EVALUATION OF F0 STYLISATION METHODS AND FUJISAKI-MODEL EXTRACTORS

Hartmut R. Pfitzinger[1] *Hansjörg Mixdorff*[2]

[1]*Inst. of Phonetics and Digital Speech Processing (IPDS), Christian-Albrechts-University Kiel, Germany*
[2]*Department of Computer Sciences and Media, BHT Berlin University of Applied Sciences, Germany*
hpt@ipds.uni-kiel.de mixdorff@beuth-hochschule.de

Abstract: Four automatic methods for estimating parameters of the Fujisaki model are evaluated and compared with three F0 stylisation methods. Although the four methods yield comparable results with respect to their total errors, they show different error distributions. Particularly, the command amplitude distributions of two methods reveal weaknesses in accent or phrase command extraction due to arbitrarily set amplitude thresholds. Also, the means of the command rates are significantly different and their standard deviations are inhomogeneous. Finally, an alignment of the commands of the extractors shows correspondences between 46% and 91% of the phrase commands and 49% and 97% of the accent commands. In summary, two of the four Fujisaki-model extractors are currently unsuitable for meaningful phonetic as well as functional analysis and should be substantially improved.

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. and thus are dependent on a reliable decomposition of the high- and low-frequency components of prosodies [16]. First, raw F0 values need to be automatically and sometimes even manually corrected. Prior to F0 parameterisation or modelling, a stylisation step is necessary usually consisting of interpolation, smoothing, and data reduction.

One well-known method for parameterising 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. Therefore, we will first review common F0 stylisation methods, then introduce the Fujisaki model and discuss the approaches compared.

2 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 [18, p. 982] already remarked: “listeners won't perceive sentence melodies to be interrupted by unvoiced speech sounds". Nooteboom 1997 [13, 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 [3, p. 18], justifies the underlying strategy by the surprisingly low sensitivity of humans to differences in pitch movements. Scheffers 1988 [18, 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 [22], 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 [4]: 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. A much simpler stylisation method is the first-order linear regression of voiced stretches of speech.

3 The Fujisaki model

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, amplitude 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 [11, 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 those presented by Pätzold 1991 [14] and by Narusawa et al. 2000 [12] are unfortunately not accessible and had to be left unconsidered in the present study albeit mentioned for reasons of completeness.

3.1 Mixdorff (2000)

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 [20]. 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 characterised 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 amplitude 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.

3.2 Kruschke (2001)

Following [6, 7], 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.

3.3 Schwarz (2009)

Following [19], an equiripple FIR high-pass filter with a 0.5 Hz cutoff frequency separates quadratic-spline interpolated F0 contours into high- (HFC) and low-frequency components (LFC). LFCs contain the phrase components and the speaker-dependent baseline frequency F_b set to the global minimum of the LFC. Accent and phrase components are extracted from the HFC and the LFC, respectively, by searching for local extremes. Since local maxima of the HFC 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 amplitudes A_p and local minima delimit the regions of phrase components. Phrase components will have at least a distance of 750 ms between them [10]. Finally, the extracted parameters are adjusted recursively in the least-squares sense. In contrast to [9] 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 HFC, each time segment given by T_1, T_2, and A_a represents an accent component that is optimized iteratively. Subtracting the fitted HFC from the F0 contour results in a modified LFC which is also optimized iteratively.

The procedure is carried out recursively as long as the MSE error of the modified LFC and the previous modified LFC is larger than 5%.

3.4 Pfitzinger

This method was developed in 2004 but is yet unpublished. All higher F0 modulation components above 3.5 Hz are removed from raw F0 values by applying sample-selective Fourier Transform [15] and successive frequency-domain low-pass filtering with a -18 dB/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, -18 dB/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.

4 Evaluation

The evaluation is based on the *IMS Radio News Corpus* [17]. 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 exist that were extracted automatically [9] and manually corrected following linguistic criteria [10] and using the interactive FujiParaEditor [8]. Although raw F0 data are provided with the corpus extracted in 10 ms steps via *get_f0* of *ESPS waves* [21], 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 resynthesize the F0 values by means of the Fujisaki model which is defined in the log F0 domain. Thus, our evaluation is based on the semitone scale.

5 Results

Fig. 1 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.

The number of extracted phrase commands was between 924 *(Pfitzinger)* and 1640 *(Schwarz)* while 1934 *(Pfitzinger)* to 4172 *(Kruschke)* accent commands were automatically detected. In the following sections these results were inspected in more detail via command alignment, and histograms of command rates, command amplitudes, and accent command durations.

5.1 Command alignment

In order to compare the agreement between the linguistically controlled parameter set *AutoFujiPos,Man* and the automatically estimated parameter sets we aligned accent and phrase com-

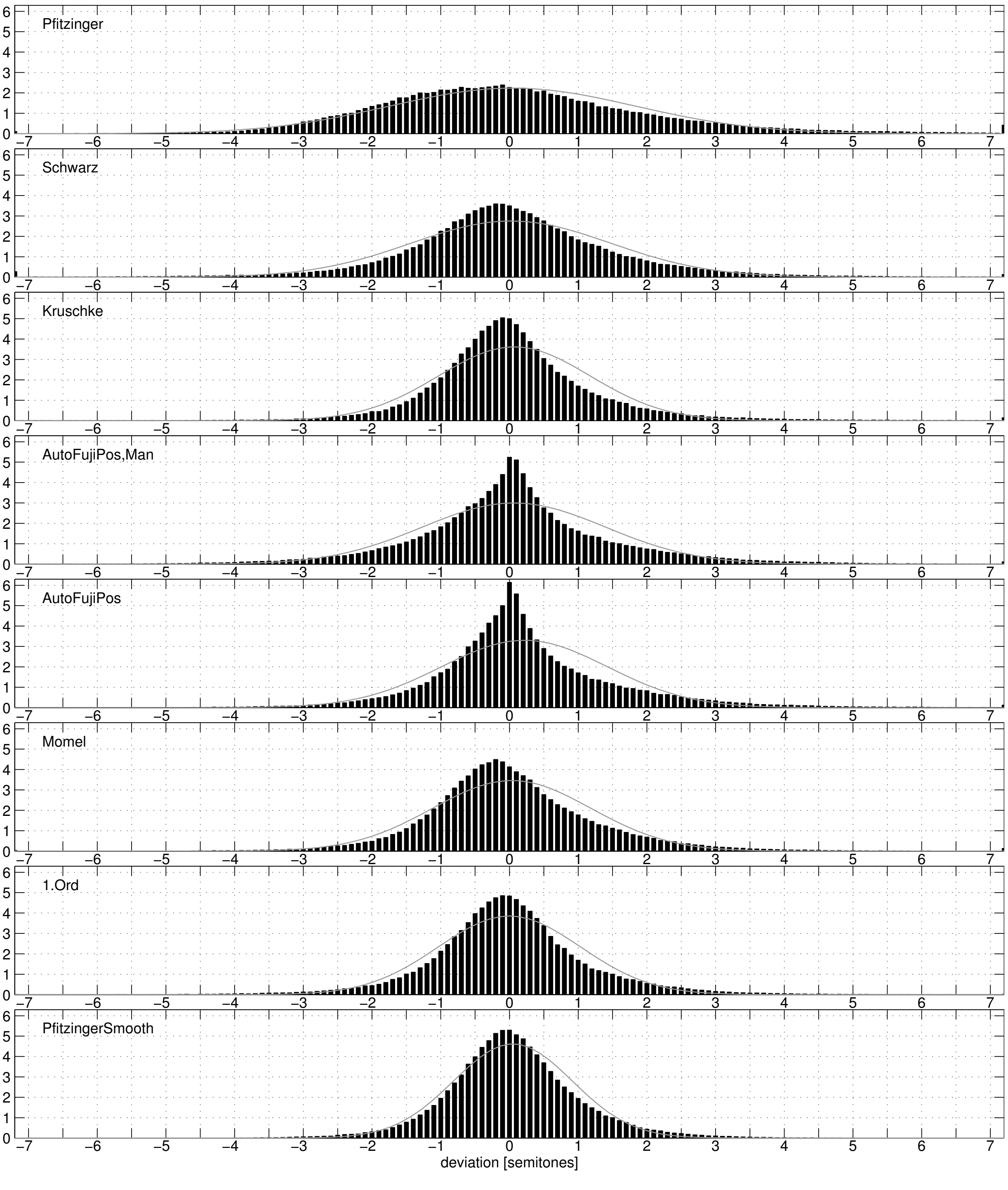

Figure 1 - Histograms of the frequency (in percent) and amount of deviation (in semitones) from the 167,039 reference F0 values.

mands from *AutoFujiPos,Man* with those produced by the other algorithms. To that end, accent commands that overlapped with accent commands from *AutoFujiPos,Man*, as well as phrase commands within a region of ±300 ms of a phrase command from *AutoFujiPos,Man* were classified as matching ones. Table 1 displays the results. It lists the total number of accent commands, the percentages of accent commands in the result from *AutoFujiPos,Man* aligned with accent commands calculated by the other approaches, the amount of overlap between commands for those cases in ms and the correlation between the values of accent command amplitude A_a for these accent commands. In the lower half, the table displays the total number of phrase commands, percentages of matching phrase commands, mean phrase command

	AutoFujiPos,Man	AutoFujiPos	Kruschke	Schwarz	Pfitzinger
Total number accent cmd.	3100	3386	4172	3201	1934
Aligned accent commands	–	97.3%	92.0%	78.0%	49.2%
Overlap (ms), mean/s.d.	–	232/136	198/110	211/136	158/89
rho(A_a)	–	0.944	0.803	0.527	0.857
Total number phrase cmd.	1227	1273	1332	1640	924
Aligned phrase commands	–	91.2%	51.3%	63.2%	45.8%
Distance (ms), mean/s.d.	–	92/74	139/87	132/84	144/86
rho(A_p)	–	0.932	0.608	0.348	0.874

Table 1 - Results of alignment between the parameter sets from *AutoFujiPos,Man* and those from the automatic algorithms *AutoFujiPos*, *Kruschke*, *Pfitzinger*, and *Schwarz*.

distances from the *AutoFujiPos,Man* reference as well as the correlation of phrase command amplitudes A_p. Since *AutoFujiPos,Man* was produced by manually editing the results from *AutoFujiPos* the match is the best for *AutoFujiPos*. *Kruschke* generally produces a high frequency of commands and therefore 92% of accent commands in *AutoFujiPos,Man* find a correspondence in *Kruschke*, whereas the amount of overlap is the largest in relationship with *Schwarz*. Matching accent commands from *Pfitzinger* yield the highest correlation as to their amplitudes. In the case of phrase commands the percentages of matches are considerably lower than for accent commands. *Kruschke*, *Pfitzinger* and *Schwarz* yield comparable mean distances. With respect to phrase command amplitudes *Pfitzinger* yields the highest correlation.

5.2 Command rates

Fig. 2 explains the overall poor performance of *Pfitzinger*: Accent and phrase command rates are generally too low compared with the other approaches. Furthermore, the variance of the accent command rate is too high which indicates that the command estimation method is not robust enough. The algorithm assigns too few commands as it ignores f0 peaks with accent amplitudes less than 0.2, which corresponds to 3.1 semitones according to

$$\mathrm{st} = A_a(12 \cdot \gamma \cdot \log_2 e).$$

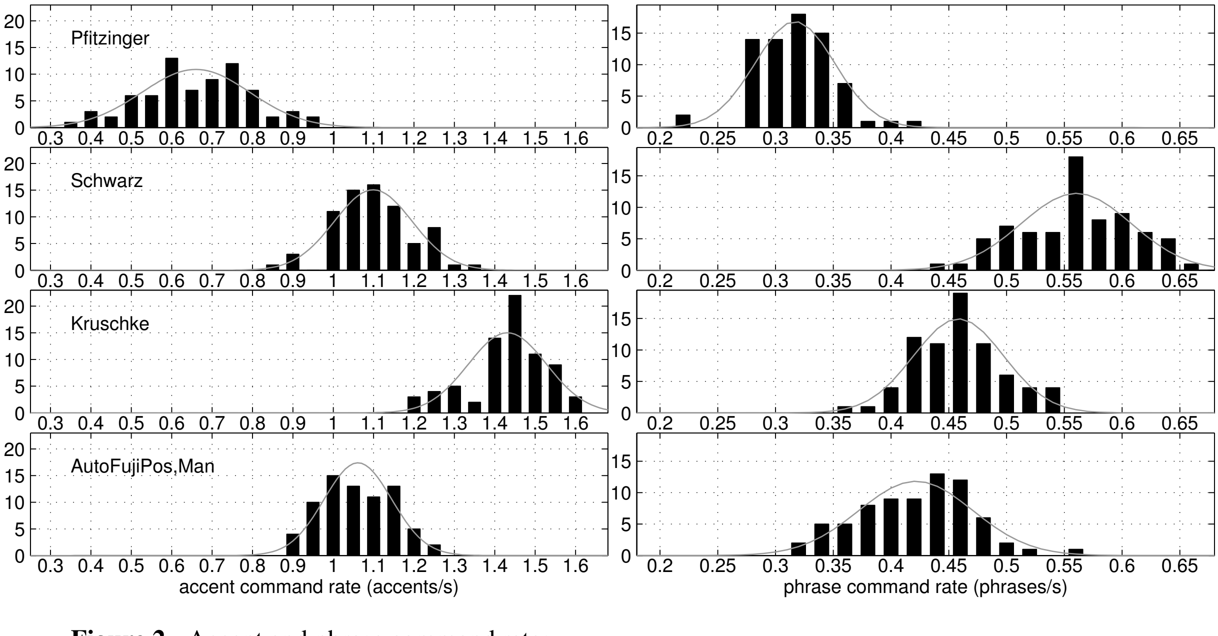

Figure 2 - Accent and phrase command rates.

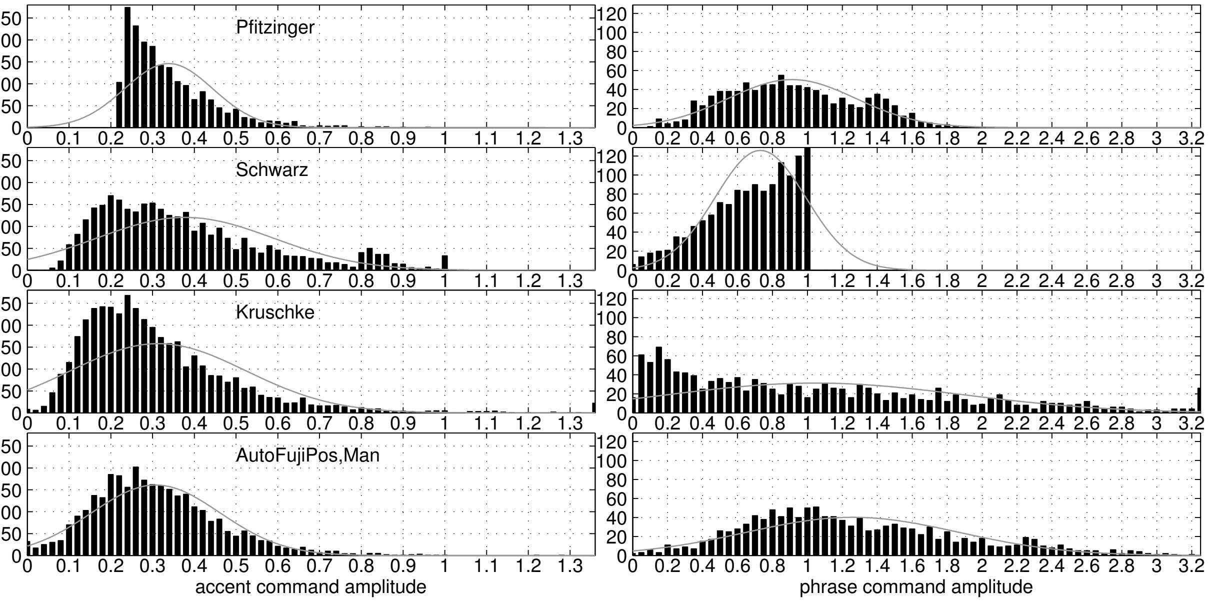

Figure 3 - Accent and phrase command amplitudes.

Lowering this threshold to 0.1, which amounts to 1.6 semitones and is, according to Isačenko & Schädlich 1964 [5], still sufficient to yield a perceivable prominence of a syllable, would significantly increase the number of commands. This consideration is confirmed by comparison with *AutoFujiPos,Man* and would yield a better F0 contour modelling.

5.3 Command amplitudes

The distribution of phrase command amplitudes shown in Fig. 3 suggests that *Schwarz* requires a high phrase command rate due to limiting the maximum phrase command amplitude to 1.0. Besides, the cluster above 0.8 in the accent command amplitude distribution might indicate accent commands that are actually utilized for supporting the modelling of the phrasal contour. In contrast to the other methods *Kruschke* frequently produces phrase commands with very low amplitudes below 0.2 as well as very high amplitudes above 3.

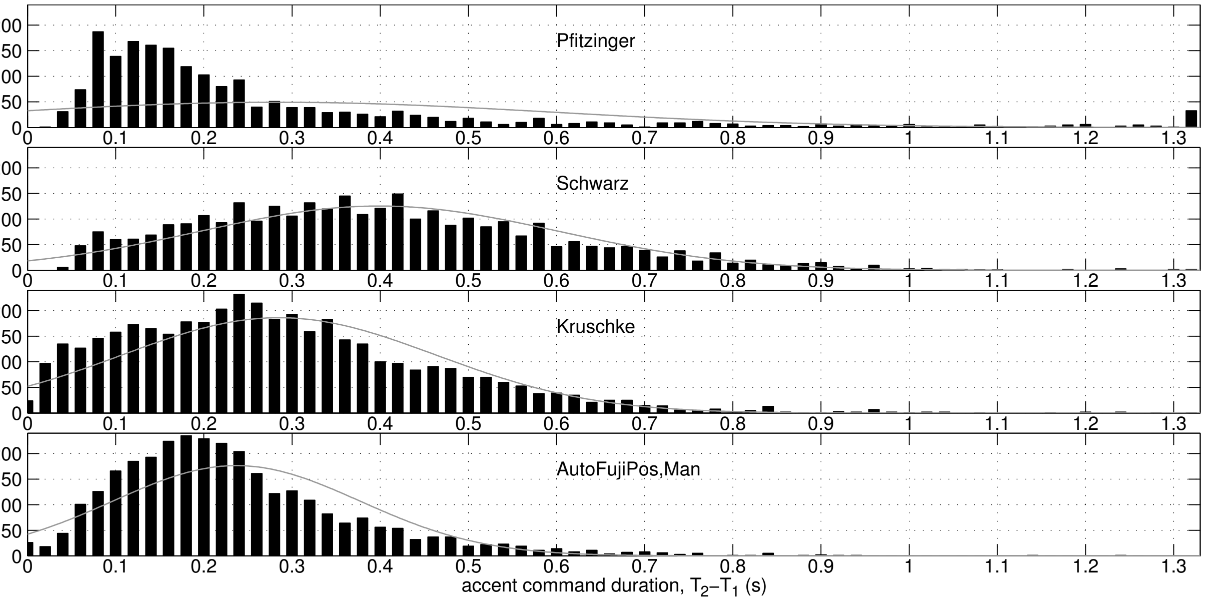

Figure 4 - Accent command durations.

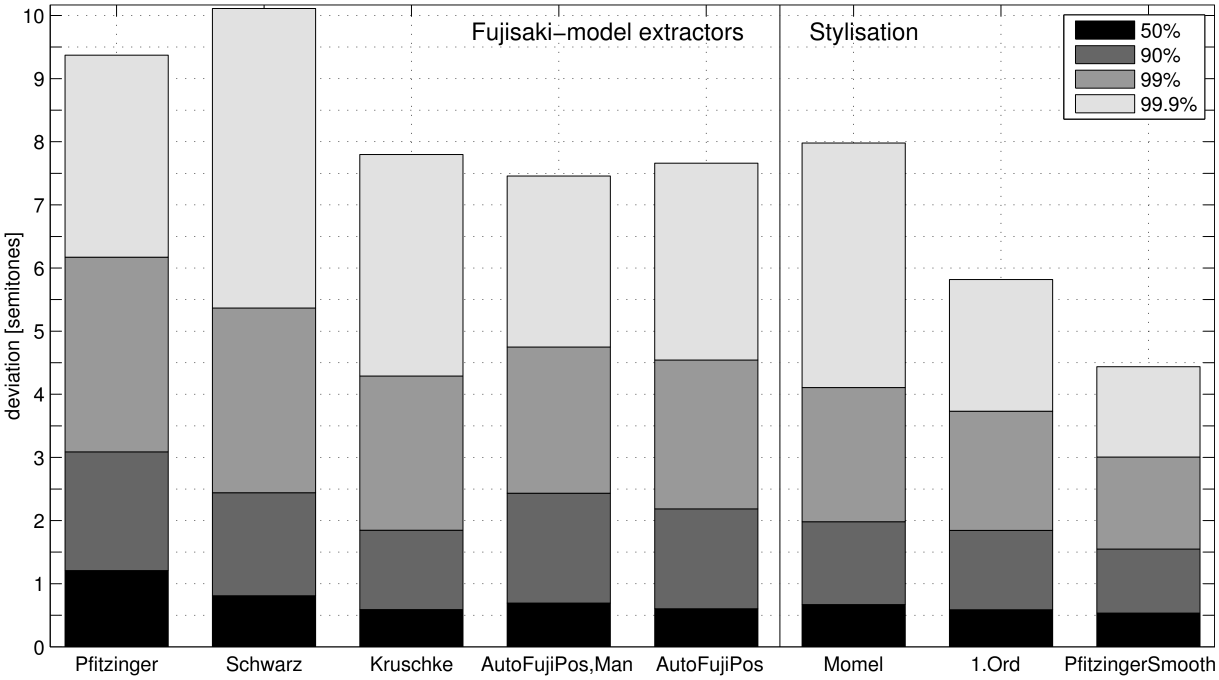

Figure 5 - Deviation in semitones between modelled F0 and reference F0. 50%, 90%, 99%, and 99.9% of all the data are below the F0 thresholds shown in the bars, respectively.

5.4 Accent command durations

Fig. 4 shows histograms of the accent command durations of four Fujisaki-model extractors. While *AutoFujiPos,Man* mainly produces accent commands with a duration of approx. 200 ms, which can be regarded as the reference, and a relatively small standard deviation, the other three extractors show larger standard deviations and significantly longer or shorter mean command durations with *Kruschke* being the closest to the reference data.

5.5 Overall error distribution

Finally, Fig. 5 presents a more condensed way of looking at the total 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-model extractors are below 2.5 semitones. 0.1% means that deviations are greater than 7.5 to 10 semitones only for 167 F0 values.

6 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 F_b, α, and β 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 the method of Schwarz depending on the particular sentence. In Kruschke's algorithm, besides F_b, also α and β are varied for each command 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:

Fujisaki-model	mean command rates ±std-dev.				model parameters		RMS	algorithmic
extractor	accents/s		phrases/s		F_b	α, β	error	complexity
Pfitzinger	0.66	±0.134	0.32	±0.035	const.	const.	1.99	low
Schwarz	1.10	±0.097	0.56	±0.048	var.	const.	1.61	very high
Kruschke	1.43	±0.097	0.46	±0.039	var.	var.	1.23	very high
AutoFujiPos,Man	1.06	±0.084	0.42	±0.049	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 [11, p. 116].

7 Conclusion and future research

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. 5, 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.

8 Acknowledgements

We are very grateful to Hans Kruschke for permitting us to use his Fujisaki-model extractor, to Hussein Hussein for running it on the reference data, to Jan Schwarz for the description and results of his extractor, and to Bernd Möbius from Bonn University for making the IMS Radio News Corpus available to us.

References

[1] Fujisaki, H.: *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, p. 347–355. Raven Press, New York, 1988.

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

[3] Hart, J. 't: *Psychoacoustic backgrounds of pitch contour stylisation*. IPO Annual Progress Report 11, Inst. for Perception Research, Eindhoven, 1976.

[4] HIRST, D. & R. ESPESSER: *Automatic modelling of fundamental frequency using a quadratic spline function.* Travaux de l'Institut de Phonétique d'Aix 15, Univ. de Provence, 1993.

[5] ISAČENKO, A. V. & H.-J. SCHÄDLICH: *Untersuchungen über die deutsche Satzintonation.* Akademie-Verlag, Berlin, 1964.

[6] KRUSCHKE, H.: *Advances in the parameter extraction of a command-response intonation model.* In *Int. Symposium on Intelligent Signal Processing and Communication Systems (ISPACS'01)*, p. 135–138, Nashville, Tennessee, 2001.

[7] KRUSCHKE, H. & M. LENZ: *Estimation of the parameters of the quantitative intonation model with continuous wavelet analysis.* In *Proc. of EUROSPEECH '03*, vol. 4, p. 2881–2884, Geneva, 2003.

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

[9] MIXDORFF, H.: *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, p. 1281–1284, Istanbul, 2000.

[10] MIXDORFF, H.: *An integrated approach to modeling German prosody.* w.e.b. Universitätsverlag, Dresden, 2002.

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

[12] NARUSAWA, S., H. FUJISAKI & S. OHNO: *A method for automatic extraction of parameters of the fundamental frequency contour.* In *Proc. of ICSLP 2000*, vol. 1, p. 649–652, Beijing, 2000.

[13] NOOTEBOOM, S. G.: *The prosody of speech: Melody and rhythm.* In HARDCASTLE, W. J. & J. LAVER (eds.): *The Handbook of Phonetic Sciences*, no 5 in *Blackwell Handbooks in Linguistics*, ch. 21, p. 640–673. Blackwell, Oxford, 1997.

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

[15] PFITZINGER, H. R.: *Removing hum from spoken language resources.* In *Proc. of ICSLP 2000*, vol. 3, p. 618–621, Beijing, 2000.

[16] PFITZINGER, H. R.: *Five Dimensions of Prosody: Intensity, Intonation, Timing, Voice Quality, and Degree of Reduction.* In HOFFMANN, R. & H. MIXDORFF (eds.): *Speech Prosody Abstract Book. Studientexte zur Sprachkommunikation*, vol. 40, p. 6–9. TUDpress, Dresden, 2006.

[17] RAPP, S.: *Automatisierte Erstellung von Korpora für die Prosodieforschung.* Arbeitspapiere (phonetikAIMS) 4(1), Inst. für Maschinelle Sprachverarbeitung, Lehrstuhl für experimentelle Phonetik der Univ. Stuttgart, 1998.

[18] SCHEFFERS, M. T. M.: *Automatic stylization of F0-contours.* In AINSWORTH, W. A. & J. N. HOLMES (eds.): *Proc. of SPEECH '88. 7th FASE Symposium*, vol. 3, p. 981–987, Edinburgh, 1988.

[19] SCHWARZ, J., M. TRAN & U. HEUTE: *Is the Fujisaki model a suitable (prosodic) model for the voice-conversion task?* In *Proc. of the Int. Conf. on Acoustics, and the 35th German Annual Conf. on Acoustics (DAGA)*, Rotterdam, The Netherlands, 2009.

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

[21] TALKIN, D.: *A robust algorithm for pitch tracking (RAPT).* In KLEIJN, W. B. & K. K. PALIWAL (eds.): *Speech coding and synthesis*, ch. 14, p. 495–518. Elsevier, New York, 1995.

[22] XU, Y. & X. SUN: *How fast can we really change pitch? Maximum speed of pitch change revisited.* In *Proc. of ICSLP 2000*, vol. 3, p. 666–669, Beijing, 2000.