Perceived Prominence Reflected by Imitations of Words with and without F0 Continuity

Hansjörg Mixdorff [1], *Angelika Hönemann*[1], *Oliver Niebuhr*[2] *and Christoph Draxler*[3]

[1] Department of Computer Science and Media, Beuth University Berlin, Germany
[2] Department of General Linguistics, ISFAS, Christian-Albrecht-University of Kiel, Germany
[3] Institute of Phonetics und Speech Processing, Ludwig-Maximilian-University Munich, Germany

{mixdorff|ahoenemann}@beuth-hochschule.de, niebuhr@linguistik.uni-kiel.de,
draxler@phonetik.uni-muenchen.de

Abstract

This paper continues our work on the perception of prominence as a function of *F0* continuity. In an earlier study the first author had shown that *F0* intervals occurring at lexically accented syllables – and measured using the amplitude of Fujisaki model accent commands – strongly contribute to the perceived prominence of that syllable. More recent work explored how *F0* continuity influenced prominence ratings of single word utterances. The outcome indicated that listeners made use of the physically available *F0* information and therefore words containing gaps in the contour were perceived as less prominent. It was also shown that subjects were able to interpolate missing parts as long as the *F0* peak was still present. The current study explores whether subjects compensate the lack of prominence in words containing *F0* gaps by asking them to produce a word with the same accent strength as that of an acoustic word stimulus, the acoustic word either being the same or different from the one they are asked to utter. We evaluated word durations, *F0* intervals and intensities of the responses as correlates of prominence and found that listeners indeed seem to adjust depending on the kind of stimulus they are presented.

Index Terms: prominence perception, Fujisaki model, word imitation

1. Introduction

It is a well-known fact that the information structure of an utterance is coded in the relative saliency of its lexical constituents. At the acoustic level we observe that accented syllables serve as anchoring points of this structure. They are emphasized or toned down by phonetic means. The perceptual correlate of this process is the so-called prominence, cf. [21]. Various segmental and supra-segmental factors have been shown to affect prominence, cf. [1,2,3]. In an earlier study [4], the first author and his co-worker investigated the relationship between perceived syllable prominence and the *F0* contour in terms of the parameters of the Fujisaki model [5]. The model was used to parameterize a subcorpus of the Bonn Prosodic Database [6]. Analysis showed that prominences labeled on a scale from 0-31 strongly correlated with the excursion of *F0* movements, as represented by the amplitude *Aa* of accent commands, but only when it was anchored to accented syllables. The fact that the prominence-lending *F0* movement does not necessarily take place inside the accented syllable indicates that the prominence judgment is partly guided by linguistic considerations. Evidence in support of this assumption has been presented for many languages, including German [7,8,9], which is the language of the present study.

Since the Fujisaki model fits natural *F0* contours continuously with a defined value for each speech frame, it smoothly interpolates or extrapolates *F0* gaps owing to unvoiced sounds. However, from a communicative point of view, the implicit claim of using the same underlying prosodic gesture for voiced and unvoiced sound sections is that listeners are also able to interpolate or extrapolate *F0* gaps. Recent evidence from a tonal scaling study [10] is inconsistent with this implicit claim. Subjects were presented with short resynthesized utterances and asked to rate the tonal height of accent-related *F0* rises. The rises led to a peak that was either present or absent due to an unvoiced stop consonant. Tonal height ratings were made and analyzed relative to reference utterances in which the *F0* rise was replaced by a flat *F0* stretch, yielding a constant tonal height. The findings of [10] suggested that the subjective continuity of pitch contours in speech is due to the fact that the auditory system simply ignores rather than fills *F0* gaps.

In [11] we examined the implications of these findings for the perception of word prominence. We investigated how gaps in the *F0* contour due to unvoiced consonants affect prominence perception, given that such gaps can either be filled or blinded out by listeners. For this purpose we created a stimulus set of real disyllabic words which differed in the quantity of the vowel of the accented syllable nucleus and the types of sub-sequent intervocalic consonant(s) and had participants rate pairs of these stimuli in a forced choice task for accent strength, that is, decide which word in a pair sounded more strongly accented. Results included, inter alia, that stimuli with unvoiced gaps in the *F0* contour are indeed perceived as less prominent. The prominence reduction was smaller for monotonous stimuli than for stimuli with *F0* excursions across the accented syllable. Moreover, in combination with *F0* excursions, it also mattered whether *F0* had to be interpolated or extrapolated, and whether or not the gap included a fricative sound. The results supported both the filling-in and blinding-out of *F0* gaps, which fits in well with earlier experiments on the production and perception of pitch.

The current experiment examines whether speakers compensate for the inherent difference in prominence when they produce words with or without continuous *F0* contours. To this end we asked participants of a production experiment to reply to an acoustic example of an isolated word and speak either the same or a different word with the same accent strength as the acoustic stimulus. Our hypothesis is that if speakers indeed compensate for the lack of sonority of a word it should exhibit a relatively higher *F0* target when speakers react to a high-sonority word than when a word with continuous *F0* contour is produced and vice-versa.

2. Stimuli and Experiment Design

The acoustic stimuli were taken from the set employed in [11]. They are shown in Table 1 with their critical segments set in bold in the SAMPA transcription.

Table 1. *The five target words and their critical segments.*

Word	SAMPA	English	Critical Segment	Energy (dB)
Rahmen	[R**a:m**@n]	frame	Long vowel (LV), voiced (vcd) nasal	74.23
Rasen	[R**a:z**@n]	lawn, to speed	LV, vcd fricative	72.10
Raten	[R**a:t**@n]	guess	LV,voiceless (vcl) plosive	68.10
Rasten	[R**ast**@n]	rest	short vowel (SV), vcl fricative+plosive	66.19
Ratten	[R**at:**@n]	rats	SV, long vcl plosive	50.68

For acoustic uniformity, the stimuli had been created using the *MBROLA* concatenative speech synthesizer driving the German male voice *de8* [12], starting with a monotonous stimuli at *F0*=100Hz. The long vowel [a:] was adjusted to duration of 244ms and the central consonant portion to 126ms. Using the *FujiParaEditor* [13] and Praat PSOLA resynthesis [14] we created additional stimuli by adding *F0* peak contours to the monotonous stimuli. The contour basis was laid by a phrase component, constant for all stimuli. One accent component with duration of 200ms was superimposed on the base contour. In [11] we had produced medial-peak and late peak versions, however, to facilitate imitation we decided to only employ medial-peak stimuli in the current study. In the long-vowel target words, the *F0* maxima of medial peaks were aligned close to the accented-vowel offset, in line with previous findings [15] and observations in citations forms.
Figure 1 displays the stimuli Rahmen, Rasen and Rasten with at *Aa*=0.6. The range of the *F0* peaks was varied in the form of three different accent command amplitudes (*Aa*): 0.4 (interval of approximately 3 semitones), 0.6 (about 3 semitones higher) and 0.8 (about 6 semitones higher). Hence, including the monotonous condition, we yielded four different acoustic versions of every word and hence a total of 100 acoustic stimulus/text pairs.

The experiment was performed using WikiSpeech [16], a framework developed at Ludwig-Maximilian-University Munich for web-based perceptual testing and speech data collection. Participants were asked to enroll on the WikiSpeech-Website and accept the download of the SpeechRecorder audio recording tool. When participants executed the program on their computers, the task was explained to them on the start-up screen. Every trial consisted of the automatic playback of the acoustic stimulus, that is, one of the synthetic stimuli, followed by the display of the word to be produced written as text. A traffic light indicating when to speak turned from red to green. After the recording the experiment continued with the next stimulus. Each of the acoustic words was either paired with its text equivalent or that of the other target words. Subjects were asked to pronounce the text word with the same accent strength as the acoustic stimulus. The duration of the recording slot was fixed at five seconds, and audio sampled at 44.1kHz/16bit.

In preparation for the experiment the subjects were first presented two sample recordings of an acoustic stimulus followed by the reply of a test subject. Subsequently a training session started involving four pairs of the same or different words.

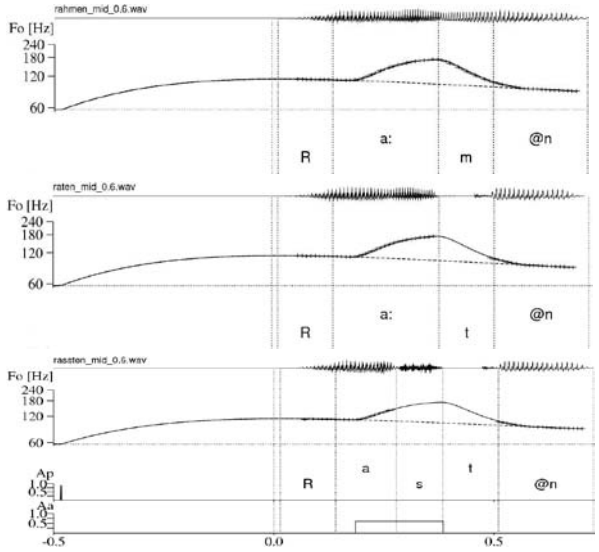

Figure 1. Examples of stimulus words Rahmen, Raten and Rasten with Aa=0.6. Panels display waveform (top), F0 contour (+++extracted, —modeled, middle), and underlying phrase/accent commands (bottom).

3. Acoustic Analysis of Stimuli

For the analysis we included recordings of 19 native speakers of German (10 females, 9 males, aged between 22 and 46), most of them students at Beuth University Berlin or Kiel University. Since the experiment was web-based we had little control of the equipment and recording environments. Although we had requested the use of headsets, many participants apparently used built-in microphones picking up environmental noise. As a consequence, the quality of recordings varied considerably with a wide range of gain settings, background noise and even audible audio compression effects. However, as we were mostly interested in prosodic parameters we sometimes admitted data which would not have qualified for fine acoustic analyses.

We first checked the audio files for the correct intended word, then admitted data sets with more than 70% correct word replies to further analysis. As a consequence, two participants were excluded (1 female, 1 male), leaving a total of 17 data sets or 1700 recordings.

Of the remaining sets 1615 contained the correct word (95%). Only three of the subjects were a 100% successful. Errors chiefly concerned the words [Rast@n] and [Rat@n]. In German, the words [Rast@n] – *"to rest"* and [Ra:st@n] – past tense of the verb 'rasen' – *"to speed"* are homographs and both spelled 'rasten'. Since we had anticipated this interference we had spelled [Rast@n] in the non-standard way 'rassten' to indicate the short vowel. However, the presence of 'Rasen' in the data set possibly triggered errors, especially when a long-vowel word had to be reacted to. Eleven of the participants produced [Rast@n] wrongly, in a total of 79 trials. [Rat@n] exhibited two-way confusions with [Ra:t@n], 31

times in this direction and seven times in the reverse. In most cases the vowel quantity of the acoustic stimulus matched the one produced erroneously. In contrast, participants reacted to the words [Ra:m@n] and [Ra:z@n] without difficulties. Many fewer cases of errors concerned empty audio files (5), stuttered words (7) or repetition of the acoustic stimulus word (8) when this was not requested.

Subsequently recordings were down-sampled to 16 kHz, and the first 0.85 seconds removed from the beginning as they often contained audible traces of the acoustic stimulus before the production of the subject and disturbed the subsequent forced alignment with the WEVOSYS LINGWAVES aligner[17]. Automatic phone segmentations were checked and corrected manually in the *PRAAT* TextGrid Editor [14]

We calculated word, syllable and phone durations based on the segmentations. *F0* values were extracted at a step of 10ms with *F0* floors and ceilings for male (50-300Hz) and female participants (120-400Hz) using *PRAAT* [14]. All *F0* contours were then subjected to Fujisaki model [5] parameter extraction [18]. Results were checked and if necessary corrected in the *FujiParaEditor*[19]. Intensity contours were extracted in *PRAAT* with default settings, and mean intensities in dB, as well as maxima employing parabolic interpolation were determined for each phone.

4. Results of Analysis

For our analysis we examined the Fujisaki model parameters *Aa* (accent command amplitude, as a measure of magnitude of *F0* excursion at the accented syllable), *Fb* (base frequency of the *F0* pattern), *Ap* (phrase command magnitude, a measure of the magnitude of *F0* reset before phrase onset), *T1rel* (the onset time of the accent command relative to the a-vowel onset). Each word exhibited at the most one phrase and one accent command, similar to the stimuli shown in Figure 1. *F0* contours of several reactions to monotonous stimuli were absolutely flat, so that neither a phrase nor an accent command was extracted.

Our measurements of the participants' productions were analyzed statistically with a three-way multivariate ANOVA based on the fixed factors *Word Heard*, *Word Realized*, and *F0 Range Heard*. The latter factor included four levels, i.e *Aa*= 0.0 (monotonous), as well as 0.4, 0.6, and 0.8 (henceforth "*F0* peak conditions"). The other two factors had five levels each, corresponding to the disyllabic target words Rahmen, Rasen, Raten, Rasten, and Ratten. The independent variables tested in the MANOVA were – for the sake of simplicity and relevance in the present paper – restricted to *Fb*, *T1rel*, *Aa*, *Ap*, *duration syllable 1*, *duration syllable 2*, as well as *duration vowel*, *mean intensity vowel*, and *max intensity vowel*, each of the latter three concerning the vowel in the first, accented syllable.

All three fixed factors had significant effects on many dependent variables. Our results section can only summarize a subset of these findings; and since the monotonous conditions differ considerably from the *F0* peak conditions, we will deal with the two conditions separately, starting with the *F0* peak conditions.

The most important finding in the *F0* peak conditions was that the *Aa* levels in the speakers' productions were highly significantly affected by both *Word Heard* ($F[4,1296]= 35.522$, $p<0.001$) and *F0 Range Heard* ($F[4,1296]= 7.582$, $p<0.001$). The four factor levels of *F0 Range Heard* were reflected in the speakers' *Aa* level. That is, an increase in the *F0 Range Heard* also resulted in a significant increase of the *Aa* level produced, independent of the target word uttered (cf. Figure 2, center). However, the speakers apparently underestimated the two higher *F0* ranges so that the *F0 Range Heard* levels of *Aa*= 0.4, 0.6, and 0.8 were on average replicated as 0.4, 0.5, and 0.6.

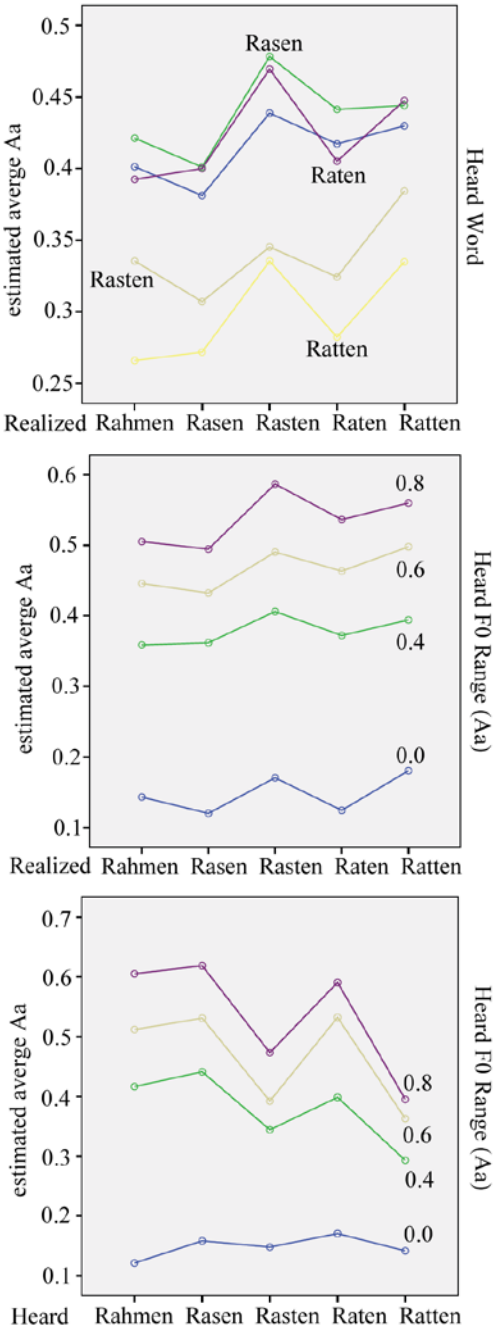

Figure 2: *Effects of* Word Heard *on average Aa levels in* Word Realized *(top), as well as combined effects of* F0 Range Heard *and* Word Realized *(center) or* Word Heard *(bottom) on the produced average Aa levels.*

In addition to these general parallels between heard and actually produced *Aa* levels, the *Aa* levels in the speakers' productions showed target word-specific differences. This effect was two-fold and is also reflected by a significant interaction between *Word Heard* and *F0 Range Heard* in

terms of *Aa* (which is at the same time the only significant two-way or three-way interaction in the MANOVA, $F[12,1296]= 3.813$, $p<0.001$). First, our speakers produced higher *Aa* levels across all target words when they previously heard the respective *F0* peaks in the more sonorous stimuli Rahmen, Rasen, and Raten than in the less sonorous stimuli Rasten and Ratten ($p<0.001$ in all cases of multiple post-hoc comparisons with Sidak correction, cf. Figure 2, bottom). The *Aa* level difference was on average 0.1. Moreover, the latter two words Rasten and Ratten yielded an additional weak *Aa* difference ($p<0.05$) with the Rasten stimuli triggering significantly higher *A*a level productions (of about 0.05) than the Ratten stimuli. Second, when we move on from the heard to the actually produced target words, we find an inverted pattern. That is, the *F0* peaks in the less sonorous target words Rasten and Ratten were produced with an *Aa* on average about 0.7 points higher than the more sonorous target words Rahmen, Rasen, and Raten ($p<0.05$ in all cases of multiple post-hoc comparisons with Sidak correction, cf. Figure 2, top and center). So, while successive desonorization and/or devoicing results in a reduction of the *F0* range at the level of perception, it seems to trigger an extension of the *F0* range at the level of production. It is not a usual finding in the area of segment-related microprosodic perturbations that production and perception findings go in opposite directions. The present outcome suggests the existence of a compensatory strategy in *F0* production.

The speakers' productions after monotonous stimuli differed considerably from those of the *F0* peak stimuli. The differences suggest that the speakers used a stylized, singing speech mode when producing target words after monotonous stimuli. For example, *Fb* (i.e. the base *F0*) was significantly higher and *Ap* significantly smaller after monotonous than after all other stimuli, which is among others reflected in the fixed factor *F0 Range Heard* ($F[3,1296]= 20.752$, $p<0.001$; $F[12,1296]= 103.065$, $p<0.001$). The same factor moreover showed that target word productions after monotonous stimuli were also softer in terms of a lower maximum intensity in the accented vowel (max_int_vowel; $F[3,1296]= 2.253$, $p<0.05$) and longer due to increased durations in both the first (dur_syl1; $F[3,1296]= 2.169$, $p<0.05$) and especially the second syllable (dur_syl2; $F[3,1296]= 5.095$, $p<0.001$).

In addition to these main findings, *Word Heard* caused an interesting additional effect, which could be characterized as "transfer" or "echo effects" on the produced target words. For example, after hearing the short-vowel stimuli Rasten and Ratten, speakers produced their target words with shorter first syllables and vowels, independently of the target word or the quantity of its accented vowel ($F[4,1296]= 10.965$, $p<0.001$; $F[4,1296]= 17.097$, $p<0.001$). So, even Rahmen was produced with a shorter first syllable and vowel when preceded by a stimulus like Rasten or Ratten. Concluding with effects of the fixed factor *Word Realized*, we found evidence for known effects of syllable structure and/or consonant type on *F0* peak timing [20], for example, in the form of a later *T1rel* timing relative to the accented vowel onset with increasing desonorization of the produced target words ($F[3,1296]= 5.309$, $p<0.001$). Finally, our measurements show that phonologically long vowels were actually produced longer and with an intrinsically greater vowel intensity than phonologically short vowels ($F[4,1296]= 580.056$, $p<0.001$; $F[4,1296]= 8.236$, $p<0.001$). This fact again underlines our speakers' competence and the validity of our findings.

5. Discussion and Conclusions

This paper presented results from an imitation study comparing reactions to word stimuli with either the same or a different word. Effects of *F0* on prominence perception are usually investigated with continuously voiced speech material. But what happens in more natural speech conditions, i.e. when parts of prominence-related *F0* movements are missing due to interruptions by voiceless sound segments? Do speakers and/or listeners compensate for the missing *F0* sections?

In accord with previous findings on English [10], our speakers' responses to the stimuli suggest that *F0* gaps are ignored rather than filled so that word intonations in which the peak maximum and/or adjacent high *F0* section are missing sounded lower and were hence imitated with lower *Aa* on the same or other words. As in our previous study, voiceless fricatives, here the Rasten/Ratten distinction, seem to be an exception. Speakers produced higher *Aa* as a reaction to Rasten stimuli than to Ratten stimuli, which suggests that they also heard more of the prominence-related *F0* peak in the Rasten stimuli, although the *F0* gap was physically equally long as in Ratten. That is, it seems that voiceless fricatives lend themselves better to a perceptual fill-in of *F0* gaps. This matches well with the notion of "segmental intonation", developed with reference to *F0* adjusted fricative productions by the third author [21].

However, even though *F0* sections masked by voiceless segments can be restored by listeners under certain circumstances, our present findings also suggest the existence of a compensatory mechanism. Speakers did not use the same *Aa* level across all target words when they imitated the prominence level of a given stimulus. Rather, they adjusted the *Aa* level of the realized target word in such a way that they used higher *Aa* levels for words with *F0* interruptions, hence compensating for their inherently lower prominence. As far as we know, such a compensatory mechanism is observed for the first time, although it is known for a long time that microprosodic variation is compensated for in speech production and/or perception. For example, listeners compensate for intrinsic *F0* variation [22] or intrinsic duration variation [23]. Such compensatory mechanisms are typically only partial; and a comparison between the *Aa* levels in perception and production/imitation suggests that the same also applies to our findings.

The fact that our speakers were, for example, able to replicate the *F0* ranges from 0.0 to 0.8 and clearly distinguished between short and long vowels supports the general validity of our data and shows moreover that the imitation-task paradigm is not just suitable for intonation, but also for prominence. Nevertheless, follow-up studies should aim at replicating the present findings, probably without the monotonous *F0* stimuli, as it is unclear how their very special character biased the prominence imitations in the *F0* peak stimulus conditions. Future work should also take up the observed "echo effects" (e.g., short/long vowels in the stimuli resulted in generally shorter/longer syllable productions), which could point to phonetic accommodation or the way in which acoustic properties are mapped onto perceptual measures, which is another promising field for imitation tasks.

6. Acknowledgements

Special thanks go to all students at Beuth University Berlin and Kiel University participating in this experiment.

7. References

[1] Fry, D.B., “Experiments in the perception of stress”, Language and Speech 1, 126-152, 1958.

[2] Gay, T., “Physiological and acoustic correlates of perceived stress. Language and Speech 21, 347-353, 1978.

[3] Koreman, J., Van Dommelen, W., Sikveland, R., Andreeva, B., Barry, W.J., ”Cross-language differences in the production of phrasal prominence in Norwegian and German”, in M. Vainio, R. Aulanko, O. Aaltonen (eds.): Nordic Prosody X. Frankfurt: Peter Lang, 139-150, 2009.

[4] Mixdorff, H., Widera, C., “Perceived Prominence in Terms of a Linguistically Motivated Quantitative Intonation Model”, Proc. Eurospeech 2001, Aalborg, Denmark, 403-406, 2001.

[5] Fujisaki, H., Hirose, K., “Analysis of voice fundamental frequency contours for declarative sentences of Japanese”, Journal of the Acoustical Society of Japan 5, 233-241, 1984.

[6] Heuft, B., “Eine prominenzbasierte Methode zur Prosodieanalyse und –synthese”, in W. Hess and W. Lenders (eds.): Computer Studies in Language and Speech, Vol. 2, Peter Lang, Frankfurt am Main, 1999.

[7] Niebuhr, O., “Interpretation of pitch patterns and its effects on accentual prominence in German”, Proc. Tone and Intonation in Europe 3, Lisbon, Portugal, 2008.

[8] Niebuhr, O., “F0-based rhythm effects on the perception of local syllable prominence”, Phonetica 66, 95-112, 2009.

[9] Kleber, F., Niebuhr, O., “Semantic-context effects on lexical stress and syllable prominence”, Proc. 5th Speech Prosody, Chicago, USA, 1-4, 2010.

[10] Barnes, J., Brugos, A., Veilleux, N., Shattuck-Hufnagel, S., “Voiceless Intervals and Perceptual Completion in F0 contours: Evidence from scaling perception in American English”, Proc. 16th ICPhS, Hong Kong, China, 108-111, 2011.

[11] Mixdorff, H. and Niebuhr, O. (2013). “The Influence of F0 Contour Continuity on Prominence Perception”, Proceedings of Interspeech 2013, Lyon, France.

[12] Dutoit, T., Pagel, V., Pierret, N., Bataille, F., van der Vreken, O., “The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes”, Proc. ICSLP, Philadelphia, USA, 1393-1396, 1996.

[13] Mixdorff, H., “FujiParaEditor”, http://public.beuth-hochschule.de/~mixdorff/thesis/fujisaki.html, 2009.

[14] Boersma, P., “Praat, a system for doing phonetics by computer”, Glot International 5, 341-345, 2001.

[15] Niebuhr, O., Ambrazaitis, G.I., “Alignment of medial and late peaks in German spontaneous speech”, Proc. 3rd Speech Prosody, Dresden, Germany, 161-164, 2006.

[16] Draxler, Chr. and K. Jänsch, “WikiSpeech - A Content Management System for Speech Databases”, Proceedings of Interspeech 2008, 1646-1649, Brisbane, 2008.

[17] http://www.wevosys.com/products/lingwaves/lingwaves.html

[18] Mixdorff H., “A novel approach to the fully automatic extraction of Fujisaki model parameters”, Proceedings of ICASSP 2000, vol. 3, 1281-1284, Istanbul Turkey, 2000.

[19] Mixdorff, H., “FujiParaEditor”, http://public.beuth-hochschule.de/~mixdorff/thesis/fujisaki.html, 2009.

[20] Wichmann, A., J. House und T. Rietveld,“Discourse constraints on F0 peak timing in English”, in A. Botinis (Hrsg.). Intonation. Dordrecht/Norwell: Kluwer Academic Publishers. 163-182, 2000.

[21] Niebuhr, O., “At the edge of intonation – The interplay of utterance-final F0 movements and voiceless fricative sounds”, Phonetica 69, 7-27, 2012.

[22] Niebuhr, O., “Intrinsic pitch in opening and closing diphthongs of German”, Proceedings of the 2nd international conference of speech prosody, Nara, Japan, 733-736, 2004.

[23] Gussenhoven, C., “Explaining two correlations between vowel quality and tone: the duration connection2”, Proceedings of the 2nd international conference of speech prosody, Nara, Japan, 179-182, 2004.