PERCEIVED PROMINENCE AND ACOUSTIC PARAMETERS IN AMERICAN ENGLISH

Thomas Portele

Institut für Kommunikationsforschung und Phonetik, University of Bonn
tpo@ikp.uni-bonn.de

ABSTRACT

This paper describes the relationships between perceived prominence as a gradual value and some acoustic-prosodic parameters. Prominence is used as an intermediate parameter in a speech synthesis system. A corpus of American English utterances was constructed by measuring and annotating various linguistic, acoustic and perceptual parameters and features. The investigation of the corpus revealed some strong and some rather weak relations between prominence and acoustic-prosodic parameters that serve as a starting point for the development of prominence-based rules for the synthesis of American English prosody in a content-to-speech system.

1. MOTIVATION

Perceived syllable prominence was interpreted as a gradual parameter by Fant & Kruckenberg [1]. Subjects rated the perceived prominence of syllables on a 30-point scale. The authors investigated a small corpus of Swedish and found linear relationships between perceived prominence and acoustic and articulatory parameters. They also investigated the consistency of their labellers and obtained high correlations; this was confirmed by de Pijper & Sanderman [2] for boundary prominence. Grover et al. [3] showed that the reliability of word prominence ratings is higher for a 10-point scale than for a 4-point scale.

Heuft et al. [4] annotated a corpus of more than 11000 German syllables with perceived prominence values between 0 and 31. For three labellers correlation coefficients (cc) of around 0.8 were obtained. The authors further investigated the relationship between prominence and acoustic parameters and found moderate but highly significant correlations for syllable duration (cc = 0.55) but rather inconclusive results for F0 peak parameters like peak height [5]. A subsequent investigation of spectral parameters like formant position and relative energy distribution revealed a number of relations to perceived prominence: Higher prominence values cooccur with formant values closeto their respective target values, and energy in the vicinity of the second formant increases with higher prominence values.

In [6] we concluded that prominence may serve as an intermediate parameter for speech synthesis that may be used to explain a number of prosodic effects. For speech synthesis, prominence was defined as

> "a quantitative parameter of a syllable or a boundary that describes markedness relative to surrounding syllables and boundaries, respectively. As a system parameter, it corresponds to the perceived prominence of a human listener. When a sequence of syllables and boundaries, each one supplied with a prominence value, is synthesized, the listener should perceive the prominence relations between the synthesized syllables and boundaries in a way that is implied by the numerical values of the system parameter." [6].

A prosody control system for speech synthesis was constructed based on the relations in the German corpus [4], it is included in the German synthesis of the VERBMOBIL face-to-face translation project [7]. Prominence values are predicted from part-of-speech tags (similar to [8]) using syntactic and semantic information from higher-level linguistic modules (Fig. 1).

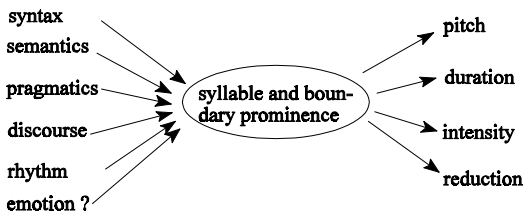

Figure 1: Structure of a prominence-based synthesis system (from [6])

Currently, a similar system for American English is being constructed [9]. In order to extend the prominence-based approach in speech synthesis to other languages it is necessary to assess the relationship between acoustic parameters and perceived prominence for each language. This paper describes the corpus, the annotation, and some results.

2. CORPUS

The corpus consists of 443 question-answer pairs. They were read by two speakers, recorded, and segmented [10]. Pitch accents and boundary tones were marked manually. Each pitch accent was described by four automatically extracted parameters [11].

Acoustic parameters like formant position and bandwidth, F0, and energy distribution were computed automatically using procedures supplied by the ESPS program. A detailed description of the corpus and its construction is given in [10].The most tedious task was the annotation of perceived prominence. As cited above, listeners' agreement is high. Therefore, only one labeller rated the perceived prominence of more than 19 000 syllables. She is a native speaker of American English but not phonetically trained.

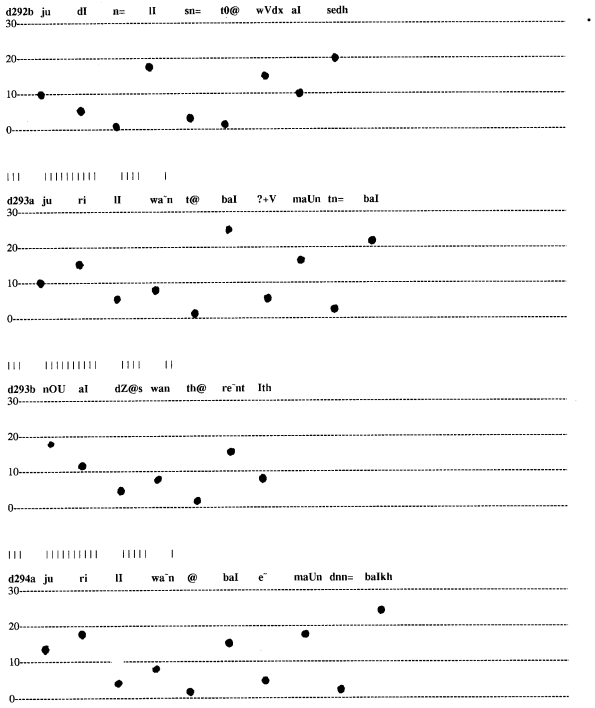

Figure 2: Data sheet used for labelling perceived prominence. The dots indicate the prominence values perceived by the labeller.

A graphical labelling scheme was used to provide an intuitive visualization of the relative nature of prominence (Fig. 2). The labels were scanned, processed and automatically converted to prominence values.

3. RESULTS

The acustic-prosodic parameters duration, F0, energy, and formant frequencies were related to the perceived prominence values. We investigated the relationships for both speakers and compared the results with those obtained for German.

3.1. Duration

Syllable duration is to a large extent explained by three parameters: prominence, number of segments, phrase finality. A simple linear regression yields r^2 values of 0.72 for speaker 1 and 0.62 for speaker 2. Correlation coefficients between syllable duration and perceived prominence are 0.63 for speaker 1 and speaker 2. The relationship between prominence and syllable duration is fairly linear (Fig. 3), and the correlation is stronger than in German [6].

A normalisation for phrase position and number of segments does not lead to higher values due to a modest correlation between segment number and prominence; longer syllables are more prominent than shorter ones (cc= 0.35 for speakers 1 and 2).

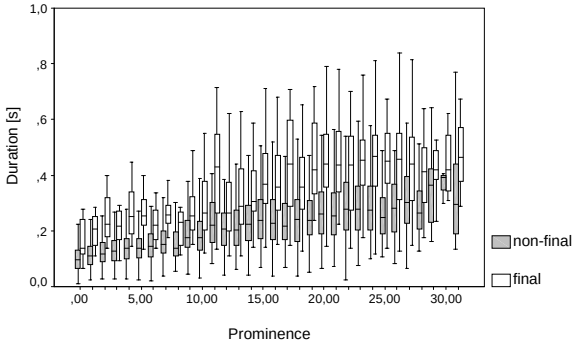

Figure 3: Boxplot of syllable duration depending on perceived prominence for final and non-final syllables (speaker 2).

The results for segment duration are similar. For speaker 1, the cc for vowels is 0.55 and for consonants 0.33, while for speaker 2 the respective values are 0.53 and 0.32. Although English does not distinguish phonologically between long and short segments, phonetic differences exist. But a normalisation with segment-specific z-score values is inadequate, because results for e.g. /ə/, which is not prominence-dependent (cc = 0.25) are mixed with results for prominence-sensitive vowels like /æ/ (cc = 0.64). Thus, for each segment that appeared more than 100 times in the corpus individual cc were computed; the mean cc for vowels is 0.42 (speaker 1) or 0.44 (speaker2), and the mean cc for consonants is 0.31 for speaker 1 and 0.32 for speaker 2. Vowels with longer inherent durations are more likely to be found in prominent syllables, while the mean individual cc for consonants is equivalent to the pertinent overall cc.

In comparison to the results obtained for German [4] the durations for American English syllables and segments are related more closely to perceived prominence.

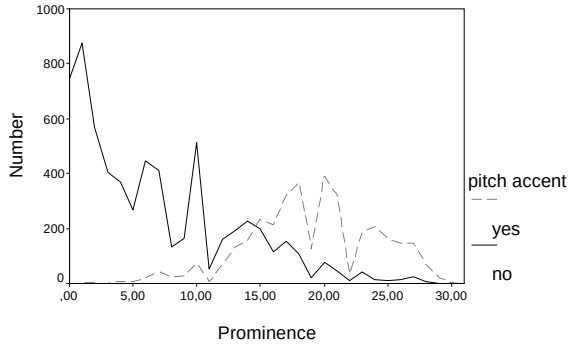

Figure 4: Number of prominence ratings depending on the presence of a pitch accent for speaker 1.

3.2. F0

There exists a clear dependency between the presence of a pitch accent and the perceived prominence of the associated syllable (Fig. 4). A Linear Discriminance Analysis predicts the presence

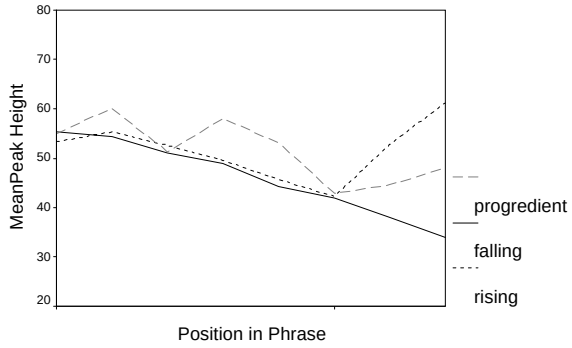

Figure 5: Effect of declination for speaker 1 depending on phrase type.

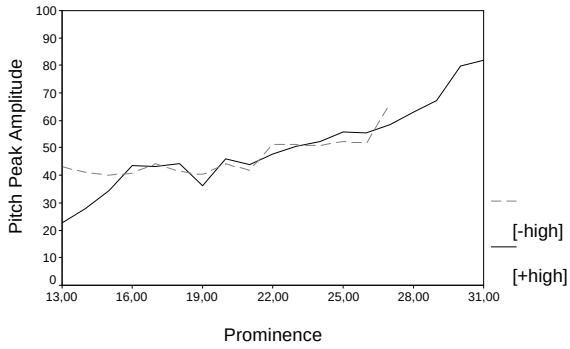

Figure 6: Mean pitch peak height for prominence values depending on the presence of the feature [+high] for speaker 1.

of a pitch accent from the prominence values correctly in 84.3 % (speaker 1) and 87.7 % (speaker 2). These values are quite high and in the range of those for German (88.6 % [6]). The mean prominence values for unaccented syllables are 6.7 (speaker 1) and 5.1 (speaker 2); accentend syllables receive a mean prominence score of 18.9 (speaker 1) and 18.4 (speaker 2). These results confirm the well-known fact that a pitch accent is an important cue for prominence.

The relationships between perceived prominence and properties of a pitch accent are less clear. Peak height is established to influence perceived prominence [12], but it is unclear what the mental reference for peak height is. The correlation coefficient between peak height and prominence is rather low (0.39 for speaker 1 and 0.24 for speaker 2). The most obvious normalisation is one for the effect of declination, as suggested by Figure 5. However, the cc of a linear regression is only 0.28 for speaker 1, and the cc between peak height and prominence increases only from 0.39 to 0.45 for the normalised values. For speaker 2, the linear regression (cc = 0.17) only increases the cc from 0.24 to 0.28. If only falling contours are analyzed by the same procedure (linear regression), no significant changes are visible. For the German corpus a normalisation for accent number (downstep) was attempted, with even less convincing results. Therefore, no normalisation was carried out for the following analysis.

Recent investigations in the perceived differences between pitch peaks [13] indicated that listeners are more sensible to differences in peak height if the pertinent peak has reached a certain height. This influenced the development of an intonation model [15] where the most prominent syllable of a prosodic phrase has a special feature [+high], and its height is related to the perceived prominence of the syllable while other pitch accents do not cue prominence with anything else than their existence. The distinction between [+high] and [-high] peaks can be found in the data for speaker 1, where the height of the most prominent peaks in a prosodic phrase (n = 787) correlate with prominence with a cc of 0.46, while theheights of the other peaks (n = 1381) and prominence have a cc of only 0.14 (Fig. 6). For speaker 2 this effect, however, is only marginal (cc = 0.15 for [-high] peaks, cc = 0.27 for [+high] peaks). For German, correlations (0.1 vs. 0.4) similar to those obtained for speaker 1 were found. No relations between peak position or peak slope [11] and perceived prominence could be established.

3.3. Spectral Parameters

Spectral parameters investigated are the positions of the first three formants, overall energy, and relative energy between 0-1 kHz, 1-2 kHz, 2-4 kHz and 4-8 kHz. No normalisations were carried out for formant frequencies because coarticulatory influences are difficult to control for in a large corpus. The same holds for the relative energy distributions, where no interactions with e.g. mean F0 or position in phrase were found. Overall energy was normalised by phrase position. Only vowels were investigated.

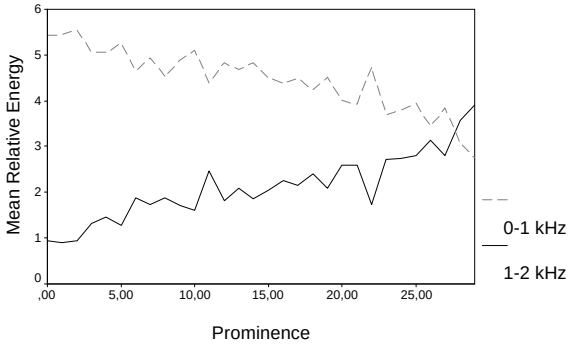

Figure 7: Relative energy depending on perceived prominence for /æ/.

Some vowels reflect prominence changes in their spectral parameters to a large extent, while others are nearly unaffected. These characteristics are consistent between both speakers. The vowel /æ/ is especially sensitive: perceived prominence correlates with position of first formant (cc = 0.6 for speaker 1, cc= 0.58 for speaker 2), energy between 0-1 kHz (cc = -0.62 for speaker 1, cc = -0.6 for speaker 2), and energy between 1-2 kHz (cc = 0.65 for speaker 1, cc = 0.58 for speaker 2) (Fig. 7). A general tendency was found for open vowels to raise F1 and for back vowels to lower F2 with increasing prominence. Relative energy between 0-1 kHz is reduced, and relative energy between 1-2 kHz is increased. Overall energy increases with perceived prominence, but the correlation is segment- and speaker-dependent (Table 1).

Segment	Speaker 1	Speaker 2
æ	0.36	0.41
a	0.07	0.35
a͡ɪ	0.29	0.23
e	0.34	0.41
e͡ɪ	0.19	0.29
i	0.09	0.25
ɪ	0.11	0.41
ɔ	0.12	0.29
ɔ͡ʊ	0.30	0.28
u	0.21	0.31
ʊ	0.36	0.38
ʌ	0.41	0.51

Table 1: Correlation coefficients between overall energy and perceived prominence for 12 vowels.

4. DISCUSSION

The results indicate that a number of relations between perceived prominence and acoustic-prosodic parameters exist. This confirms the prominence-based approach for speech synthesis [6]. Prominence can be used as an intermediate parameter; pitch accents and syllable durations can be assigned with high reliability. Segment durations is related a number of factor [14]s, and prominence is one of them. The same holds for spectral parameters like overall energy, spectral balance, and formant positions, but here, other influences can be more important (e.g. coarticulatory influences).

The relation between features of a pitch accents and perceived prominence is not easy to describe[12]. The differentiation between high and ordinary peaks may be a step towards a better understanding of this relation.

Currently, rules based on prominence values are formulated and integrated into the Verbmobil synthesizer for American English.

5. ACKNOWLEDGEMENTS

I thank two reviewers for their comments on the initial summary. Anja Elsner coordinated the corpus construction where the whole synthesis group played an active role. Jörg Bröggelwirth computed a number of parameters. Petra Wagner helped me with the investigation and works on the rule system. I want to thank them and the whole group at the IKP.

6. REFERENCES

[1] Fant G., Kruckenberg A.: "Preliminaries to the study of Swedish prose reading and reading style" *STL-QPRS* **2/89**, Stockholm, 1-83, 1989

[2] de Pijper J., Sanderman A.: "On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues" *J. Acoust. Soc. Am.* **96**, 2037-2047, 1994

[3] Grover C., Heuft B., van Coile B.: "The reliability of labelling word prominence and prosodic boundary strength" *Proc. ESCA-Workshop on Intonation*, Athens, 165-168, 1997

[4] Heuft B., Portele T., Höfer F., Krämer J., Meyer H., Rauth M., Sonntag G.: "Parametric description of F0-contours in a prosodic database" *Proc. XIII: ICPhS*, Stockholm, 2:378-381, 1995

[5] Heuft B., Portele T.: "Synthesizing prosody - a prominence-based approach" *Proc. ICSLP 96*, Philadelphia, 1996

[6] Portele T., Heuft B.: "Towards a prominence-based synthesis system" *Speech Communication* **21**, 61-72, 1997

[7] Bub T., Schwinn J.: "VERBMOBIL: The Evolution of a Complex Large Speech-to-Speech Translation System" *Proc. ICSLP 96*, Philadelphia, 1996

[8] Widera C., Portele T., Wolters M.: "Prediction of Word Prominence" *Proc. Eurospeech 97*, Rhodes, 999-1003, 1997

[9] King S., Portele T., Höfer F.: "Speech Synthesis Using Non-Uniform Units in the Verbmobil Project" *Proc. Eurospeech 97*, Rhodes, 569-572, 1997

[10] Elsner A., Wolters M., Portele T., Rauth M., Sonntag G.: "Designing and labelling a prosodic database for American English" *Proc. Workshop on Language Resources*, Granada, 1998

[11] Portele T., Heuft B.: "The maximum-based description of F0 contours and its application to English" (these proceedings)

[12] Gussenhoven C., Repp B.H., Rietveld A., Rump H.H., Terken J.: "The perceptual prominence of fundamental frequency peaks" *J. Acoust. Soc. Am.* **102**, 3009-3022, 1997

[13] Portele T.: "Perceptual evidence for accent categories: preliminaries and first results" *Proc. ESCA-Workshop on Intonation*, Athens, 271-274, 1997

[14] van Santen J.P.H.: "Contextual effects on vowel duration" *Speech Communication* **11**, 513-546, 1992

[15] Portele T.: "Eine perzeptiv motivierte Beschreibung der Intonation des Deutschen" (to appear in *Proc. ESSV 98*, Dresden)