Perceived Prominence in Terms of a Linguistically Motivated Quantitative Intonation Model

Hansjörg Mixdorff
Faculty of Computer Sciences

Berlin University of Applied Sciences
mixdorff@tfh-berlin.de

Christina Widera
Institut für Kommunikationsforschung und Phonetik
University of Bonn, Germany
widera@ikp.uni-bonn.de

Abstract

The current study investigates the relationship between perceived syllable prominence and the *F0* contour in terms of the parameters of a linguistically motivated model of German intonation based on the Fujisaki formula. A subcorpus of the Bonn Prosodic Database was analyzed using the *F0* model, and normalized log syllable durations were calculated. Analysis shows that, for accented syllables, prominences strongly correlate with the amplitude *Aa* of accent commands underlying the *F0* movements in these syllables, whereas comparable *F0* movements in unaccented syllables have little effect on prominence. The influence of *Aa* versus syllable duration on prominence is greater in higher prominence classes. The fact that the *F0* movement does not necessarily take place in the accented syllable proper, indicates that the prominence judgment is partly guided by linguistic considerations. The results also show that *F0* modeling in TTS needs to be especially accurate in accented syllables which supports the main rationale of the *F0* model.

1. Introduction

It is an undisputed fact that the naturalness of synthesized speech strongly depends on its prosody. Still the mechanisms and relative contributions of prosodic features are far from being fully understood.

One important function of prosody is the highlighting of linguistic units. Investigations show that the perceived prominence of these units can be regarded as a gradual parameter. It is suited for describing the emphasis assigned to linguistic units in relation to their environment.

This paper focuses on the relationship between the perceived prominence of a syllable and two important prosodic features assigned to the syllable. These features are (1) the interval of a major *F0* transition connected to the syllable[1], as yielded by parametrising the *F0* contour with a quantitative model, (2) normalized log syllable durations.

1.1. Prosodic features examined in this study

Earlier work by Mixdorff was dedicated to a model of German intonation which uses the well-known quantitative Fujisaki formula [1] for parametrising *F0* contours, the **M**ixdorff-**F**ujisaki Model of **G**erman **I**ntonation (short **MFGI**, [2]). The model has been shown to be capable of producing close approximations to a given contour from two kinds of input commands: phrase commands (impulses) and accent commands (stepwise functions). A major attraction of the Fujisaki formula is the physiological interpretation which it offers for connecting *F0* movements with the activity of intrinsic larynx muscles [3].

In the framework of MFGI, following the works by Isačenko & Schädlich [4] and Stock & Zacharias [5], a given *F0* contour is described as a sequence of linguistically motivated tone switches, major rises and falls, which are modelled by onsets and offsets of accent commands connected to accented syllables, or by so-called *boundary tones* before prosodic boundaries. Hence the interval of a tone switch readily relates to the accent command amplitude *Aa* assigned to it. Tone switches constitute functionally distinct intonational elements, so-called 'intonemes'. In the scope of this study, we concentrate on the classes N↑ intoneme ('non-terminal intoneme' at phrase-medial accents, rising tone switch), and I↓ intoneme ('information intoneme' at declarative-final accents, falling tone switch).

1.2. Prominence of syllables

The notion of prominence that we follow is based on Fant. and Kruckenberg [6]. Three labelers had to judge the degree of prominence on the syllable level relative to the surrounding syllables on a scale from 0 to 31. Between subjects, the labeled prominences correlate strongly (rho > 0.8; [7]).

Earlier investigations show that the relation between prominence ratings and syllable duration, as well as *F0* peaks, described by parameters of a maximum based description of *F0* contours [8], are linear. However, prominence is also related to linguistic features (i.e. word class, position in a phrase, and focus). Furthermore, perceived prominence is reliably predicted from linguistic features, as well as from acoustic features [9].

Thus perceived prominence can be regarded as a gradual parameter integrating linguistic features and acoustic parameters.

[1] i.e. a rise and/or fall during the syllable proper or in the preceding or following one.

Since the Fujisaki model is inherently **production**-based, one major issue in this study is to establish the relationship between the amplitude parameter *Aa* and the **perceived** prominence of a syllable. Furthermore the implicit claim underlying MFGI that not all parts of the *F0* contour are 'equally important' is investigated. If the claim is correct, linguistically motivated *F0* transitions, i.e. tone switches, should strongly contribute to the perceived prominence of a syllable, whereas so-called 'pitch-interrupters' (Isačenko), *F0* transitions at non-accent syllables, should not.

2. Speech Material and Method of Analysis

The speech material was taken from the Bonn Prosodic Database (BPD, [10]). The BPD contains read speech of three German speakers. The subset is composed of isolated sentences, question-answer pairs, and short stories of one female speaker, and contains a total of 3401 syllables. Every syllable is assigned information about its position and its number in higher-level units (i.e. content words bearing lexical stress), its nucleus, as well as the number of phones it consists of.

The syllables are annotated with their word class and lexical word stress, as well as their prominence scaled from 0 to 31, as judged by three phoneticians. The prominence of a syllable is taken to be the median of the judgments.

Log syllable durations were computed from phone labels in the BPD and normalized to their phone count and the property of the nuclear vowel, being either schwa or non-schwa, the most important intrinsic features as shown in [11].

F0 contours were extracted at intervals of 10 ms, and the Fujisaki parameters determined using an automatic multi-stage approach [11]. All parameter sets were then checked visually as well as auditorily and errors corrected.

Tone switches were assigned to syllables by evaluating the timing of accent commands with respect to syllabic boundaries. In the case of potentially accented syllables (i.e. word accent syllables of content words) also the preceding and the following syllable were taken into account.

3. Results of Analysis

Figure 1 shows an example of analysis from the database displaying the utterance "Ist das die einzige Möglichkeit? - Ja, so ist es."-*"Is this the only possibility? - Yes, it is."* The figure displays from top to bottom: the speech waveform, the extracted and model-generated *F0* contours, the duration contour in terms of the syllabic z-score drawn as horizontal lines of the length of the respective syllable, the SAMPA transcription of the utterance, the underlying phrase and accent commands and the median perceived prominence. It can be seen that syllables with the highest prominence are accented syllables connected to tone switches (rising on [Ist], [aIn] and [m2:k], falling on [ja:] and [Ist]). High prominence is assigned to an accented syllable, even if the *F0* movement starts late in the syllable as in [m2:k] or in the following syllable as in [aIN].

A boundary tone can be observed on the syllable [keit] whose interval corresponds to the amplitude difference between the 3rd and 4th accent command in the utterance. Pre-boundary syllables, such as [kaIt] and [Es] exhibit relatively long syllable durations compared with accented or even unaccented syllables. [ja:] is a case of a syllable that is both accented and in a pre-boundary location, showing high prominence, high *Aa* as well as long duration.

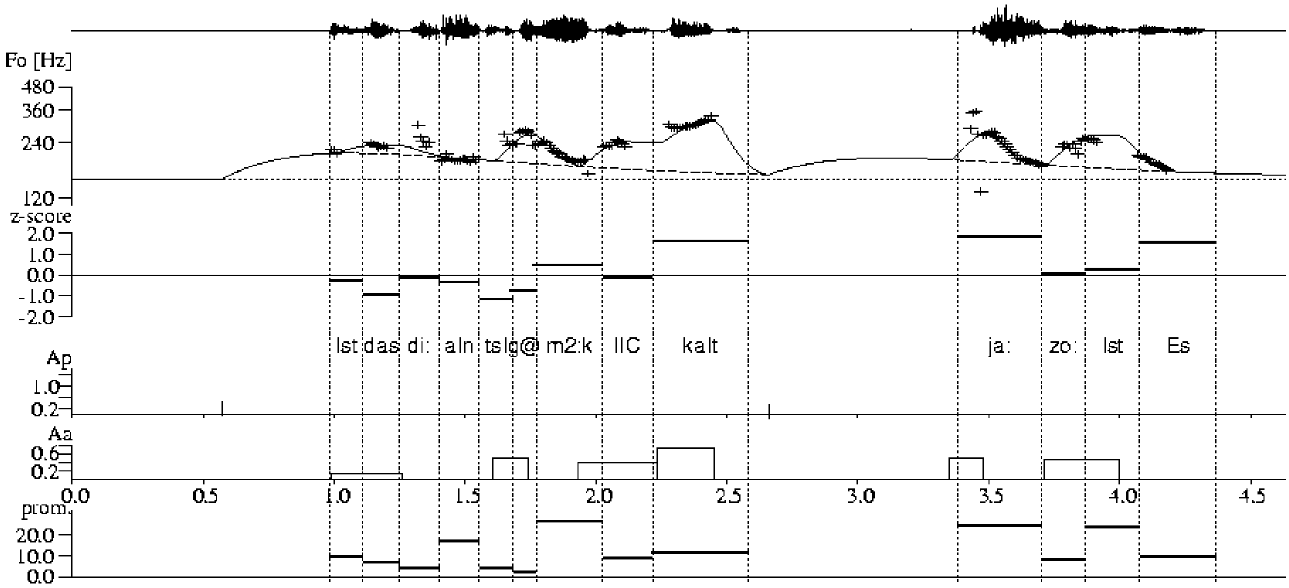

Figure 1: Example of of analysis from the database. Utterance: "Ist das die einzige Möglichkeit? - Ja, so ist es."-"Is this the only possibility? - Yes, it is." From top to bottom: speech waveform, extracted and model-generated F0 contours, duration contour (syllabic z-score), SAMPA transcription, underlying phrase and accent commands and median perceived prominence.

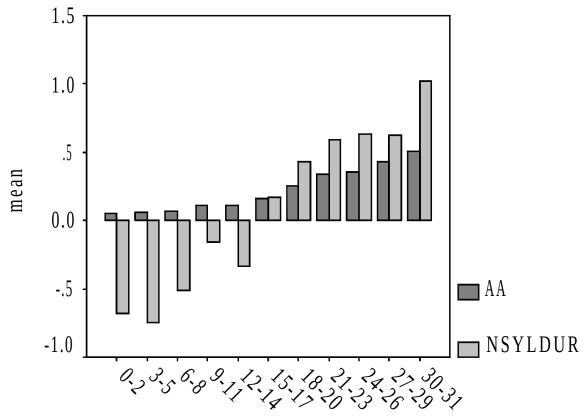

Figure 2: Mean accent commands amplitude (Aa) and normalized log syllable duration (NSYLDUR) depending on grouped prominence values.

3.1. Perceived prominence and acoustic parameters

Perceived prominences are evaluated in relation to the acoustic parameters *Aa* and normalized log syllable duration (*nsyldur*). The correlation over all syllables (rang correlation coefficient *ρ*) is about 0.5 for *Aa* (α <.01, n = 3401) and about 0.4 for *nsyldur* (α<.01, n = 3399).

These relatively low values may be explained by other influences such as phrase-final lengthening and boundary tones. If we only include syllables with lexical word accent, the correlation between prominence values and *Aa* (ρ = 0.6, α<.01, n = 2043) as well as *nsyldur* (ρ = 0.5, α<.01, n = 2043) increases. However, *Aa* is mostly related to higher prominence values (>15). In contrast to *Aa*, *nsyldur* correlates more strongly with lower prominence values (<16). Hence, weak perceived prominence grading is associated with duration and strong perceived prominence is mostly related to *F0* movements.

The relationship between perceived prominence and the two acoustic parameters can be regarded as nearly linear. This becomes more obvious when the prominence values are grouped (*Figure 2*).

3.2. Perceived prominence and tone switches

In this section we examine whether the linguistic notion of tone switches is reflected by prominence values. Prominence values and acoustic parameters of the linguistically motivated tone switches (I↓ intonemes and N↑ intonemes) are compared with the values of non-linguistic *F0* movements, i.e. rising and falling pitch interrupters (c.f. *Figure 3* and *Figure 4*).

The comparison of falling pitch interrupters (n = 294) with I↓ intonemes (n = 395) yields significant differences with respect to their mean prominence values (mean_P-interrupter = 10.9, mean_ I↓ = 22.6, t = 20.80, df = 43.72, α < 0.01) and their acoustic values (*Aa*: mean_P-interrupter = 0.35, mean_ I↓ = 0.39, t = 3.11, df = 687, α < 0.01; *nsyldur*: mean_P-interrupter = 0.22, mean_ I↓ = 0.58, t = 4.768, df = 554.83, α < 0.01). I↓ intonemes more strongly contribute to prominence than falling pitch interrupters. Comparable results are also found for the N↑ intonemes (n = 422) and rising pitch interrupters (n = 272). N↑ intonemes are associated with stronger prominence than not linguistically motivated *F0* rises (mean_ N↑ = 19.8, mean_P-interrupter = 13.0, t = 11.71, df = 411.51, α < 0.01). However, no significant differences are established for *nsyldur*.

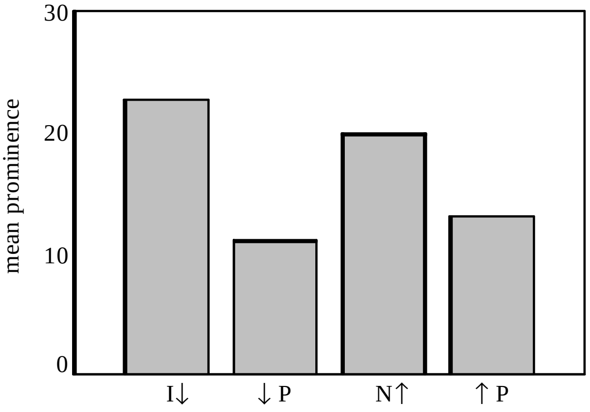

Figure 3: Mean prominence of tone switches (information intoneme (I↓), non-terminal intoneme (N↑)) and pitch interrupters (falling (↓P) and rising pitch interrupter (↑P)).

Furthermore the results show that the average prominence value in N↑ intonemes is lower than in I↓ intonemes. The significant difference between the prominence values (t = -7.74, df = 813.40, α < 0.01) is also reflected by the acoustic parameters *Aa* (mean_ N↑ = 0.37, mean_ I↓ = 0.39, t = -3.37, df = 788.98, α < 0.01) and *nsyldur* (mean_ N↑ = 0.30, mean_ I↓ = 0.58, t = -4.96, df = 762.67, α < 0.01).

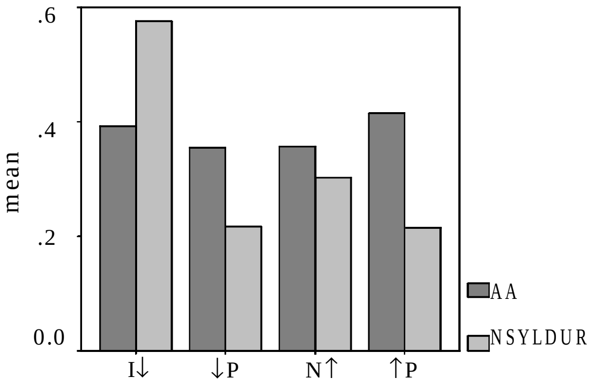

Figure 4: Mean Aa and normalized log syllable duration (NSYLDUR) depending on tone switches (information intoneme (I↓), non-terminal intoneme (N↑)) and pitch interrupters (falling (↓P) and rising pitch interrupter (↑P)).

Detailed analysis shows that prominences correlate with the timing of the accent command with respect to the syllable. Accented syllables of I↓ intonemes with an accent command on- and offset in the syllable proper are perceived more prominent (mean = 23.5, n = 135) than those with command on- and offset in the preceding (mean = 21.5, n = 63; t = 2.89, df = 196, $\alpha < 0.01$). Hence the position of the accent command with regard to the accented syllable influences prominence ratings. This result also indicates that, although the *F0* movement takes place in an unaccented syllable, the perceived prominence is assigned to the following accented syllable. In contrast, for N↑ intonemes, at least in our data, prominence judgments are not significantly affected by the timing of the accent command.

In conclusion, the comparison between prominence values assigned to different classes of *F0* transitions indicates that prominence ratings reflect a linguistically motivated association and interpretation of *F0* movements.

4. Discussion and Conclusions

Our results show that, for accented syllables, prominences strongly correlate with the amplitude *Aa* of accent commands underlying the *F0* movements in these syllables, whereas comparable *F0* movements in unaccented syllables have little effect on prominence.

Comparison between the influences of *Aa* and syllable duration on prominence shows a greater contribution of *F0* in higher prominence levels.

The fact that the *F0* movement does not necessarily take place in the accented syllable proper, indicates that the prominence judgment is partly guided by linguistic considerations. These findings are in accordance with those of earlier studies of prominence.

As we saw, the accent command amplitude parameter *Aa* of the production-based Fujisaki model is a very strong correlate of perceived prominence *wherever F0 movements can be motivated linguistically*. We may tentatively interpret this relationship as follows: While *Aa* - inter alia - reflects the 'relative importance' of accented constituent words in an utterance as intended by the speaker, prominence reflects the 'performance structure' of the utterance as perceived by the listener.

In the context of TTS our results indicate that *F0* modeling needs to be especially accurate in accented syllables which supports the main rationale of the *F0* model.

Future work should be dedicated to more closely controlled perception experiments evaluating prominence ratings on stimuli which are resynthesized with modeled *F0* contours. Especially the influence of accent command timing on prominence needs to be examined in more detail.

5. References

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

[2] Mixdorff, H., *Intonation Patterns of German - Model-based. Quantitative Analysis and Synthesis of F0-Contours.* PdD thesis TU Dresden, (http://www.tfh-berlin.de/~mixdorff/thesis.htm), 1998.

[3] 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 Physiology: Voice Production, Mechanisms and Functions* (pp. 347-355). Raven Press Ltd., New York, 1998.

[4] A.V. Isačenko and H.J. Schädlich, *Untersuchungen über die deutsche Satzintonation.* Akademie-Verlag, Berlin, 1964.

[5] E. Stock and C. Zacharias, *Deutsche Satzintonation*. VEB Verlag Enzyklopädie, Leipzig, 1982.

[6] Fant, G. and Kruckenberg, A., "Preliminaries to the study of Swedish prose reading and reading style". *Speech Transmission Laboratory – Quarterly Progress and Status Report, KTH Sockholm*, 2:1-83, 1989.

[7] Heuft, B. and Portele, T. "Synthesizing prosody: A prominence-based approach", *Proceedings ICSLP'96*, 1361-1364, Philadelphia, 1996.

[8] Heuft, B., Portele, T., Höfer, F., Krämer, J., Meyer, H., Rauth, M., and Sonntag, G., "Parametric description of F0-contours in a prosodic database", *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Stockholm KTH, 378-381, 1995.

[9] Portele, T., Heuft, B., Widera, C., Wagner, P., and Wolters, M., "Perceptual Prominence", in W. Sendlmeier (ed): *Speech and Signals. Aspects of Speech Synthesis and Automatic Speech Recognition.* Dedicated to Wolfgang Hess on his 60th birthday (pp.97-115), Theo Hector, Frankfurt am Main, 2000.

[10] 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.

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

[12] Mixdorff, H. and Fujisaki, H., "A quantitative description of German prosody offering symbolic labels as a by-product", in *Proceedings ICSLP 2000*, vol. 2., 98-101. Beijing, China, 2000.