone scale.

5. Results

5.1. Automatic vs manual initialization

As mentioned above, there are two possible ways of creating prototypes to initialize the automatic method: from parameters estimated automatically or corrected manually. We explore the two alternatives for tonal based method, without appreciable differences, and with results comparable to the manual method.

The difference in performance was 0.03 ST, as can be seen in Table 1. In this table, we can see the root mean square error (RMSE) in semitones (ST) and the average command density per second of the different experiments. We have also included the standard deviation as measured dispersion values.

Table 1: *Results for the tonal method with two initialization methods. The RMSE is given in ST and the rates in commands per seconds. Standard deviation is included as scattering measure.*

	RMSE		Ap rates		Aa rates	
Manual	1.48	0.19	0.44	0.05	1.12	0.08
Automatic	1.88	0.45	0.45	0.05	1.11	0.09
Tonal (Manual)	1.45	0.28	0.49	0.06	1.29	0.14
Tonal (Automatic)	1.42	0.25	0.49	0.06	1.30	0.13
Tonal Prototype			0.49	0.06	1.53	0.16

This small difference between the performances of both approaches can be explained by analyzing the histograms of Fig. 1 and Fig. 2. The first and second rows of Figures 1 and 2 show the histograms of the parameter values estimated by the Mixdorff methods, manual and automatic respectively. Third and fourth rows show the histogram for the proposed method using tonal accents for initialization as reference, when we used statistics from manual and automatic methods respectively. As we can see, there is no appreciable difference in the histograms of the two new approaches.

In Table 1, we have also added information from the prototypes generated to initialize the methods.

As can be seen the density of phrase commands remains unalterable.

The density of accent commands on the other hand is similar for both alternatives of initialization, but is lower than that of the prototype initialization. We can assume that the method does not rule out those accents that are useful.

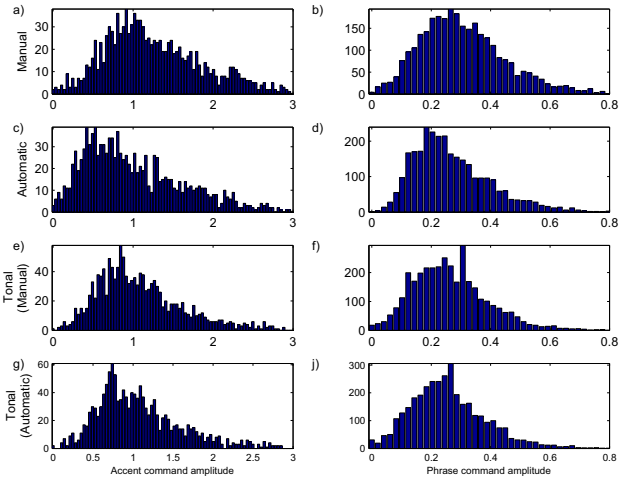

Figure 1: *Histogram of accent and phrase amplitude command values for four models. a) to b) Mixdorff's manual commands. c) to d) Mixdorff's automatic commands. e) to f) Tonal's commands with Mixdorff manual initialization. g) to j) Lexical's commands with Mixdorff automatic initialization.*

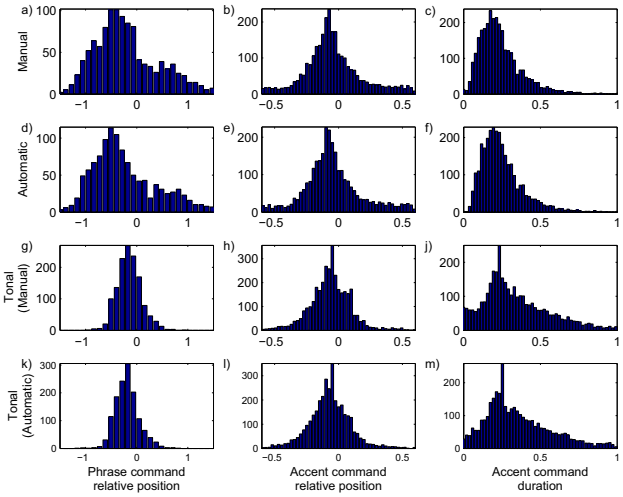

Figure 2: *Histogram of relative positions and durations values of accent and phrase command for four models. a) to c) Mixdorff's manual commands. d) to f) Mixdorff's automatic commands. g) to j) Tonal's commands with Mixdorff manual initialization, and k) to m) with Mixdorff automatic initialization.*

5.2. Lexical accent

Another alternative was to use the lexical accents to create the prototype initialization of the parameter estimation method. The results of the experiments with this approach are shown in Table 2. When we were gathering these results, only a small number of sentences labeled with lexical accent was available. Therefore our results are restricted to the given subset of data. For this reason, we include the results of the other estimation methods on the same subset of sentences in this table.

Table 2: *Results for the lexical method on a reduced data set. The RMSE is given in ST and the rates in commands per seconds. Standard deviation is included as scattering measure.*

	RMSE		Ap rates		Aa rates	
Manual	1.72	0.16	0.47	0.06	1.15	0.07
Automatic	2.21	0.27	0.48	0.05	1.14	0.07
Tonal	1.81	0.35	0.50	0.05	1.59	0.18
Lexical	1.65	0.32	0.50	0.05	1.51	0.16
Lexical Prototype			0.50	0.05	1.72	0.16

Figure 3a) shows an example of fundamental frequency generated from Mixdorff methods, manual and automatic, and the methods presented in this paper, tonal and lexical. We have also included the syllabic phonetic labeling and the original intonation contour for comparison. In the Figures 3b) to 3e) we can see the accent and phrase commands inserted by the four methods. Also shown is the RMSE in ST of the whole sentence for each method.

6. Discussion

The biggest advantage of the model used, and at the same time its biggest disadvantage, is that there are no restrictions on the values of its parameters. This allows us to select them in such a way that they fit the real intonation contour with the desired accuracy. In this study we linked the tonal movement, read as oscillations in the contour of the fundamental frequency generated by model commands, with tonal and lexical accents.

While our methods assign a command for each tonal or lexical accent and one for each juncture, the method takes 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 of the prototype accents that may lead to the elimination of one of them. Yet, in comparison with other reference methods, our methods still have a higher accent density [7].

7. Conclusions and future work

The results from this study confirm our hypotheses about the location of the commands.

Moreover, as we have demonstrated a double bond between *lexical accents - accent commands - tonal accents*, we can use this information to assist in an automatic tonal tagger.

In this work we have lifted restrictions on the amplitudes, overlaps, minimum/maximum parameter values of commands that were imposed on the original method. 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. This is possible because the model is based on the anatomy and physiology of the vocal apparatus.

Finally it is also important to adapt and test the hypotheses for other languages than German.

8. 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.

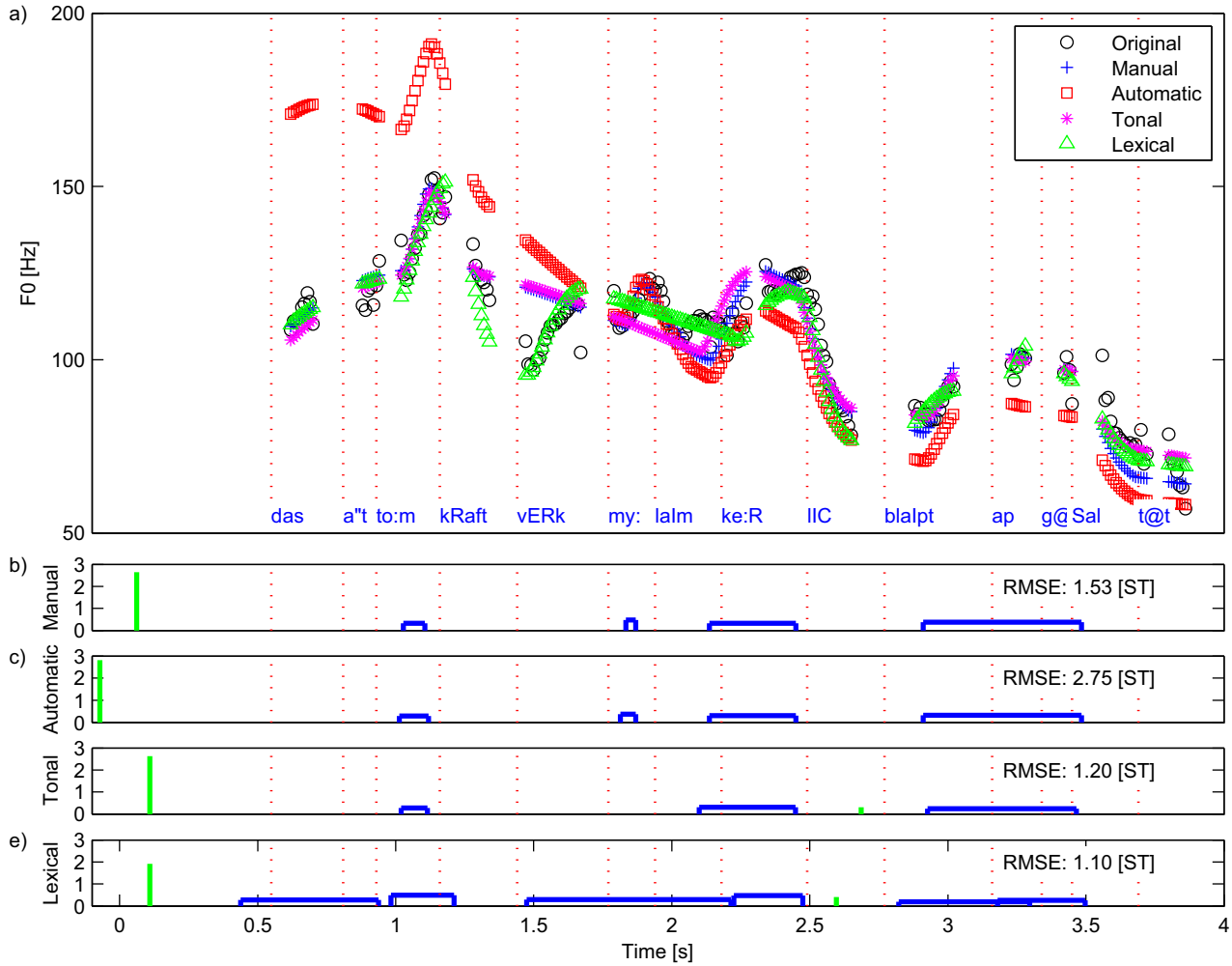

Figure 3: *Example of F0 estimation of the four methods. a) F0's and phonetic syllable transcription. b) and c) Commands for Mixdorff manual and automatic methods. d) and e) Commands for new tonal and lexical method of estimation. RMS error in ST for a whole sentence were included. The sentence corresponds to a portion of a bigger sentence, and for the text* "Das Atomkraftwerk Mühlheim-Kehrlich bleibt abgeschaltet" *(*"The nuclear power plant Mühlheim-Kehrlich stays shut down"*).*

9. References

[1] Fujisaki, H. and Hirose, K., "Analysis of voice fundamental frequency contours for declarative sentences of Japanese", Journal of Acoustic Society, 5(4):233-242, 1984.

[2] Torres, H. and Gurlekian, J., "Parameter estimation and prediction from text for a superpositional intonation model", In Proc. of the 20 Konferenz Elektronische Sprachsignalverarbeitung, 238-247, Dresden, 2009.

[3] Zervas, P., Mporas, I., Fakotakis, N., and Kokkinakis, G., " Employing Fujisakis Intonation Model Parameters for Emotion Recognition", In Advances in Artificial Intelligence, 443-453, Springer Berlin/Heidelberg, 2006.

[4] O'Reilly, M., and N Chasaide, A. "Analysis of intonation contours in portrayed emotions using the Fujisaki model", In Cowie, R., de Rosis, F. (eds), Doctoral Consortium, The Second International Conference on Affective Computing and Intelligent Interaction, Lisbon, 102-109, 2007.

[5] Mixdorff, H., "A novel approach to the fully automatic extraction of Fujisaki model parameters", In Proc. of ICASSP, 3:12811284, Istanbul, 2000.

[6] Mixdorff, H., "An Integrated Approach to Modeling German Prosody", w.e.b Universitätsverlag, Dresden, 2002.

[7] Pfitzinger, H. and Mixdorff, H., "Evaluation of F0 stylisation methods and Fujisaki-model extractors", In Proc. of the 20 Konferenz Elektronische Sprachsignalverarbeitung, 228-237, Dresden, 2009.

[8] Strom, V., "Detection of accents, phrase boundaries and sentence modality in German with prosodic features", In Proc. of EUROSPEECH'95, 3:2039-2041, Madrid, 1995.

[9] Beckman, M.; Ayers Elam, G., "Guidelines for ToBI labeling", Version 3. Ohio State University, 1997.

[10] Rosenberg, A. and Hirschberg, J., "Detecting Pitch Accent Using Pitch-corrected Energy-based Predictors", In Proc. of Interspeech,2777-2800, Antwerp, 2007.

[11] Torres, H. and Gurlekian, J., "Automatic Determination of Phrase Breaks for Argentine Spanish", In Proc. of 2nd International Conference on Speech Prosody, 553-555, Nara, 2004.

[12] Rapp, S., "Automatisierte Erstellung von Korpora für die Prosodieforschung. Arbeitspapiere (phonetikAIMS), 4(1):1-167, Inst. für Maschinelle Sprachverarbeitung, Lehrstuhl füexperimentelle Phonetik der Univ. Stuttgart.

[13] Mixdorff, H., FujiParaEditor: http://www.tfh-berlin.de/ mixdorff/thesis/fujisaki.html. TFH Berlin University of Applied Sciences, 2009.