INTERSPEECH 2009 BRIGHTON

Comparison of Fujisaki-Model Extractors and F0 Stylizers

Hartmut R. Pfitzinger[1] *Hansjörg Mixdorff*[2] *Jan Schwarz*[3]

[1] Institute of Phonetics and Digital Speech Processing (IPDS)
Christian-Albrechts-University of Kiel, Germany
[2] Department of Computer Sciences and Media
BHT Berlin University of Applied Sciences, Germany
[3] Institute for Circuit and System Theory (LNS), Faculty of Engineering
Christian-Albrechts-University of Kiel, Germany

hpt@ipds.uni-kiel.de mixdorff@beuth-hochschule.de js@tf.uni-kiel.de

Abstract

This study compares four automatic methods for estimating Fujisaki-model parameters. Since interpolation and smoothing are necessary prerequisites for all approaches their fitting accuracies are also compared with that of a novel stylisation method. A hand-corrected set of results from one of the methods which was created on linguistic grounds served as a second benchmark. Although the four methods yield comparable results with respect to their total errors, they show different error distributions. The manually corrected version provided a poorer approximation of the F0 contours than the automatic one.

Index Terms: prosody, intonation, Fujisaki model, automatic parameter estimation, F0 stylisation, evaluation

10.21437/Interspeech.2009-313

1. Introduction

Over the last decades various techniques were developed for post-processing raw F0 contours in order to reduce redundancy, to increase their degree of abstraction, and to derive from them perceptually or functionally relevant content. These methods are of great importance, since error-prone automatic detection methods extract F0 contours which are affected by a complex superposition of microprosodic disturbances such as jitter, laryngealization, vowel-intrinsic pitch, and aerodynamic fluctuations. In addition, these contours are interrupted by voiceless stretches of speech. Most intonation studies are concerned with macroprosodic components of F0, e.g. pitch-accent, boundary tones, declination etc. Therefore, a reliable pre-processing and decomposition of raw F0 contours is crucial for the development of higher level models. Fig. 1 shows the process leading from raw F0 via stages of stylisation and component separation towards more abstract symbolic representations of intonation.

One well-known method for parameterizing F0 contours is the Fujisaki model which will be the focus of the current study. Our goal is to compare four available automatic extractors for Fujisaki model parameters by means of a reference database. F0 stylisation methods serve as a baseline to assess the error between original and reconstructed F0. Since F0 contour interpolation and smoothing are necessary prerequisites for all existing Fujisaki model extractors we will first review common F0 stylisation methods, then introduce the Fujisaki model and discuss the approaches compared.

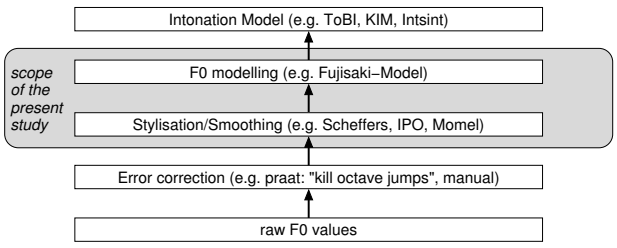

Figure 1: *F0 processing hierarchy.*

2. Algorithms

2.1. F0 stylisation methods

F0 stylisation methods have two main goals: (1) microprosodic disturbances should be removed from the F0 contour without affecting perception, and (2) F0 contours should be interpolated during voiceless stretches of speech, as Scheffers 1988 [14, p. 982] already remarked: "listeners won't perceive sentence melodies to be interrupted by unvoiced speech sounds". Nooteboom 1997 [10, p. 644] specified that "interruptions [...] are only perceived [...] when they are longer than, roughly, 200 ms. [...] When the pitch after a silent interval is considerably higher or lower than before, the listener perceives a rise or fall in pitch, as if human perception unconsciously bridges the silent gap by filling in the missing part of the pitch contour".

One of the pioneers of F0 contour stylisation, t'Hart 1976 [16, p. 18], justifies the underlying strategy by the surprisingly low sensitivity of humans to differences in pitch movements. Scheffers 1988 [14, p. 981] points out that simple low-pass filtering is not sufficient to remove from F0 contours irregularities that have no relation to the perceived intonation, since it "will affect the slope and onset and offset moments of the important movements". However, (1) electromyographic investigations of the *pars recta* and *pars obliqua* of the cricothyroid muscle which are mainly responsible for F0 movements [1, 2], (2) production studies concerning the maximum speed of F0 changes [18], and (3) preliminary results from recent production studies give rise to the assumption that the modulation frequency of functionally motivated F0 changes hardly exceeds 3.5 Hz.

A popular stylisation method is *Momel* (modelling melody) introduced by Hirst & Espesser 1993 [3]: A quadratic spline aligned to so-called target points along the F0 contour yields a smoothed version that is perceptually indistinguishable from the original and supposedly void of microprosodic fluctuations.

2.2. Fujisaki model, properties, and extraction methods

The well-known Fujisaki model [2] reproduces a given F0 contour by superimposing three components in the log F0 domain: A speaker-individual base frequency F_b, a phrase component and an accent component. The phrase component results from impulse responses to impulse-wise phrase commands associated with prosodic breaks. Phrase commands are described by their onset time T_0, magnitude A_p and time constant α. The accent component results from step-wise accent commands associated with accented syllables. Accent commands are described by on- and offset times T_1 and T_2, amplitude A_a and time constant β. Typical values for α and β are 3 and 20/s, respectively. Möbius 1993 [8, p. 115] chose a ratio of 1:5 for his studies rather than 1:7 or even 1:10 which were commonly observed when analysing actual F0 contours. Earlier extractors such as Pätzold 1991 [11] and Narusawa et al. 2000 [9] are unfortunately not accessible and had to be left unconsidered in the present study albeit mentioned for reasons of completeness.

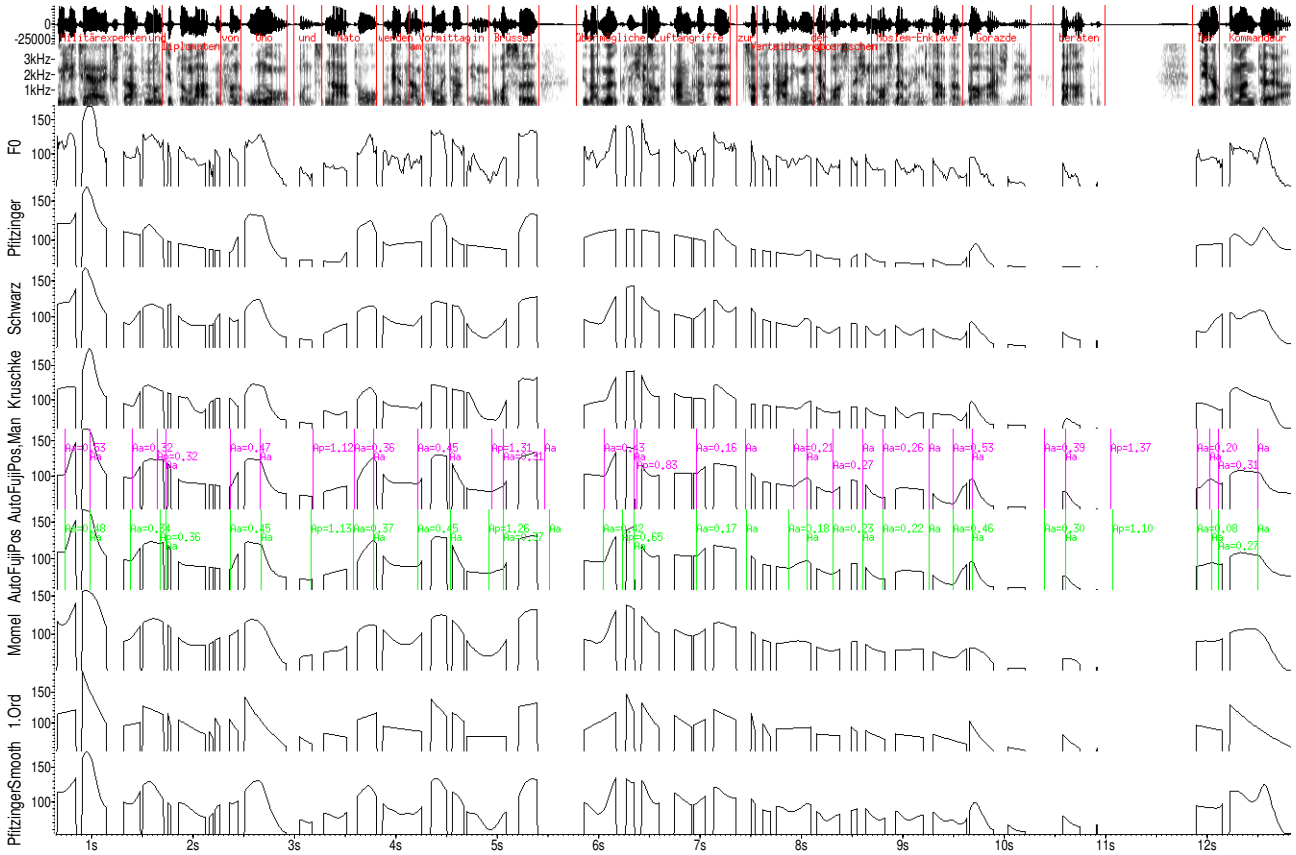

Figure 2: *Signal dlf951121.1200.n2.wav with reference F0 and the F0 contours resulting from Fujisaki extractors and from F0 stylizers.*

2.2.1. *Mixdorff*

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 [15]. 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 components and F_b. The latter is initially set to the overall minimum of the LFC. The HFC is searched for consecutive minima delimiting potential accent commands whose A_a is initialized to reach the maximum of F0 between the two minima. Since the onset of a new phrase command is characterized by a local minimum in the phrase component the LFC is searched for local minima, applying a minimum distance threshold of 1 s between consecutive phrase commands. For initializing the magnitude value A_p assigned to each phrase command the part of the LFC after the potential onset time T_0 is searched for the next local maximum. A_p is then calculated in proportion to the F0 at this point considering contributions of preceding commands. The Analysis-by-Synthesis procedure is performed in three steps, optimizing the initial parameter set iteratively by applying a hill-climb search for reducing the overall mean-square-error in the log F domain. At the first step, phrase and accent components are optimized separately, using LFC and HFC, respectively, as the targets. Next, phrase component, accent component and F_b are optimized jointly, 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 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.

2.2.2. *Kruschke*

As in [4], after piecewise polynomial interpolation and smoothing the lowest F0>0 is selected as a first approximation of F_b, and subtracted from the logarithmic F0 contour. Then a Wavelet Transform using a Mexican hat wavelet is applied to the residual signal F0res1(t). From the left to the right the first marked maximum in the resulting scalogram is searched and picked as the maximum of a detected accent. The preceding marked minimum is selected as a starting value for T_1. T_2, the point where the smoothed accent command reaches 0 is set to the next F0 minimum. The initial values of the parameters A_a, β and T_2 are obtained in a pattern comparison, i.e. within specific ranges A_a, β and T_2 are successively incremented to match the local F0 contour around the accent. The parameter set with the smallest RMSE is taken as a first approximation of the parameters A_a, β and T_2. Accent detection continues by searching the next marked maximum after T_2. Then the resulting parameters are optimized in an A-b-S procedure, which is controlled by an evolutionary strategy. An F0 contour is generated from the accent commands and subtracted from the contour F0res1(t). The resulting residual contour F0res2(t) is smoothed and used for detecting the phrase commands, again by Wavelet Transform using the Mexican hat wavelet. Each marked maximum in the scalogram is assigned to a phrase. The point in time 200 ms before a maximum at the beginning of the F0 contour is chosen as a first approximation of T_0 and the lowest F0 value between two extremes is selected as a starting value of T_0. A_p, α and T_0 are estimated by a procedure similar to that for accents. The algorithm continues until the parameters of the last phrase have been estimated. Finally the parameters of all phrase and accent commands are optimized jointly.

2.2.3. Schwarz

As in [6], an equiripple FIR high-pass filter with a 0.5 Hz cutoff frequency separates quadratic-spline interpolated F0 contours into high-pass (HPC) and low-pass components (LPC). LPCs contain the phrase components and the speaker-dependent baseline frequency F_b set to the global minimum of the LPC. Accent and phrase components are extracted from the HPC and the LPC, respectively, by searching for local extremes. Since local maxima of the HPC roughly correspond to the accent components and their amplitudes, consecutive local minima are used to define regions related to the onset T_1 and the offset T_2 time. T_1 is set to the local minimum and T_2 to a position 200 ms before the next minimum. The local maxima of phrases correspond to the magnitudes A_p and local minima delimit the regions of phrase components. Phrase components will have at least a distance of 750 ms between them [7]. Finally, the extracted parameters are adjusted recursively in the least-squares sense. In contrast to [6] the parameters are optimized segmentally, i.e., the number of phrase components is given by the number of extracted time segments and will not be changed, and accent components are allowed to be merged or to be cancelled. Starting from the HPC, each time segment given by T_1, T_2 and A_a represents an accent component that is optimised iteratively. Subtracting the fitted HPC from the F0 contour results in a modified LPC which is also optimized iteratively. The procedure is carried out recursively as long as the MSE error of the modified LPC and the previous modified LPC is larger than 5%.

2.2.4. Pfitzinger

This method was developed in 2004 but yet unpublished. All higher F0 modulation components above 3.5 Hz are removed from raw F0 values by applying sample-selective Fourier Transform [12] and successive frequency-domain low-pass filtering with a -18dB/oct slope, to avoid the deformation of slopes, onsets, and offsets of F0 contours. The stylised F0 contour is resynthesized from frequency-, amplitude- and phase-locked sinusoids without interruptions at voiceless stretches of speech. This new smoothing is included in the evaluation as *PfitzingerSmooth*. The smooth contour is passed through a low-pass filter (0.35 Hz cutoff, -18dB/oct slope) whose output contour maxima are regarded as the phrase command amplitudes and positions. Subtracting this contour from the 3.5-Hz-filtered one leads to the signal which serves as the basis for accent command extraction. Schmitt triggering with a threshold of 0.2 and 10% hysteresis followed by delaying the achieved positions by -85 ms yields the accent command amplitudes and times.

3. Evaluation

The evaluation is based on the *IMS Radio News Corpus* [13]. It consists of German news texts read by professional speakers. The extractors by Mixdorff and Kruschke were both developed on this corpus. Thus, reference data for the Fujisaki model exists that was extracted automatically [6] and manually corrected following linguistic criteria [7] and using the interactive Fuji-ParaEditor [5]. Although raw F0 data are provided with the corpus extracted in 10 ms steps via *get_f0* of *ESPS waves* [17], 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 phrase and accent command amplitudes and positions produced by the four Fujisaki-model extractors as well as F_b, α and β were used to resynthesise the F0 values by means of the Fujisaki model which is defined in the log F0 domain. Thus, our evaluation also uses a semitone scale.

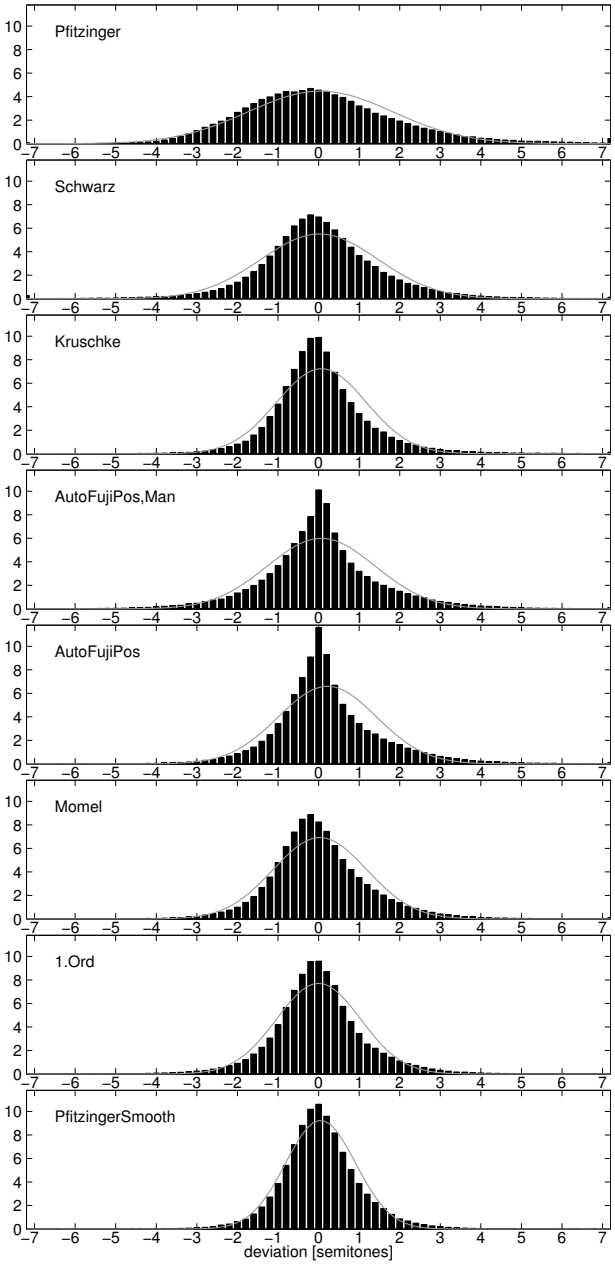

Figure 3: *Histograms of the frequency (in percent) and amount of deviation from reference F0 values in semitones.*

4. Results

Fig. 2 shows an example of analysis from the database. From the top to the bottom: The speech waveform, the underlying text, the spectrogram, the reference F0 contour, Fujisaki model-based contours for Pfitzinger, Schwarz, Kruschke, Mixdorff with manual post-processing and without, furthermore contour stilisations based on *Momel*, first order stylisation, and the novel smoothing approach by Pfitzinger. Visual inspection confirms that all methods capture the essential F0 movements adequately and yield very similar smoothed versions of the original. Fig. 3 displays histograms of fitting errors measured in semitones for all methods examined. Of all Fujisaki model extractors the one by Kruschke yields the smallest standard deviation and hence the best overall fit. It is followed by the automatic and the manually corrected versions of Mixdorff, and the methods by Schwarz and Pfitzinger whose error distributions resemble

more the shape of the corresponding Gaussian (drawn with a grey line) whereas Kruschke's and Mixdorff's algorithms produce error distributions that are more Laplace-shaped and yield a proportionally larger number of very small errors. *Momel* is slightly worse than the first order stylisation. The closest approximation, however, is reached by the novel smoothing approach. Fig. 4 presents a more condensed way of looking at the error distributions by displaying error thresholds for 50, 90, 99 and 99.9% of the data, respectively. For example, it shows that 90% of the deviations of three Fujisaki extractors are below 2.5 semitones. 0.1% means that deviations are greater than 7.5 to 10 semitones only for 167 F0 values.

5. Discussion

In order to interpret our results we have to bear in mind that the number of accent and phrase commands as well as variability of the (theoretical) model constants α, β and F_b have a direct influence on the accuracy of approximation. The more commands are employed, the better the fitting of an observed F0 contour becomes. As a consequence, however, the resulting parameters will become more and more difficult to interpret, since they will ultimately model micro-prosodic fluctuations and not accented syllables or phrasal declination. Hence, moving from the automatic to the manually post-processed version of Mixdorff, the fitting accuracy decreases, because only those commands remain that can be motivated by accented syllables and prosodic phrase onsets. As an additional restriction, the manually post-processed version employs constant F_b, α and β for one and the same speaker, whereas F_b is adjusted in Schwarz depending on the particular sentence. In Kruschke, besides F_b, also α and β are varied and therefore lead to a smaller error. Since, however, F_b, A_p and α, as well as A_a and β are related through the model formulation, A_p and A_a become more difficult to compare when F_b, α and β are treated as variables. The following table summarizes the main properties of the four extractors:

	command rates		model parameters			algorithmic
Method	accents/s	phrases/s	F_b	α, β	RMSE	compl.
Pfitzinger	0.66	0.32	const.	const.	1.99	low
Schwarz	1.10	0.56	var.	const.	1.61	very hi
Kruschke	1.43	0.46	var.	var.	1.23	very hi
AutoFujiPos,Man	1.06	0.42	const.	const.	1.48	high

With respect to the evaluation of the approaches compared we are aware that objective differences such as RMSE cannot replace psycho-acoustic experiments regarding either the perceptual or — as a somewhat relaxed criterion — functional-semantic equivalence of original, stylised, and modeled F0 contours, an argument already raised by Möbius 1993 [8, p. 116].

The best way of ensuring that the Fujisaki model parameters reflect the underlying linguistic units and structures of an utterance would be by introducing such knowledge already at the stage of parameter extraction. Applying these restrictions, as can be seen when comparing *AutoFujiPos,Man* and *AutoFujiPos* in Fig. 4, might lead to poorer approximations. However, from the stand-point of intonation research we are not so much interested in just noticeable differences between F0 contours, but rather in the functional differences. Therefore, the ultimate goal should not be the closest approximation to automatically extracted F0 values, which by nature is an unreliable reference, but rather the derivation of an interpretable set of parameters that can be related to the meaning conveyed by an utterance.

Future work will concern perceptual evaluations of the contours generated for the current study, as well as efforts towards the integration of linguistic knowledge into the model parameter estimation procedure properly.

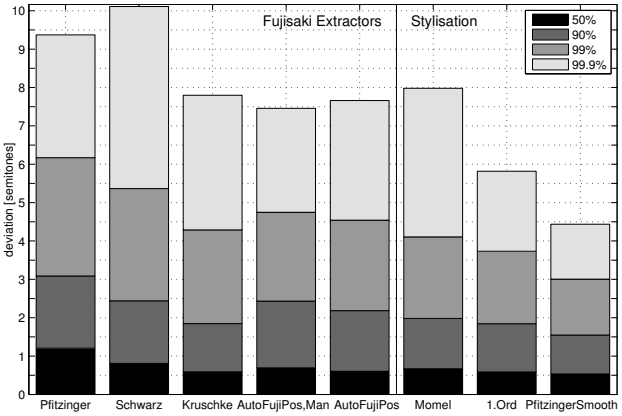

Figure 4: *Deviation in semitones between modelled F0 and reference F0. 50%, 90%, 99%, and 99.9% of all the data are below the F0 thresholds shown in the bars, respectively.*

6. References

[1] Fujisaki, H. 1988. A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In Fujimura, O., ed., *Vocal Fold Physiology. Voice Production, Mechanisms and Function*, vol. 2, pp. 347–355. Raven Press, New York.

[2] Fujisaki, H. 2004. Information, prosody, and modeling with emphasis on tonal features of speech. In *Proc. of the 2nd Int. Conf. on Speech Prosody*, pp. 1–10, Nara; Japan.

[3] Hirst, D.; Espesser, R. 1993. Automatic modelling of fundamental frequency using a quadratic spline function. Travaux de l'Institut de Phonétique d'Aix 15, pp. 71–85, Univ. de Provence.

[4] Kruschke, H.; Lenz, M. 2003. Estimation of the parameters of the quantitative intonation model with continuous wavelet analysis. In *Proc. of EUROSPEECH '03*, vol. 4, pp. 2881–2884, Geneva.

[5] Mixdorff, H. 1/10/2009. FujiParaEditor: http://www.tfh-berlin.de/∼mixdorff/thesis/fujisaki.html. TFH Berlin University of Applied Sciences.

[6] Mixdorff, H. 2000. A novel approach to the fully automatic extraction of Fujisaki model parameters. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP2000)*, vol. 3, pp. 1281–1284, Istanbul.

[7] Mixdorff, H. 2002. *An integrated approach to modeling German prosody*. w.e.b. Universitätsverlag, Dresden.

[8] Möbius, B. 1993. *Ein quantitatives Modell der deutschen Intonation: Analyse und Synthese von Grundfrequenzverläufen*. Niemeyer, Tübingen.

[9] Narusawa, S.; Fujisaki, H.; Ohno, S. 2000. A method for automatic extraction of parameters of the fundamental frequency contour. In *Proc. of ICSLP 2000*, vol. 1, pp. 649–652, Beijing.

[10] Nooteboom, S. G. 1997. The prosody of speech: Melody and rhythm. In Hardcastle, W. J.; Laver, J., eds., *The Handbook of Phonetic Sciences*, Nr. 5 in Blackwell Handbooks in Linguistics, chap. 21, pp. 640–673. Blackwell, Oxford.

[11] Pätzold, M. 1991. Nachbildung von Intonationskonturen mit dem Modell von Fujisaki. Implementierung des Algorithmus und erste Experimente mit ein- und zweiphrasigen Aussagesätzen. Master's thesis, Univ. Bonn.

[12] Pfitzinger, H. R. 2000. Removing hum from spoken language resources. In *Proc. of ICSLP 2000*, vol. 3, pp. 618–621, Beijing.

[13] Rapp, S. 1998. Automatisierte Erstellung von Korpora für die Prosodieforschung. Arbeitspapiere (phonetikAIMS) 4(1), pp. 1–167, Inst. für Maschinelle Sprachverarbeitung, Lehrstuhl für experimentelle Phonetik der Univ. Stuttgart.

[14] Scheffers, M. T. M. 1988. Automatic stylization of F0-contours. In Ainsworth, W. A.; Holmes, J. N., eds., *Proc. of SPEECH '88. 7th FASE Symposium*, vol. 3, pp. 981–987, Edinburgh.

[15] Strom, V. 1995. Detection of accents, phrase boundaries and sentence modality in German with prosodic features. In *Proc. of EUROSPEECH '95*, vol. 3, pp. 2039–2041, Madrid.

[16] 't Hart, J. 1976. Psychoacoustic backgrounds of pitch contour stylisation. IPO Annual Progress Report 11, pp. 11–19, Inst. for Perception Research, Eindhoven.

[17] Talkin, D. 1995. A robust algorithm for pitch tracking (RAPT). In Kleijn, W. B.; Paliwal, K. K., eds., *Speech coding and synthesis*, chap. 14, pp. 495–518. Elsevier, New York.

[18] Xu, Y.; Sun, X. 2000. How fast can we really change pitch? maximum speed of pitch change revisited. In *Proc. of ICSLP 2000*, vol. 3, pp. 666–669, Beijing.