Towards the Automatic Extraction of Fujisaki model Parameters for Mandarin

Hansjörg Mixdorff, Hiroya Fujisaki**, Gao Peng Chen ***, and Yu Hu ****
*Faculty of Computer Science, Berlin University of Applied Sciences, Germany
mixdorff@tfh-berlin.de
**Professor Emeritus, University of Tokyo, Japan
fujisaki@alum.mit.edu
***University of Science and Technology of China, Hefei, Anhui, China
gpchen@mail.ustc.edu.cn; yuhu@iflytek.com

Abstract

The generation of naturally-sounding *F0* contours in TTS enhances the intelligibility and perceived naturalness of synthetic speech. In earlier works the first author developed a linguistically motivated model of German intonation based on the quantitative Fujisaki model of the production process of *F0*, and an automatic procedure for extracting the parameters from the *F0* contour which, however, was specific to German. As has been shown by Fujisaki and his co-workers, parametrization of *F0* contours of Mandarin requires negative tone commands, as well as a more precise control of *F0* associated with the syllabic tones. This paper presents an approach to the automatic parameter estimation for Mandarin, as well as first results concerning the accuracy of estimation. The paper also introduces a recently developed tool for editing Fujisaki parameters featuring resynthesis which will soon be publicly available.

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 first 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. A main attraction of the Fujisaki model is its underlying 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'.

Mandarin is a well known tone language and has four different lexical tones plus a so-called 'light' tone in unstressed syllables. Fujisaki and his co-workers have shown that modeling *F0* contours of Mandarin requires accent commands (referred to as 'tone commands' in the case of Mandarin) of positive and negative polarity (Table 1). The 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. Earlier the first author developed a multi-stage approach consisting of a quadratic spline smoothing, contour filtering, accent command initialization and a three-pass Analysis-by-Synthesis procedure[4]. Since this approach, however, was specific to German in that it only assumed accent commands of positive polarity and applied certain thresholds concerning the distance between consecutive accent commands it could not be applied to Mandarin directly, but had to be modified.

Table 1: Mandarin tones with tone command assignment.

tone	*F0* contour	tone commands assigned
1	high	positive
2	rising	negative/positive
3	falling-rising	negative
4	falling	positive/negative

The current paper discusses the resulting algorithm. The method was applied to part of a corpus of read speech compiled by USTC of a total of 14450 sentences. The corpus designed for a unit selection TTS system was uttered by a female professional radio announcer and contains boundary labels on the phone and syllable levels, as well as syllabic tones and break indices. The subcorpus of 100 sentences used for the present study contains a total number of 2245 syllables and has a total duration of about 21 minutes. 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 at the final stage of the modeling procedure.

2. Modeling Procedure

2.1 Interpolation and Smoothing

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 are smoothed out, as the Fujisaki model explicitly deals with macroprosody only. In the current approach, a cubic interpolation and smoothing algorithm was incorporated which was originally developed for a Fujisaki parameter estimation of Japanese [5]. First experiments showed that the smoothing interval had to be reduced from 200 to 100

ms due to the faster changes in the *F0* contours of Mandarin as compared to Japanese. Figure 1 (top) shows the initial part of an utterance from the database displaying the extracted (+ signs) and the interpolated smoothed contours (solid line), as well as syllable labels and boundaries. As can be seen from the figure, silent intervals are interpolated by straight lines, in order to avoid the extreme excursions sometimes observed with cubic interpolation.

2.2 High-Pass Filtering and Component Separation

The Fujisaki model as applied to Mandarin 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 tone component made up by the faster movements in the *F0* contour

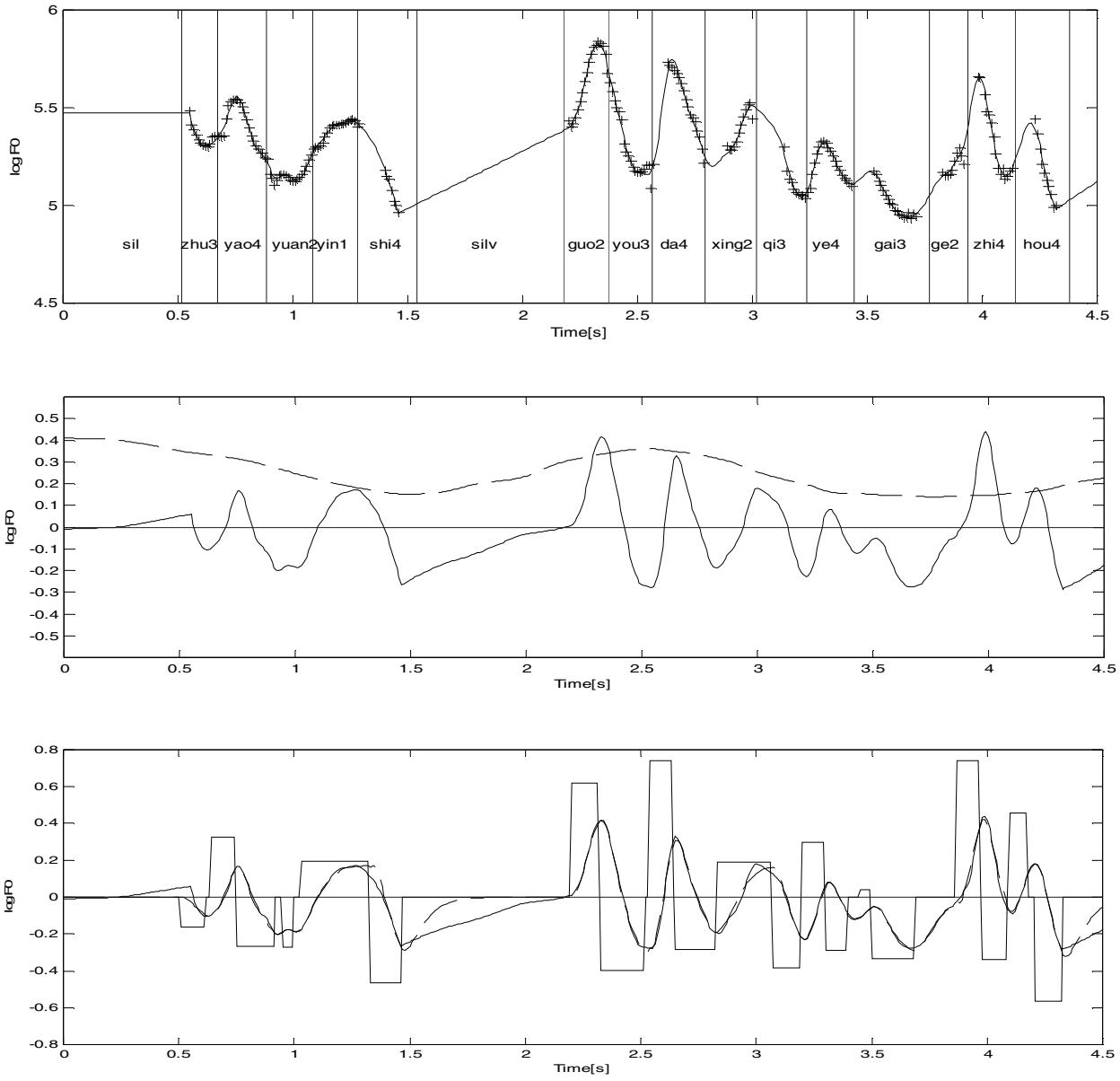

Figure 1. Top: Example of extracted (+signs) and interpolated/smoothed (solid line)*F0* contours from the corpus. Center: HFC (solid) and LFC (dashed) reduced by *Fb*. The HFC oscillates around zero, with excursions suggesting underlying positive or negative tone commands. Local minima in the LFC indicate the approximate positions of subsequent phrase commands, except for the first phrase command which is located before the onset of the utterance. Bottom: Tone command sequence and resulting model contour (dashed) as compared with the HFC (solid line) after the first optimization stage.

connected with syllabic tones, and *Fb*, a speaker-individual constant. In order to separate the tone component from the phrase component and *Fb*, the smooth 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 smooth contour yielding a 'low frequency contour' (LFC), containing the sum of phrase component and *Fb*. The latter, unless it is not explicitly set to a constant value at the beginning of the analysis is initially set to the overall minimum of the LFC. Hence, partial contours roughly corresponding to phrase and tone components are determined, as shown in Figure 1 (center). In the case of the current speaker, a value of *Fb* of 160 Hz was found to be optimal.

2.3 Fujisaki-Model Command Initialization

It was shown in [4] that 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 (compare Figure 1, center), 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. The time constant α is set to 2.0/s, a value found appropriate after a series of preliminary trials. In order to yield an optimal initialization of phrase commands, the *F0* contour is examined for pauses longer than 400 ms. Since the minima of the LFC do not provide the exact locations of upcoming phrase commands, phrase commands found in the vicinity of a pause are readjusted and aligned with with the pause.

For initializing the appropriate number, polarity and onset times *T1* and offset times *T2* of tone commands, the cubically smoothed contour pertaining to the part of the wave file between first and last voiced frame, is subdivided into segments with positive or negative gradient, respectively. These contour segments are searched for points where the derivative exhibits a maximum, that is, the inflection points of the cubically smoothed curve. Inflection points on contour segments with rising slope are associated with the offset of a negative tone command and the subsequent onset of a positive tone command, and inflection points on contour segments with falling slope are associated with the offset of a positive tone command and the subsequent onset of a negative tone command. Hence, an alternating sequence of positive and negative commands is yielded initially.

The HFC is basically DC-free and therefore oscillates around 0. For initializing the tone command amplitude *Aa* the positive or negative maximum in the HFC between the initial settings of *T1* and *T2* is determined and *Aa* set in proportion to the frequency value found at this point. Tone commands are not continued across major pauses in the speech signal. The tone command time constant β is set to a value of 20/s.

2.4 Analysis-by-Synthesis

The Analysis-by-Synthesis procedure is performed in three steps, in the course of which the initial parameter configuration 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 tone components are optimized separately, taking the LFC and HFC, respectively, as the targets. Figure 1 (bottom) displays the result after this step with respect to the tone commands. Next, phrase component, tone component and *Fb* are optimized jointly, taking the smooth contour proper as the target.

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 2 shows the resulting model contour and the underlying model commands. As an option, during the analysis, tone commands with extreme shape, that is, very long commands with extremely low amplitude or very short commands with extremely high amplitude can be deleted.

3. Preliminary Results

The Fujisaki parameters calculated for the corpus were examined for errors, that is, for instance, wrong polarity or insertions/deletions. In a sequence of tones, however, consecutive tone commands of the same polarity are allowed to merge. That is, a sequence of two tone 1 syllables, for instance, can exhibit a single long positive command. Analogously, the negative tone command in a tone 4 syllable can merge with the negative command of a following tone 3 syllable. These mergers were not regarded as errors because they reflect tone coarticulation.

In the entire corpus of 2245 syllables, 152 deletions of actually required tone commands were observed, as well as 76 insertions. 10 of these insertions were related to *F0* extraction errors. There were also 55 cases in which two consecutive short tone commands of the same polarity which actually should have been merged were assigned to a single syllable. Of the 594 phrase commands which were extracted 45 were not closely associated with major syntactic breaks. Furthermore 19 required phrase commands were missing. As a consequence, the amplitudes of superimposed tone commands in these parts were incorrect. Generally speaking, pauses are good candidate locations for new phrase commands, but commands in inter-pause stretches are somewhat more difficult to assign. In these cases information on break indices might be useful.

4. Integration into Editing Tool

In order to provide a graphical and acoustic feedback and a means of controling the Fujisaki parameter extraction at any stage, an editing and resynthesis tool, the *FujiParaEditor,* was developed by Patavee Charnvinit (CRSLP, Chulalongkorn University, Bangkok, Thailand), Gao Peng Chen and Yu Hu. Along with full click-and-drag control over Fujisaki parameters, the automatic extraction routine can be called from within the editor and started from and run up to any of the analysis steps. Hence, intermediate parameter configurations can be displayed immediately and if necessary adjusted. The *FujiParaEditor* also provides a real-time Fujisaki model based

resynthesis feature by remote procedure call to the PSOLA resynthesis of the PRAAT *ManipulationObject* (© P. Boersma). The software will soon be made freely available to the research community.

5. Summary

The current paper introduced a novel approach to extracting Fujisaki model parameters of Mandarin employing a cubically smoothed contour as the intermediate approximation target. The method takes into account the need for tone commands of both positive and negative polarity. By integrating the algorithm into an editing tool intermediate parameter settings can be modified and the analysis rerun. Future work will include a closer analysis of the relationship between model parameters and linguistic information, especially the identification of tones based on the underlying tone command configurations.

6. 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. D.Eng. thesis TU Dresden, 1998.

[3] Fujisaki, H., Hallé, P. and Lei, H., "Application of F_0 contour command-response model to Chinese tones," *Reports of Autumn Meeting, Acoustical Society of Japan,* **1***:* 197-198, 1987.

[4] Mixdorff, H., "A novel approach to the fully automatic extraction of Fujisaki model parameters," in *Proceedings ICASSP 2000*, vol. 1, 1281-1284, Istanbul, Turkey, 2000.

[5] Fujisaki, H. and S. Narusawa, "Automatic extraction of model parameters from fundamental frequency contours of speech," *Proc. 2001 2nd Plenary Meeting on Prosody and Speech Processing*, pp. 133-138, Tokyo, 2001.

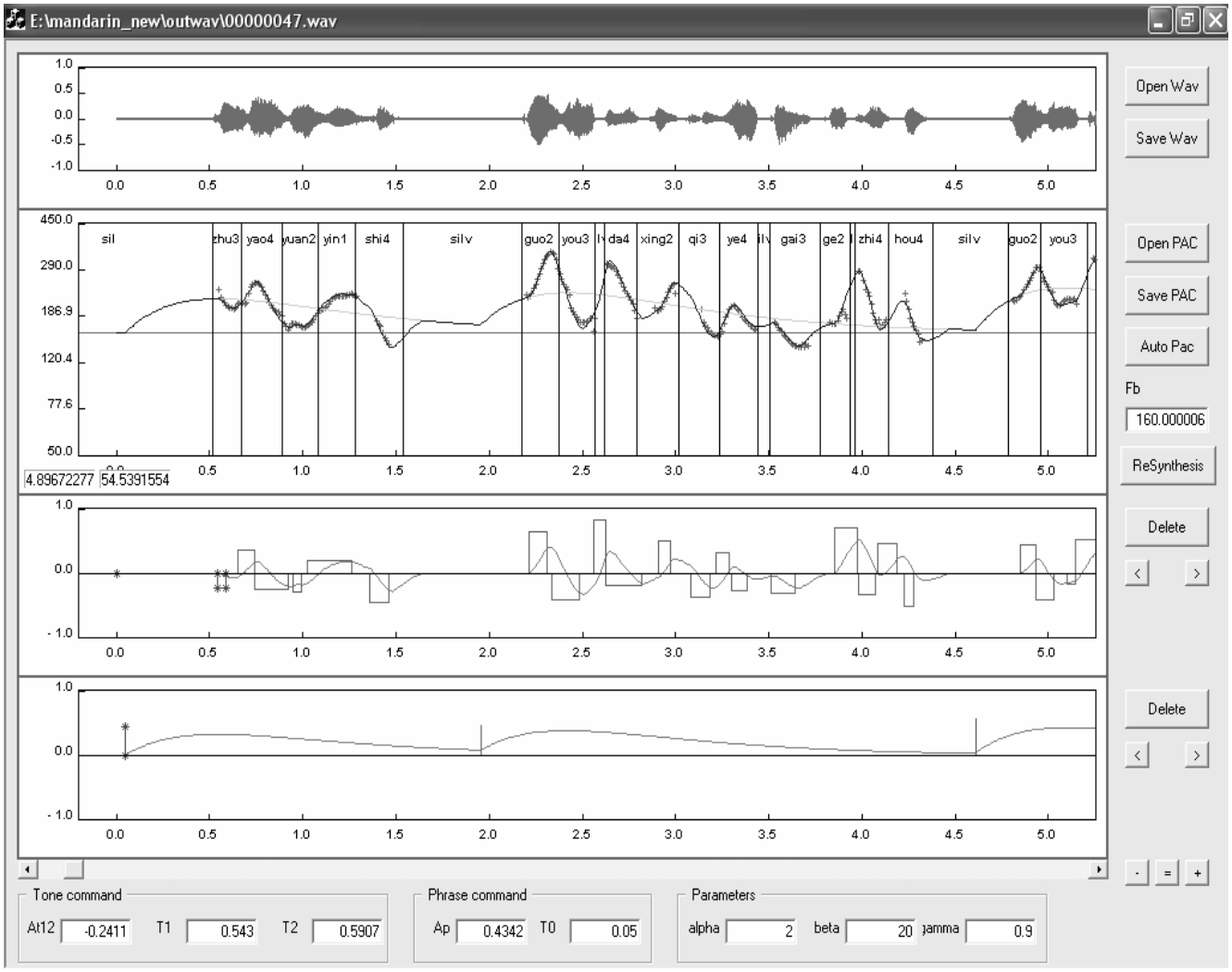

Figure 2. Resulting parameter configuration as displayed in the *FujiParaEditor*. From top to bottom: Speech waveform, *F0* contour, tone commands and resulting tone component, and phrase commands and resulting phrase component. The *FujiParaEditor* features real-time resynthesis based on Fujisaki model contour and integrated automatic extraction ('Auto Pac').