A NOVEL APPROACH TO THE FULLY AUTOMATIC EXTRACTION OF FUJISAKI MODEL PARAMETERS

Hansjörg Mixdorff

Dresden University of Technology
Mommsenstr. 13
01062 Dresden, Germany

ABSTRACT

The generation of naturally-sounding F0 contours in TTS is important for the intellegibility and perceived naturalness of synthetic speech. In earlier works the author developed a linguistically motivated model of German intonation based on the quantitative Fujisaki model of the production process of F0. The extraction of parameters for this model from the extracted F0 contour, however, poses problems since model components are superimposed in a particular contour and cannot be calculated directly. The current paper introduces a novel fully-automatic multi-stage approach which was applied to a larger speech database of German. After explaining the modeling procedure in detail, the paper presents first results concerning the relationship between the Fujisaki parameters and the linguistic information underlying an utterance.

1. INTRODUCTION

The generation of naturally-sounding F0 contours is an important issue crucially influencing the intelligibility and perceived naturalness of synthetic speech. In earlier studies by the author a model of German intonation was developed which uses the quantitative Fujisaki-model of the production process of F0 [1] for parametrizing a given F0 contour. The contour is described as a sequence of linguistically motivated tone switches, major rises and falls, which are modeled by onsets and offsets of accent commands connected to accented syllables. Prosodic phrases correspond to the portion of the F0 contour between consecutive phrase commands [2]. The model was integrated into a German TTS system and proved to produce a high naturalness compared with other approaches [3]. The main attraction of the Fujisaki-model lies in the fact that it offers a physiological interpretation connecting F0 movements with the dynamics of the larynx, a viewpoint not inherent in any other of the currently used intonation models which mainly aim at breaking down a given F0 contour into a sequence of 'shapes' [4][5].

The direct estimation of parameters for the Fujisaki-model from the extracted F0 contour, however, poses problems since its components are superimposed in a particular contour and difficult to be inferred directly. Furthermore, determining the appropriate number of model commands underlying a given F0 contour requires a trade-off between fitting accuracy and linguistic meaningfulness. As a consequence, methods for determining model parameters were either limited to short utterances [6] or required user interaction flawing the objectiveness of the analysis. An earlier approach developed by the author made use of ToBI-labels in a prosodically labeled speech data base for determining and aligning the necessary number of model commands [7]. The current paper introduces a novel multi-stage approach consisting of a quadratic spline smoothing, contour filtering, accent command initialization and a three-pass Analysis-by-Synthesis procedure. The corpus used in the study is part of a German corpus compiled by the Institute of Natural Language Processing, University of Stuttgart and consists of 48 minutes of news stories read by a male speaker [8].

The corpus contains boundary labels on the phone, syllable and word levels and linguistic annotations such as part-of-speech. Furthermore it is supplied with ToBI-labels following the Stuttgart System [9]. The F0 values are provided for intervals of 10 ms, along with frame-wise energy- and degree-of-voicing-measures. The latter are used for weighting the F0 contour in the final phase of the modeling procedure.

2. MODELING PROCEDURE

2.1 Quadratic Spline Stylisation

Prior to modeling a given F0 contour, two tasks are performed: (1) Intermediate F0 values for unvoiced speech segments and short pauses are interpolated from the extracted F0 contour, (2) Microprosodic variations caused by the influence of individual speech sounds (plosion, friction, etc.) are smoothed out, as the Fujisaki model explicitly deals with macroprosody only. One method successfully applied to the two tasks mentioned is the MOMEL model [10] which converts a given F0 contour into a sequence of target points used as a reference for performing a spline interpolation of the contour. It has been shown that MOMEL can be applied regardless of the particular language.

Figure 1 (top) shows the initial part of an utterance from the database displaying the extracted (+ signs) and the spline contours (solid line).

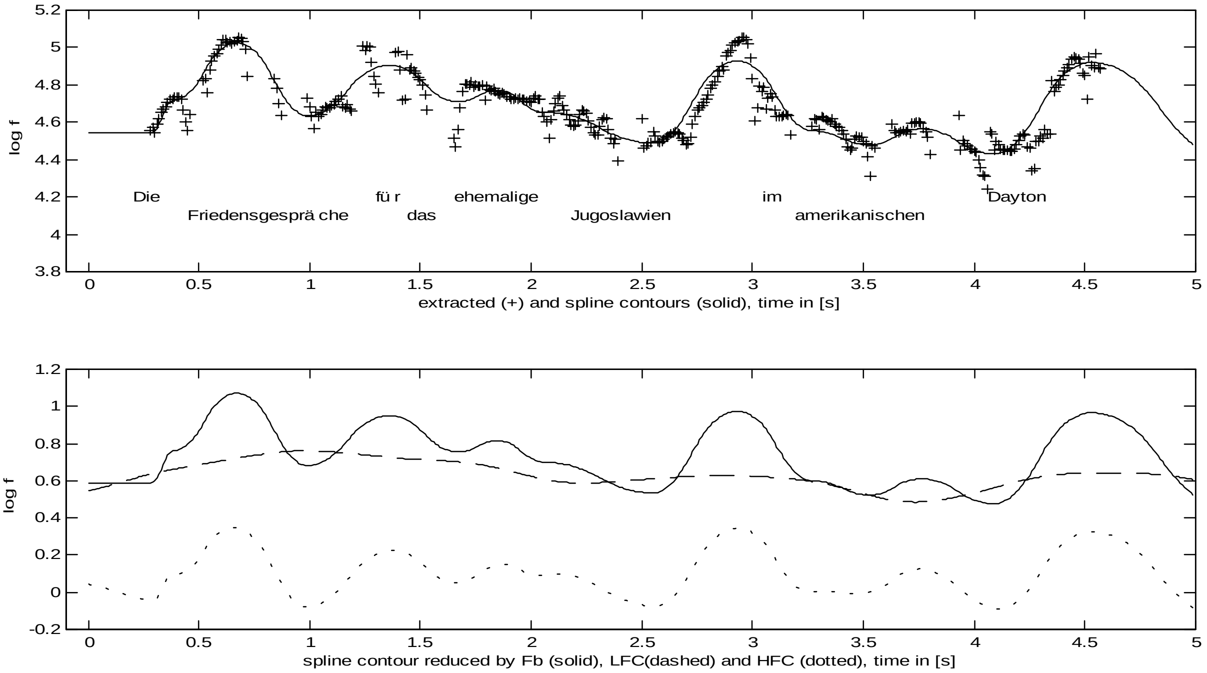

Figure 1. Top: Example of extracted (+signs) and spline contours (solid line), top). Bottom: LFC (dashed) and HFC (dotted), the solid line indicates the spline contour reduced by Fb.

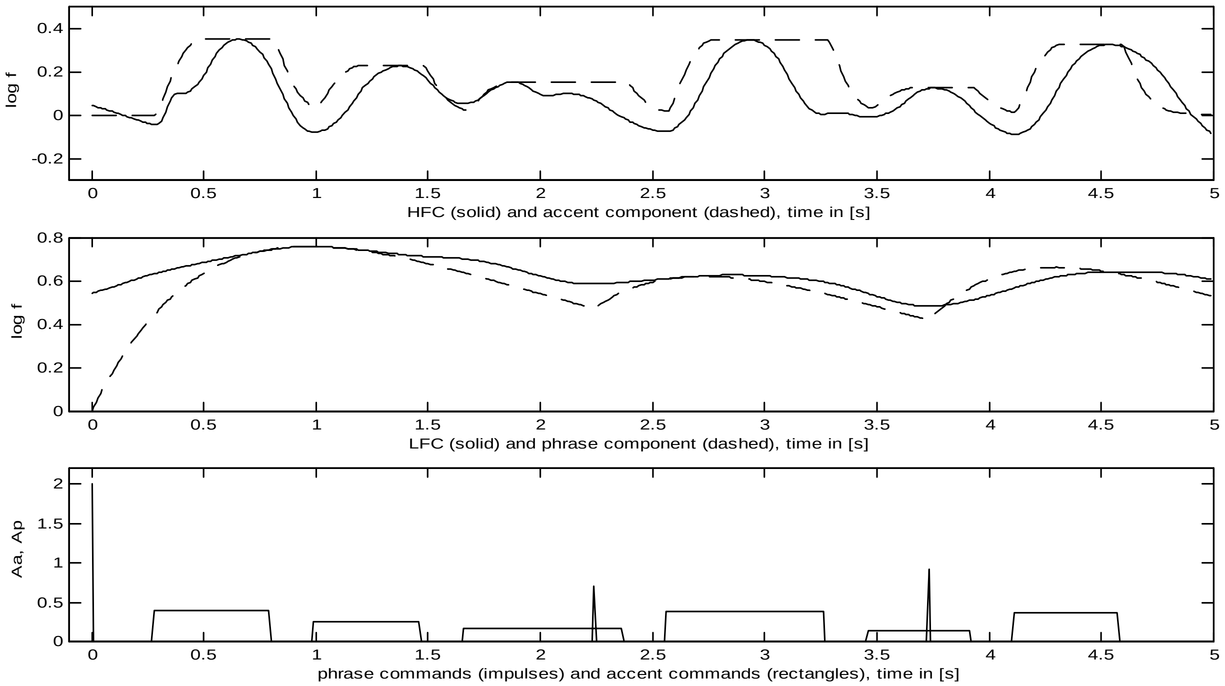

Figure 2. Initial Fujisaki model parameter configuration, bottom: phrase and accent commands, center: LFC and resulting phrase component, top: HFC and resulting accent component.

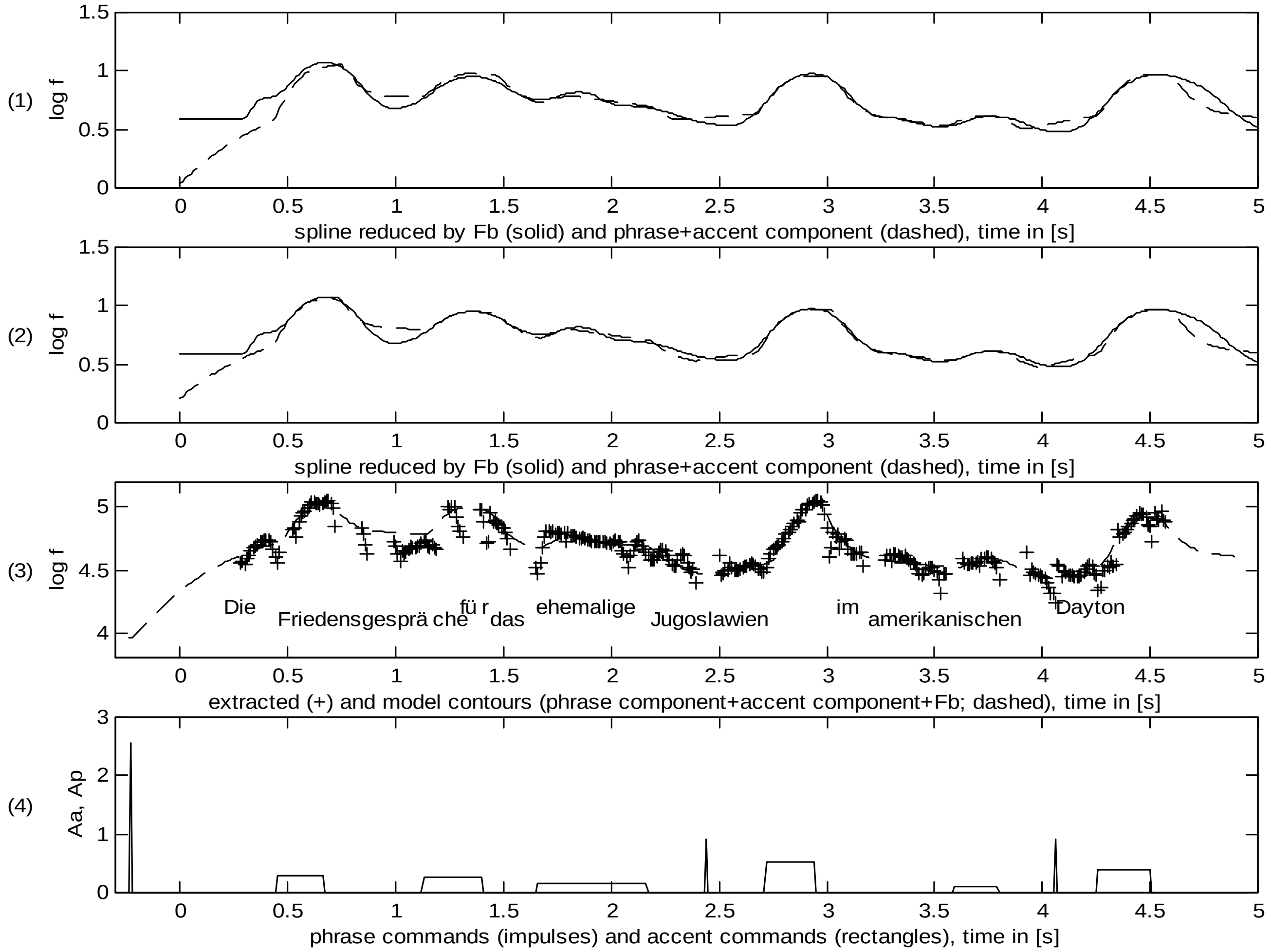

Figure 3. Intermediate and final results of optimization procedure. From top to bottom: (1) Spline contour and phrase plus accent component (1) after separate optimization (2) after fitting to spline contour; (3) extracted and final model-generated contours and (4) resulting Fujisaki model parameters.

2.2 High-Pass Filtering and Component Separation

The Fujisaki-model produces a particular F0 contour in the log F domain by superimposing three components: The phrase component which corresponds to the phrase-wise slow overall declination line, the accent component made up by the faster movements in the F0 contour connected with accents and boundary tones, and Fb, a speaker-individual constant.

In order to separate the accent component from the phrase component and Fb, the spline contour is passed through a high-pass filter with a stop frequency at 0.5 Hz. The output of the high-pass (henceforth called ‘high frequency contour’ or HFC) is subtracted from the spline contour yielding a ‘low frequency contour’ (LFC), containing the sum of phrase component and Fb. The latter is initially set to the overall minimum of the LFC. Hence, partial contours roughly corresponding to phrase and accent components are determined, as shown in Figure 1 (bottom).

2.3 Fujisaki-Model Command Initialization

The initialization procedure makes use of the characteristics of phrase and accent command responses making up phrase and accent components, respectively.

The phrase command response has its onset with the occurrence of an impulse-wise phrase command, rises to a maximum and then decays slowly according to the associated time constant α. Hence, in a sequence of phrase commands, the onset of a new command is characterized by a local minimum in the phrase component. Consequently, the LFC is searched for local minima, applying a minimum distance threshold of 1 s between consecutive phrase commands. For initializing the magnitude value Ap assigned to each phrase command the part of the LFC after the potential onset time T0 of a phrase command is searched for the next local maximum. Ap is then calculated in proportion to the frequency value found at this point. As responses of several phrase commands may add up in the phrase component, contributions of preceding commands must be taken into account when calculating Ap, which is reduced accordingly (see Figure 2, center). A full phrase command reset occurs at inter-phrase boundaries accompanied by a longer pause (> 500 ms). The time constant α is initially set to 1.0/s, a value found appropriate after a series of preliminary trials.

The accent command response is a smoothed square function rising from a value of 0 at T1 to a maximum which is sustained until the offset time T2 when it starts decaying. For initializing the appropriate number, onset times T1 and offset times T2 of accent commands, the HFC is searched for local minima, whose vicinity (+/- 100 ms) is scanned for even lower F0 values in order to avoid picking saddle points. Two subsequent local minima each are associated with a new accent command. Since the accent command response requires some time to decay to 0 after T2, T2 is set back to 200 ms before the local minimum. The accent command time constant β is set to a initial value of 20/s. For initializing the accent command amplitude Aa, the maximum in the HFC between T1 and T2 is determined and Aa set in proportion to the frequency value found at this point (see Figure 2, top). Accent commands are not continued across major pauses in the speech signal, as is the case for the rightmost accent command in Figure 2, bottom.

2.4 Analysis-by-Synthesis

The Analysis-by-Synthesis procedure is performed in three steps, in the course of which the initial parameter configuration (Figure 2, bottom) is subsequently optimized by applying a hill-climb search for reducing the overall mean-square-error in the log F domain. Each step terminates when the improvement between subsequent iterations drops below a set threshold.

At the first step, phrase and accent components are optimized separately, taking the LFC and HFC, respectively, as the targets. Figure 3, panel (1) shows the joint result of this step which already yields a quite close approximation of the spline contour. Next, phrase component, accent component and Fb are optimized jointly, taking the spline contour itself as the target (see Figure 3, panel (2) for the resulting approximation).

In the final step, the parameter configuration is further fine-tuned by making use of a weighted representation of the extracted original F0 contour. The weighting factor applied is the product of degree of voicing and frame energy for every F0 value, which favors ‘reliable’ portions of the contour. Figure 3, panels (3) and (4) show the resulting model contour and the underlying model commands.

3. PRELIMINARY RESULTS

For the sake of conciseness we only state a few observations. The mean approximation error yielded with this approach on the entire speech data base amounts to 3.1 % when taking into account all voiced frames, but is considerably lower for ‘reliable’ parts of the contour, such as stable vowel portions (mean error: 1.7 %).

About 95.6% of ToBI-labeled accents (H*L, L*H etc.) are associated with an accent command, ‘down-stepped’ accents exhibit a significantly lower mean Aa (0.219) than regular accents (0.293). About 54.8 % of break index 3- and 96.2 % of break index 4-labeled-boundaries are aligned with the onset of a phrase command, with mean Ap of 0.67 and 1.32, respectively.

4. SUMMARY

The current paper introduced a novel approach to extracting Fujisaki model parameters by applying a spline contour as intermediate approximation target. The method facilitates the analysis of larger corpora of speech data with a high degree of objectiveness. First results show that command configurations yielded can be readily related to linguistic information such as boundary strength and accent prominence. Future work will include a closer analysis of the relationship between model parameters and the underlying linguistic information, as well as efforts towards the development of an F0-predictor for TTS based on the current method.

5. REFERENCES

[1] Fujisaki, H. and Hirose, K. “Analysis of voice fundamental frequency contours for declarative sentences of japanese”. In *Journal of the Acoustical Society of Japan (E)*, 5(4): pp. 233-241, 1984.

[2] Mixdorff, H. Intonation Patterns of German - Model-based. Quantitative Analysis and Synthesis of F0-Contours. PdD thesis TU Dresden, 1998.

[3] Mixdorff, H. and Mehnert, D. “Exploring the Naturalness of Several German High-Quality-Text-to-Speech Systems”, *Proceedings of Eurospeech '99*, vol.4, pp.1859-1862, Budapest, Hungary, 1999.

[4] Taylor, P.A.. “The Rise/Fall/Connection Model of Intonation”. *Speech Communication*, vol. 15 , pp. 169-186, 1995.

[5] Portele, T.; Krämer, J.; Heuft, B. “Parametrisierung von Grundfrequenzkonturen”. *Fortschritte der Akustik - DAGA '95*, Saarbrücken, pp. 991-994, 1995.

[6] Möbius, B., Pätzold, M. and W. Hess. “Analysis and synthesis of German F0 contours by means of Fujisaki's model”, *Speech Communication*, vol.13, pp. 53-61. 1993.

[7] Mixdorff, H. and Fujisaki, H. “Automated Quantitative Analysis of F0 Contours of Utterances from a German ToBI-labeled Speech Database“. *Proceedings of Eurospeech '97*, vol. 1, pp 187 – 190, Rhodes, Greece. 1997.

[8] Rapp, S. Automatisierte Erstellung von Korpora für die Prosodieforschung, PhD thesis Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung. 1998.

[9] Mayer, J. Transcription of German Intonation: The Stuttgart System. Technischer Bericht, Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung. 1995.

[10] Hirst, Daniel & Espesser, Robert. “Automatic modelling of fundamental frequency using a quadratic spline-function”. *Travaux de l'Institut de Phonétique d'Aix* 15, pp. 71-85, 1993.