On Automatic Prominence Detection for German

Tamburini, Fabio; Wagner, Petra

On Automatic Prominence Detection for German

Abstract

Perceptual prominence is an important indicator of a word's and syllable's lexical, syntactic, semantic and pragmatic status in a discourse. Its automatic annotation would be a valuable enrichment of large databases used in unit selection speech synthesis and speech recognition. While much research has been carried out on the interaction between prominence and acoustic factors, little progress has been made in its automatic annotation. Previous approaches to German relied on linguistic features in prominence detection, but a purely acoustic method would be advantageous. We applied an algorithm to German data that had been previously used for English and Italian. Both the algorithm and the data annotation encode prominence as a continuous rather than a categorical parameter. First results are encouraging, but again show that prominence perception relies on linguistic expectancies as well as acoustic patterns. Also, our results further strengthen the view that force accents are a more reliable cue to prominence than pitch accents in German.

On Automatic Prominence Detection for German

Fabio Tamburini [1], *Petra Wagner* [2]

[1] Department of Linguistics and Oriental Studies, University of Bologna, Italy
[2] Institute of Communication Sciences, University of Bonn, Germany
fabio.tamburini@unibo.it, pwa@ifk.uni-bonn.de

Abstract

Perceptual prominence is an important indicator of a word's and syllable's lexical, syntactic, semantic and pragmatic status in a discourse. Its automatic annotation would be a valuable enrichment of large databases used in unit selection speech synthesis and speech recognition. While much research has been carried out on the interaction between prominence and acoustic factors, little progress has been made in its automatic annotation. Previous approaches to German relied on linguistic features in prominence detection, but a purely acoustic method would be advantageous. We applied an algorithm to German data that had been previously used for English and Italian. Both the algorithm and the data annotation encode prominence as a continuous rather than a categorical parameter. First results are encouraging, but again show that prominence perception relies on linguistic expectancies as well as acoustic patterns. Also, our results further strengthen the view that *force accents* are a more reliable cue to prominence than *pitch accents* in German.

Index Terms: prosody, prominence, German

1. Introduction

Perceptual prominence of linguistic units such as syllables or words can be regarded as the unit's degree of standing out of its environment [1]. The phonological equivalences of prominence are linguistic *pitch accents* and *force accents* [2, 3, c.f. below]. In speech technology, automatic annotation of prominence is useful for both recognition and synthesis applications. In recognition, prominence detection can be crucial because it fulfils important linguistic functions such as indicating semantic or pragmatic focus, lexical stress or boundaries [4]. State-of-the-art unit selection synthesis relies on databases too large to be manually annotated, both on segmental and suprasegmental level. In this paper, the automatic prominence annotation is performed on a read-speech database for German. The results are compared with manual annotations of prominence performed on a continuous scale.

1.1. Prominence and acoustic parameters

One of the major challenges in predicting syllable prominence is the disentangling of various sources of influence such as fundamental frequency excursions, duration, intensity related parameters and listeners' linguistic expectancies.

The automatic prominence detection system used to investigate the relationships between acoustic parameters and perceived prominence in German is based on a global model of these phenomena proposed in the works of Kohler [2, 3].

In his view there are two main 'actors', at linguistic-prosodic level, playing a relevant role in supporting sentence prominence: the first, *pitch accent*, coincides with the concept first introduced by Bolinger [5] and it concerns specific movements in F0 profile, while the second, *force accent*, is completely independent from intonational profiles and it is connected with different acoustic phenomena, such as intensity, segmental durations and possibly others. Both phenomena seem to play a relevant role in supporting prominence perception at utterance level, without establishing specific hierarchical roles, but reinforcing their contribution to each other.

The relationship between acoustic parameters expressing *force accents* and those expressing *pitch accents* is complex, e.g. a pitch accented syllable also tends to be longer than an unaccented one [6] and is produced with increased intensity as well [7]. In [8, 9], an algorithm was introduced that disentangles the relative impact of the two major types of influence on perceptual prominence. This approach will now be applied to German data.

1.2. German prominence patterns

A very stable acoustic cue to prominence in German is an increase in duration [10] which can be interpreted as caused by *force accent*. However, *pitch accents* – if present – have a stronger impact on prominence [11]. Pitch accents are caused by the syntactic and semantic structure of an utterance and are not present on every word perceived as prominent. Therefore, force accent related acoustic parameters as duration and intensity might be more reliable indicators of prominence in German. [12] suggest two areas of prominences, an area of low and moderate prominence mainly determined by duration, and an area of high prominence mainly determined by F0. Overall intensity seems to less reliable factor in the signalling of prominence [13]. Another intensity dimension [14] is *spectral emphasis* and refers to the energy increase in the higher frequency parts of the spectrum. It has been related to vocal effort and can be employed by a speaker to express a force accent. [15] found spectral emphasis to be a good indicator of German lexical stress, while [7] remain sceptical of its significance. In [6] prominence was found to correlate slightly with intensity in different frequency bands and formant frequencies.

A main problem in detecting the relevant acoustic parameters influencing perceptual prominence is that listeners are much guided by linguistic expectancies [16]. In German, listeners tend to perceive a syllable as prominent when they *expect* it to be – rather than relying on acoustic cues. However, in normal, unreduced speech, linguistic expectancies and acoustic cues to prominence tend to be in harmony [17]. Another major problem for automatic prominence detection is that acoustic and perceptual phenomena are not always perfectly aligned. Late pitch accents indicate a strong prominence [18, 6] but they often reach their peak *after* the prominent syllable [19].

1.3. Automatic prominence detection on German

There exist approaches to automatic prominence detection in German by [11, 20], both based on Classification and Regression Trees (CART). In both papers, prominence was regarded as having a continuous nature. Both approaches show the influence of linguistic – rather than acoustic – cues to prominence detection.

[11] reached a correlation between observed and predicted prominences similar to best inter-listener agreement (>0.86). In their classification, they used linguistic and acoustic features of the speech signal. They claim that they still reach a high agreement between predicted and perceived prominence relying on acoustic data only (>0.77). Their classification tree revealed that the most important feature was the "presence of a pitch peak" which they regarded as an acoustic feature. Since this feature has been annotated manually it cannot be straightforwardly integrated into an automatic classification. Also, the manual annotation of relevant pitch peaks is not possible without linguistic interpretation, therefore its status as an acoustic feature is at least dubitable. In the CART-based prominence prediction described in [20] it was tested how prominence prediction works with a minimal set of linguistic and acoustic features on a synthesis database that had been segmented and labelled for pitch accents automatically. Again, the "presence of a pitch accent" turned out to be the most important influential factor and correlations were high. But since this approach also relied on linguistic information, it is still unclear how German prominence can be determined relying on acoustic input only.

2. The database

The database used for prominence detection is the *Bonner Prosodische Datenbank* (henceforth: BPD) [6]. It is identical to the one used for training and classification in [11] and was used as a training set in [20]. It consists of sentences and short stories read by 3 native speakers of German. In our investigation, 100 phonetically balanced sentences from each speaker were examined. The data has been manually annotated for syllable and boundary prominence by three trained phoneticians based on the procedure described in [21], who operationalised prominence as a continuous rather than a categorical parameter. I.e., prominence was annotated on a continuous scale ranging from 0-31. The inter-labeller agreements were high and their correlations ranged between 0.74 and 0.86. After labelling, the median prominences were calculated out of the three labellers' prominence ratings for each syllable. The medians are used as reference values of perceptual prominence in our subsequent experiments.

3. Automatic Prominence Detection

As outlined in section 1.1 there are a number of acoustic parameters that support prominence perception. Table 1 depicts the parameters considered in this study as well as a brief reference on the actual computation of them.

Starting from these acoustic parameters and following the relationships outlined before we can introduce a prominence function able to assign a continuous prominence level to each syllabic nucleus using only acoustic information:

$$\text{Prom}^i = W_{FA} \cdot \left[SpEmph^i_{SPLH-SPL} \cdot dur^i \right] + W_{PA} \cdot \left[en^i_{ov} \cdot \left(A^i_{event}(at_M, at_m) \cdot D^i_{event}(at_M, at_m) \right) \right]$$

where $SpEmph_{SPLH-SPL}$ is the spectral emphasis, *dur* is the nucleus duration, en_{ov} is the overall energy in the nucleus and A_{event} and D_{event} are the parameters derived from the TILT model as a function of the maxima alignment type – at_M – and the minima alignment type – at_m (see figure 1). All parameters are referred to the generic syllable nucleus *i*.

Acoustic Parameter	Description
Nucleus Duration (*dur*)	Time duration of the syllable nucleus normalised by considering the mean and variance duration of the syllable nuclei in the utterance (z-score), computed using the manual segmentation available in the database.
Spectral emphasis ($SpEmph_{SPLH-SPL}$)	Normalised SPLH-SPL parameter [22] (z-score).
Pitch movements	TILT model [23] representation of pitch movements derived from a pitch contour computed using the *ESPS get_f0* program [24].
Overall intensity (en_{ov})	RMS energy computed in the frequency band 50-5000 Hz normalised to the mean and variance of intensity inside the utterance (z-score).

Table 1: Acoustic parameters used by the prominence identification algorithm.
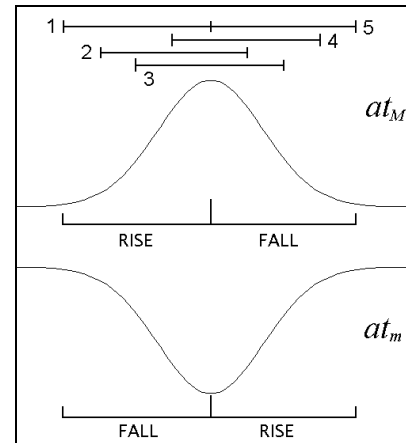

Figure 1: Alignment type parameters between pitch accents and syllable nuclei.

The body of the function *Prom* contains nine parameters, five of them can be considered as supporting the prominence phenomenon from a cross-linguistic point of view ($SpEmph_{SPLH-SPL}$, *dur*, en_{ov}, A_{event} and D_{event}), while the other four, represented in the vector **W** = (W_{FA}, W_{PA}, at_M, at_m), can be seen as language specific. In our model, W_{FA} and W_{PA} weight the contribution of the two different accent types, while at_M and at_m model the different pitch accent alignments specific for each language. For example, if $at_M = 1$ and $at_m = 3$ the rise section of the maxima and the center of the minima in the F0 profile will be taken as reference points to assign the pitch accent to the corresponding syllable nucleus.

Figure 2 shows a computed prominence profile compared with the manual annotation for an utterance taken from the BPD database.
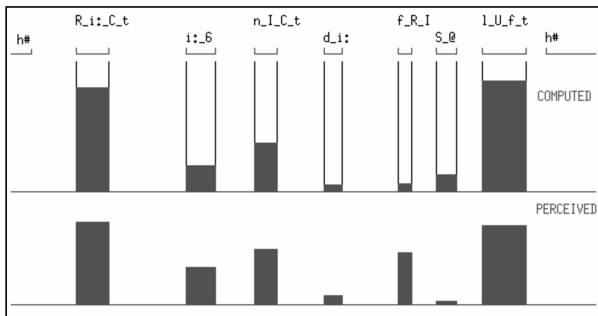
Figure 2: Computed and perceived prominence profiles for the utterance "Riecht ihr nicht die frische Luft? (Don't you smell the fresh air?)". The intervals represent the syllable nuclei in the utterance.

It is relevant to underline that all parameters involved in the *Prom*-function computation are normalised inside the utterance, thus the contributions of different numeric ranges were factored out.

4. Experiments

A parametric scanning in the search space of each **W** component allowed us to determine the optimal combination that gives the maximum agreement with manually annotated data for prominence identification using a specific performance measure.

The performances of the automatic detection system are measured by comparing the continuous prominence level identified by the system with the manual annotation through a local normalization process and the Spearman Rank Correlation Coefficient (SRCC).

Listeners certainly interpret in a different way acoustic/prosodic phenomena belonging to various parts of the utterance and keep the syntagmatic prominence evaluation on a local basis. In order to compare continuous prominence values in a meaningful way, we have to consider that listeners perceive prominence levels in relation with the neighbouring syllables. For this reason every prominence value has been normalised considering the maximum prominence in a local domain, defined as the two neighbourhood nuclei and the examined syllable, trying to model the listener's judgments and keep the necessary normalizations on a local basis.

For a quantitative evaluation of prominence profile similarities, it is more reasonable to analyse them as configurations or patterns instead of comparing the corresponding prominence levels using precise quantitative measures such as the Pearson Correlation Coefficient. Also for this reason we preferred to introduce the local normalisation described above and compare the continuous prominence values using the SRCC measure.

Applying the parametric scanning described above to our complete database (composed of 2431 syllabic nuclei), we obtained a combination of the **W** components that leads to the best performances (SRCC = 0.71):

$$\mathbf{W} = (\ W_{FA} = 0.9,\ W_{PA} = 0.4,\ at_M = 1,\ at_m = 4).$$

The relevant information gathered from the parametric scanning regarding the relationships between force accents and pitch accents are captured by the ratio W_{FA} / W_{PA} and not by W_{FA} and W_{PA} absolute values. Figure 3 shows the variation of the performance measure (SRCC) as a function of this ratio (keeping as constants $at_M = 1$ and $at_m = 4$). The curve exhibit a clear maximum for values of the ratio W_{FA} / W_{PA} near to 2.25 (= 0.9/0.4).

Figure 4 shows SRCC variation as a function of different alignment types keeping W_{FA} and W_{PA} as constants ($W_{FA} = 0.9$, $W_{PA} = 0.4$). Looking at these results, it is interesting to note that, in German, maxima in the pitch profile seems to be more relevant than minima to identify prominent syllables: variations in at_m values does not seem to affect the final performances noticeably.
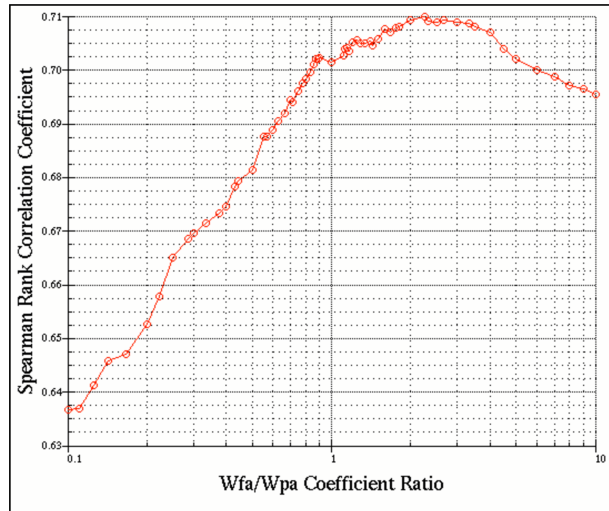

Figure 3: Variation of the performance measure (SRCC) as a function of the ratio W_{FA} / W_{PA} ($at_M = 1$ and $at_m = 4$) expressed on a logarithmic scale.
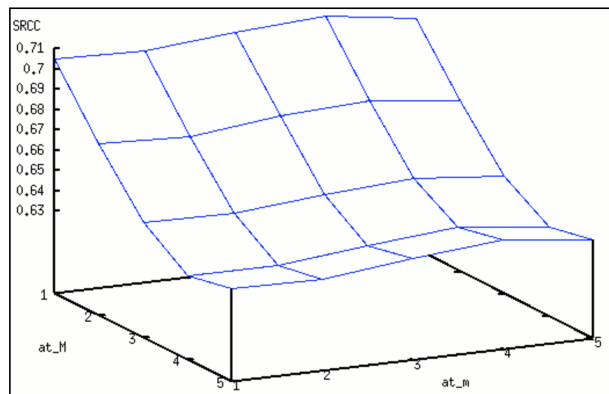

Figure 4: Variation of the performance measure (SRCC) as a function of at_M and at_m / ($W_{FA} = 0.9$ and $W_{PA} = 0.4$).

5. Discussion and Conclusions

The prominence detector reaches good correlations between perceived and predicted prominence patterns relying purely on acoustic input without any manual prosodic annotation being necessary. In order to detect possibilities for further improvements of the algorithm, an in-depth comparison of perceived and computed prominence patterns was performed. This revealed two main sources of mismatch between computed and perceived patterns:

- Syllables with late pitch accents reaching their maximum *after* the prominent syllables tend to be computed less prominent than perceived.
- Listeners prefer to perceive an alternating stress pattern that is not mirrored in the acoustic data. This indicates a systematic top-down

mechanism revealing a rhythmical bias. This phenomenon is illustrated in Fig. 2: in the antepenultimate syllable of the utterance listeners perceive an increase in prominence which cannot be detected acoustically.

Despite these sources of error, our approach proved to be successful in the majority of cases. Therefore, it can be used as an indicator of weighting the relative importance of *force* vs. *pitch accents* in German. Our results thus strengthen the view that force accent related parameters are more reliable cues to prominence in German than pitch accent related parameters. This may come as a surprise since pitch accents previously have been shown to have a major impact on prominence perception in German (cf. 1.2). However, keeping in mind that in their absence, fundamental frequency has only marginal influence on prominence, while force accent parameters seem to keep some of their influence in the presence of a pitch peak, the results are explicable. We take this to be further evidence of the view that in German, a distinction between *force accents* and *pitch accents* can be useful both on the functional and acoustic level. Before drawing further conclusions, it remains to be shown that our optimized model parameters are general enough to be applicable to other (German) data.

Future work will concentrate on two main issues:
- Our normalization procedure builds on the assumption that prominence judgments are performed locally rather than globally. In order to validate this hypothesis further, a perception experiment will be carried out.
- The rhythmical bias revealed in the perceived prominence patterns will be integrated into the detection algorithm. Such a bias might be language specific and stronger in so-called stress timed languages (e.g. German, English) compared to syllable timed languages (e.g. Spanish, Italian) – even if such a language classification still has to be proven convincingly.

6. References

[1] Terken, J., "Fundamental Frequency and perceived prominence parameters", J. Acoust. Soc. Amer., 87, 1991:1768–1776.

[2] Kohler K.J., 2003, "Neglected categories in the modelling of prosody - Pitch timing and non-pitch accents". In: Proceedings of ICPhS'03, Barcelona, 2925-2928.

[3] Kohler K.J., 2005, "Form and Function of Non-Pitch Accents". In: Prosodic Patterns of German Spontaneous Speech, AIPUK, 35a: 97-123.

[4] Nöth, E., Batliner, A., Kiessling, A., Kompe and R., Niemann, H., "VERBMOBIL: the use of prosody in the linguistic components of a language understanding system", IEEE Transactions on Speech and Audio Processing, 8(5), 2000: 519–532.

[5] Bolinger D., 1958, "A theory of pitch-accent in English", Word, 14:109-149.

[6] Heuft, B., "Eine pominenzbasierte Methode zur Prosodieanalyse und -synthese", Peter Lang, Frankfurt: 1999.

[7] Mooshammer, C. and Harrington, J., "Linguistic prominence and loudness: a systematic comparison between lexical word stress, sentence accent and vocal effort", Abstract of Talk presented at BeST, Leiden, 2005.

[8] Tamburini F., 2006, "Reliable Prominence Identification in English Spontaneous Speech". In Proc. Speech Prosody 2006, Dresden, PS1-9-19.

[9] Tamburini F., "Prominenza frasale e tipologia prosodica: un approccio acustico". In Proc. XL congresso internazionale di studi, Società di Linguistica Italiana, Vercelli, settembre 2006, in press.

[10] Jessen, M., Marasek, K., Schneider, K. and Claßen, K. "Acoustic correlates of word stress and the tense/lax opposition in the vowel system of German", Proc. of ICPhS 13 Stockholm, Vol. 4: 428–431.

[11] Portele, T., Heuft, B., Widera, C., Wagner, P. and Wolters, M. "Perceptual Prominence" in Sendlmeier, W. "Speech and Signals", Frankfurt a.M.: Hektor, 2000: 97–115.

[12] Mixdorff, H., Widera, C., "Perceived prominence in terms of a linguistically motivated quantitative intonation model", Proc. of EUROSPEECH, Aalborg, Denmark, 2001:403–406.

[13] Nöth, E., Batliner, A., Kuhn, T., Stallwitz, G., "Intensity as a predictor of focal accent", Proc. of ICPhS, Aix en Provence, France, 1991: 230–233.

[14] Sluijter, A. , van Heuven, V. "Spectral balance as an acoustic correlate of linguistic stress", 100, 1996: 2471–2485.

[15] Claßen, K., Dogil, G., Jessen, M., Marasek, K. , Wokurek , "Stimmqualität und Wortbetonung im Deutschen". Linguistische Berichte, 174, 1998: 202-245.

[16] Eriksson, A., Grabe, E., Traunmüller, H., "Perception of Syllable Prominence by Listeners with and without Competence in the Tested Language" , Proc. Speech Prosody, Aix en Provence, France, 2002, http://aune.lpl.univ-aix.fr/sp2002/pdf/eriksson-grabe-traunmuller.pdf

[17] Wagner, P., "Great Expectations – Introspective vs. Perceptual Prominence Ratings and their Acoustic Correlates", Proc. of INTERSPEECH, Lisbon, Portugal, 2005: 2381–2384.

[18] Kohler, K., "Categorical pitch perception", Proc. of ICPhS 11, Tallinn, Estonia, Vol. 5: 331-333.

[19] Niebuhr, O., Ambrazaitis, G., "Alignment of Medial and Late Peaks in German Spontaneous Speech", Proc. of Speech Prosody, Dresden, Germany, 2006.

[20] Wagner, P., Breuer, S., Stöber, K., "Automatische Prominenzetikettierung einer Datenbank für die korpusbasierte Sprachsynthese", Fortschritte der Akustik, DAGA, Oldenburg, 2000.

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

[22] Fant G., Kruckenberg A., Liljencrants, J., "Acoustic-phonetic Analysis of Prominence in Swedish". In: Botinis, A. (Ed.), Intonation, Kluwer Academic Publisher, 2000: 55–86.

[23] Taylor, P.A., Analysis and Synthesis of Intonation using the Tilt Model, J. Acoust. Soc. Amer., 107(3), 2000: 1697–1714.

[24] Talkin, D., "A robust algorithm for pitch tracking (RAPT)", In: W.B. Kleijn & K.K. Paliwal (Eds.), Speech coding and synthesis, New York: Elsevier, 1995: 495–518.