Automatic Prominence Annotation of a German Speech Synthesis Corpus: Towards Prominence-Based Prosody Generation for Unit Selection Synthesis

Andreas Windmann[1], *Petra Wagner*[1], *Fabio Tamburini*[2], *Denis Arnold*[3], *Catharine Oertel*[1]

[1] Faculty of Linguistics and Literature, University of Bielefeld, Germany
[2] Department of Linguistics and Oriental Studies, University of Bologna, Italy
[3] Institute of Communication Sciences, University of Bonn, Germany

andreas.windmann@uni-bielefeld.de, petra.wagner@uni-bielefeld.de, fabio.tamburini@unibo.it, dar@ifk.uni-bonn.de, c.oertel@uni-bielefeld.de

Abstract

This paper describes work directed towards the development of a syllable-prominence-based prosody generation functionality for the German BOSS unit selection speech synthesis system. A general concept for syllable-prominence based prosody generation in unit selection synthesis is proposed. As a first step towards its implementation, an automated syllable prominence annotation procedure based on acoustic analyses has been performed on the BOSS speech corpus. The prominence labeling has been evaluated against an existing annotation of lexical stress levels and manual prominence labeling on a subset of the corpus. We discuss methods and results and give an outlook on further implementation steps.

1. Introduction

State-of-the-art unit selection speech synthesis systems that operate on large speech corpora allow for the production of a highly natural speech output under ideal circumstances. An important factor that constrains the design of a prosody generation functionality for this technique is the general philosophy of the corpus-based synthesis approach: Unit selection speech synthesis draws its potential for naturalness from exploiting the inherent variation of the data in the speech corpus. Interference with the data should be kept at minimum [1]. Prosody generation should be faithful to this principle. A good way to ensure this is to model prosodic structures by selecting units that best fit a given prosodic specification, rather than by employing signal manipulation techniques that alter the speech signal.

As for the realization of this concept, there are two approaches that one could think of: One possibility is to model the individual prosodic parameters in a very direct way, for example by using algorithms that predict specific F0 targets and segmental durations for the speech output, and to select units that best fit these specifications. Another approach would be to treat prosody as a "black box". In this approach, prosody is represented in terms of abstract perceptual prominence levels of linguistic units, such as syllables, rather than specified acoustic-prosodic parameters. For this purpose, prominence can be defined as the degree to which the syllable is perceived as standing out relative to its environment [2]. The realization of a certain prominence pattern could be ensured by selecting units from the speech corpus that match the predicted prominence levels of the corresponding units in the desired speech output as closely as possible. It could be argued from a theoretical perspective that in the context of unit selection synthesis, representing prosody in terms of abstract prominence levels might have the advantage of better exploiting the inherent variation within the speech data. For example, a specific predicted pitch target for a certain position within an utterance to be synthesized might not be available in the speech corpus. In a prominence-based framework, this could be made up for by selecting a unit that exhibits a high value for another prosodic parameter at this position, resulting in the same abstract level of perceptual salience that the pitch target would have assigned. The prominence-based approach could thus be an efficient way of prosody modeling, although it has to be acknowledged that there certainly are contexts in which a specific pitch profile is crucial.

A possible architecture for syllable-prominence-based prosody prediction in unit selection synthesis could be thought of in terms of three components: First, the text-to-speech component of the synthesis system needs to have implemented a syllable prominence model that predicts prominence levels for each syllable in the desired target utterance to be synthesized. This could be realized largely based on the pronunciation lexicon of the system, in which prominence patterns for all the entries would have to be specified. Extra rules would be needed to account for effects of prosodic rules or focal structures on sentence level and maybe to model phenomena such as stress shifts [3]. Second, each syllable in the speech corpus of the system needs to be assigned a prominence value based on its relative perceptual salience. Third, these prominence values will have to be included into the system either as target or as transition costs, so that they can be considered in unit selection. This paper is concerned with the second step, the assignment of prominence values on syllable level to a speech corpus for unit selection synthesis. The rest of the paper is organized as follows: In section 2, we discuss previous research and motivate our method. The Bonn Open Synthesis System (BOSS), in which our prosody prediction functionality is to be implemented, is introduced in section 3. In the fourth section, the prominence model and the prominence detection algorithm are introduced and special challenges for its application in a synthesis system are discussed. In section 5, we present the results of the evaluation procedures we applied to the automatic prominence labeling. In section 6, conclusions are drawn and perspectives for further work are addressed.

2. Previous Work

An early example of work on the explicit modeling of prominence in speech synthesis is described in [4]. However, the representation of prosody in terms of abstract prominence labels is restricted to the analysis of the text input in this system. Rules are then applied which translate the prominence

values into individual prosodic parameters. More recently, it has been demonstrated that the modeling of prominence improves the quality of synthetic speech in unit selection synthesis [5]. In their implementation, every word in the synthesis corpus is assigned a prominence value, which basically expresses the probability of it carrying a pitch accent. These probabilities have been obtained from other corpora, which have been manually annotated for pitch accents.

Since unit selection corpora tend to be too large to be manually annotated, automatic methods have to be applied. There are various approaches towards the automatic assignment of prominence labels to speech data. In a number of studies, prominence has been identified in speech material based on linguistic information such as word classes. For German, [6] and [7] have demonstrated the integration of acoustic as well as linguistic information in automatic prominence classification. While good results have been obtained by automatic prominence annotation based on linguistic criteria, it entails two problems: First, it requires that the speech corpus be enriched with explicit linguistic meta-data and second, it is questionable whether prominence labeling based on linguistic criteria will always correspond to individual production in corpus recordings, e. g. due to influences of prosodic focus or contextual deaccentuation.

Another possibility is to train machine learning algorithms on manual annotations of a subset of the corpus to be labeled. There are some recent studies that have obtained promising results applying this methodology [8, 9]. Classification mostly works on segmental and acoustic information that is either inherently required in a speech synthesis context and is therefore available anyway, or can be automatically added with relatively little effort. Prominence has been classified in these studies in terms of small numbers of discrete categories.

A third approach towards automatic prominence annotation is to rely on acoustic analyses of the speech data. This strategy is based on a large body of experimental evidence that suggests a relationship between perceptual prominence ratings and a number of acoustic-phonetic parameters [10, 11]. Currently, there are various implementations of acoustically-based prominence identification algorithms for a number of languages [12, 13, 14, 15]. All these systems include components for automatically determining the positions of syllables or syllable nuclei in running speech, so there should theoretically be no need to supply additional information besides the speech data itself. For our present analysis, we employed an algorithm that is based on work by Tamburini [12, 13, 16, 17], more specifically an adaption for German data, which is described in [12]. It detects syllable nuclei in running speech and performs analyses of a number of acoustic cues. Details on the algorithm are given in section 4.

One problem with this approach is that listeners do not exclusively rely on acoustic cues in interpreting syllable prominence. There is evidence that linguistic expectancies of listeners play a substantial role in assigning prominence values to syllables [11, 18, 19]. Specifically, [20] have observed that linguistic cues introduce systematic deviations from acoustic prominence patterns to listeners' ratings in French. [12] themselves report that human annotators who have been employed for evaluating the automatic prominence tagging show a "rhythmical bias" in rating syllable prominence. Moreover, recent studies have found that priming effects may influence the perception of syllable prominence, and that the overall pattern of prominence values in an utterance may have an impact on the relative importance of the individual acoustic parameters [21, 22]. A prominence tagging algorithm that takes these factors into account would be a desirable achievement. For the moment, we have to rely on the purely acoustic detection method, which, despite the problems mentioned, has been shown to reach good correlations with annotations by human labelers.

3. Synthesis system and corpus

The Bonn Open Synthesis System (BOSS) is a non-uniform unit selection speech synthesis system based on the Verbmobil synthesis concept [23]. A simplified representation of the system architecture is shown in Figure 1. As can be seen, the system is separated into three components. The first one, the client, receives the user input and performs text preprocessing. In the current configuration, BOSS employs a generic TTS client which is described in [24]. It performs text normalization on the user input and creates an XML representation of the normalized user input which is sent to the server. The server module performs linguistic analysis of the user input, as well as unit selection and the actual synthesis.

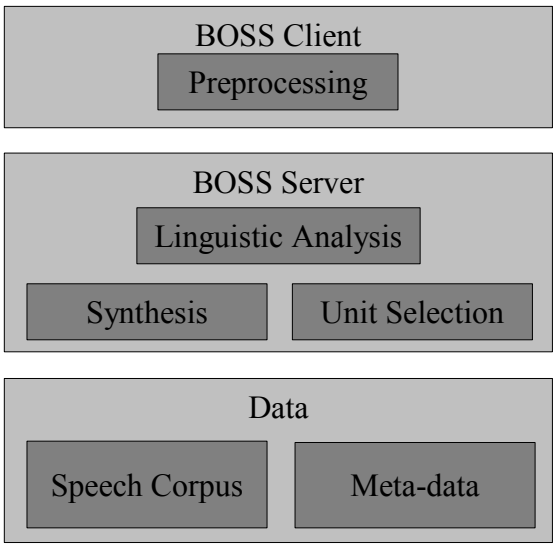

Figure 1: *Schematic representation of BOSS system architecture.*

BOSS currently employs the Verbmobil speech corpus, which is described in [25]. It consists of 4545 sentences and approximately 10,000 words. The sentences are taken from transcriptions of actual planning dialogs in the domain of meeting scheduling and hotel bookings. They were read by a professional female speaker. The related meta-data is provided in an SQL database, comprising relevant segmental and acoustic information about the speech data on sentence, word, syllable, phone and half phone level. In the unit selection process, the information stored in this database is retrieved by the unit and the transition cost function of the system, in order to select the best-fitting candidate unit in a given position. The prominence labels obtained in this study will have to be fed into the corpus meta-data in order to make them accessible for unit selection.

There is currently no specific computation of pitch targets in BOSS, but a duration prediction module based on classification and regression trees has been added to the linguistic analysis component [26]. Moreover, the system features some more general functionalities for prosodic control. The unit selection algorithm takes into account the position of target units relative to phrase boundaries in the

utterances they are part of, so that ideally, a rough match of the desired prosodic context is provided for. This information is also utilized by the duration prediction functionality. There is a further prosodic control in terms of lexical stress levels: The phonetic transcription module within the server utilizes the BOMP pronunciation dictionary [27] in order to generate phonetic transcriptions of the orthographic input. Within the dictionary entries, syllable boundaries are specified and it is indicated for each syllable whether it bears primary, secondary or no lexical stress. The speech corpus has been annotated for lexical stress on syllable level using the same three-way distinction. This annotation is taken into account by the unit selection algorithm, so that the stress pattern of the output will ideally match the pattern of the input [27]. However, the annotation of the speech corpus for lexical stress is based on the standard pronunciation of the individual words. Thus, it reflects how strongly each syllable in the corpus *should be* stressed rather than their actual realization in terms of prominence. It is hoped that the application of acoustically-motivated prominence labels will allow for a more fine-grained prosody modeling and better correspond to actual perception.

4. Automatic prominence annotation

Previous studies on the relationship between perceptual prominence and acoustic parameters have identified four central factors that contribute to the perception of a syllable as being prominent: Duration, overall intensity, pitch movements and spectral emphasis, the latter being related to the spread of energy over high frequency bands [2, 10, 12, 17]. These cues are utilized by the prominence identification algorithm we applied. Prominence values on a continuous scale, in most cases ranging from 0 to 1, are assigned to each syllable in the speech data. This is an important point, since prominence is often understood in discrete terms, distinguishing between prominent and non-prominent syllables only. The prominence values in the algorithm we applied are based on the following computation of the acoustic parameters [12]:

$$\text{Prom}^i = W_{FA} * [SpEmph^i_{SPLH\text{-}SPL} * dur^i] + W_{PA} * [en^i_{ov} * (A^i_{event}(a_{tM}, a_{tm}) * D^i_{event}(a_{tM}, a_{tm}))]$$

Detailed descriptions on the computation of the individual parameters are given in [12]. $SpEmph^i_{SPLH\text{-}SPL}$ represents the spectral emphasis parameter [28] and en^i_{ov} is the overall intensity during the generic syllable nucleus *i*. The dur^i parameter represents the duration of that nucleus. A^i_{event} and D^i_{event} are parameters derived from a TILT model representation of the F0 movements within the syllable nucleus *i* [29]. This way of computing syllable prominence reflects the finding that most of the relevant acoustic information for perceiving a syllable as prominent is concentrated in the nucleus [13]. The further components W_{FA}, W_{PA}, a_{tM} and a_{tm} are language-specific weighting factors, which have been introduced by [12]. W_{FA} and W_{PA} are meant to weight the individual contributions of *force accents*, related to spectral emphasis and duration versus *pitch accents*, related to F0 movements and overall intensity in a given language [12], whereas a_{tM} and a_{tm} represent different types of temporal alignment of pitch maxima and minima to syllable nuclei.

As for the settings of these parameters, we applied the findings reported in [12]. W_{FA} and W_{PA} were set to 0.9 and 0.4 respectively, paying attention to the finding that force accents play a more important role in marking a syllable as prominent than do pitch accents in German. The configuration of a_{tM} and a_{tm} was chosen so as to align pitch maxima with the syllable nucleus that exhibits the greatest overlap with their rise section, while pitch minima are aligned with the syllable nucleus that exhibits the greatest overlap with a period of time starting shortly before the end of the fall section and covering approximately 75% of the rise section of the minimum, which [12] report to be the optimal configuration for temporal alignment of pitch events to syllable nuclei in German.

The automatic syllable nucleus detection that is implemented in the tagging software is very delicate and error-prone and will introduce a severe element of uncertainty to tagging results. However, if reliable information on the temporal positions of syllable nuclei within the speech data under consideration can be provided, the software is able to utilize it instead of relying on the automatic nucleus detection. We were able to obtain this information from the phonetic transcription of the BOSS speech corpus that is provided within its meta-data. A list of all possible syllable nuclei was prepared and compared to the phonetic transcription of the corpus. Whenever a phone label from the transcription matched an entry in the list, the information on its temporal position within the utterance it is part of was extracted. The labeling conventions that have been applied in the creation of the BOSS corpus make this procedure very straightforward, as combinations of phones that constitute syllable nuclei in German, such as a vowel followed by an /ʁ/ realized as a vocalic [ɐ] or a combination of syllabic consonants in a reduced syllable, are represented as single phone units in the phonetic transcription. Thus, reliable segmentation into syllable nuclei could be obtained for every utterance within the BOSS speech corpus.

In the computation of the syllable prominence values, the automatic algorithm normalizes the values of the individual parameters for every syllable nucleus *i* to the mean and variance of the utterance that *i* is part of. As listeners can be expected to evaluate the prominence value of a given syllable against its neighboring syllables [12], this procedure is well-motivated with regard to perceptual adequacy. However, it is not clear whether it is ideal in a speech synthesis context: If, for example, a given syllable has been produced with a relatively high intensity compared to the mean and variance of the utterance it is part of, it will be assigned a high value for this parameter. If, on the other hand, the overall intensity level of this utterance is relatively low compared to the other utterances in the corpus, the computed prominence value for the syllable under consideration may be too high compared to syllables from other utterances. It might be speculated that this could lead to deviations from the intended prominence pattern of the speech output if elements from different utterances are combined in the synthesis process.

One possible answer to this objection is that with carefully controlled speech data such as a synthesis corpus, variation in terms of the acoustic parameters on utterance level should not be too dramatic. In order to substantiate this assumption, we performed a correlation analysis on prominence values and non-normalized durations of all syllable nuclei in the corpus. The rationale behind this procedure is that if normalization on utterance level is unproblematic, a stable relationship between prominence labels and the individual acoustic parameters should be observable not only within the individual utterances, but also over the whole speech data. We found a correlation of 0.49 (Pearson's r; $p<0.05$) for prominence and nucleus duration over the whole data range. This result lies within the range of what could be expected, since prominence is, of

course, also affected by the other acoustic parameters and cannot be predicted based on durations alone. The fact that a substantial correlation exists indicates that the relationship between prominence and acoustic parameters throughout the whole corpus is not affected to a problematic extent by normalization on utterance level.

5. Evaluation

Assuming that the above-mentioned annotation of the BOSS corpus for lexical stress levels at least roughly corresponds to perceptual prominence levels, it was used for a preliminary performance assessment of the automatic prominence labeling. Mean prominence values for the syllables in the three lexical stress categories, "Primary", "Secondary" and "None", were computed. They are shown in Figure 2. Comparing the mean prominence values of the syllables with primary, secondary and no lexical stress in a one-way ANOVA, we found a significant interaction between lexical stress category and prominence value ($F = 4534.34$, $p < 0.0005$). This preliminary result indicates that the prominence labeling algorithm has performed well on the BOSS corpus. It also provides further evidence for the assumption that the prominence normalization on utterance level does not alter the relationship between prominence and acoustic parameters throughout the whole corpus to a problematic extent .

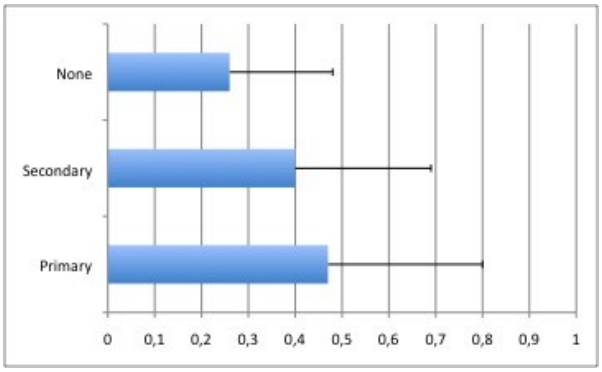

Figure 2: *Mean prominence values of syllables from the BOSS corpus with primary, secondary and no lexical stress.*

Evaluation against manually-tagged speech data is a well-established procedure for testing automatic prominence tagging algorithms [12, 14]. We recruited seven subjects for prominence labeling on a subset of the BOSS corpus, consisting of 15 sentences. All subjects were native speakers of German and trained in phonetics. A reimplementation of the graphical interface described in [22] was used for the labeling task. Orthographic representations of the sentences to be labeled were shown to the subjects on a computer screen, with a slider above each syllable. Subjects listened to the sentences over computer speakers or headphones and were asked to indicate the perceived prominence of each syllable by the position of the respective slider. Before the actual labeling procedure, subjects went through a short training phase, in which an experimenter explained the procedure based on example sentences, comprising instances of very prominent and non-prominent syllables. Subjects were instructed to move sliders up to the top of the rating scale if they perceived a syllable as maximally prominent, and to the bottom if they perceived a syllable as minimally prominent. They were allowed to listen to each sentence as often as they liked.

There has been considerable discussion in the literature with regard to the optimal rating scale for prominence labeling tasks. Scales with fine-grained discrete gradations, such as values from 0 to 30, have been used in previous works [10], but it has been argued that this kind of scale might be hard to handle for raters [30]. In other studies, continuous rating scales without any gradation have been applied [31]. Since the prominence tagger produces a continuous output, we decided to follow this approach and did not give subjects any indication of a rating scale in addition to the sliders. System-internally, prominence levels were encoded in terms of values between 0 and 100. Figure 3 shows a picture of the graphical user interface.

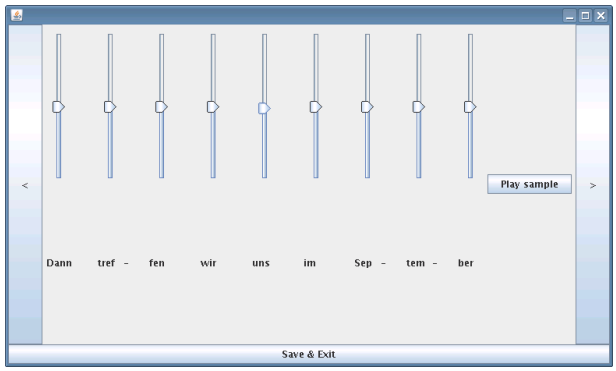

Figure 3: *Graphical user interface used for manual prominence labeling. Each syllable is rated with one slider. The "Play Sample" button is use to play the audio file. The arrow button on the right is used to proceed to the next sentence.*

Agreement between the individual annotators was computed by taking the average over pairwise correlation coefficients (Spearman ρ). Results are shown in Table 1. The mean correlation among annotators was 0.61 on the test set.

Table 1: *Pairwise correlation coefficients (Spearman ρ) between human annotators (S1-7) averaged over the 15 test sentences.*

	S1	S2	S3	S4	S5	S6	S7	**Ø all**
S2	0.70							
S3	0.58	0.58						
S4	0.79	0.72	0.55					
S5	0.56	0.49	0.38	0.59				
S6	0.79	0.72	0.59	0.74	0.58			
S7	0.58	0.55	0.44	0.64	0.60	0.62		
Ø	0.67	0.62	0.52	0.67	0.53	0.67	0.57	**0.61**

In order to compare computed and perceived prominence, we calculated the median of the perceived prominence ratings for each syllable in the test set. Thus, a perceived prominence profile for each sentence in the test set was obtained. We found a median correlation of 0.62 between perceived and computed prominence profiles on the test set (Spearman ρ; $p<0.05$). This result suggests that the automatic prominence labeling on the whole comes close to human inter-rater agreement. Yet, it also confirms some of the problems that have been highlighted in previous studies. For example, the

rhythmical bias reported in [12] was also found in our perceived prominence profiles, causing notable deviations from the computed prominence profiles in a number of cases. A closer inspection of the results revealed that, whereas in most sentences, perceived and computed prominence profiles match quite well, correlations were only marginal in some other sentences. Our work thus delivers further evidence for the finding that prominence perception is not always perfectly in line with acoustic parameters. More research is necessary to identify the sources of mismatch between acoustics and perception and find ways by which they could be taken into account by automatic prominence detection algorithms. An example of a computed and a perceived prominence profile of an utterance from the BOSS corpus is shown in Figure 4.

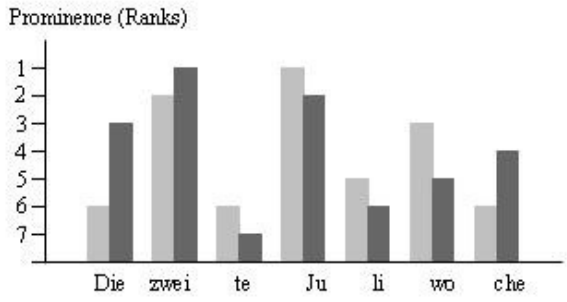

Figure 4: *Perceived (light) and computed (dark) prominence profiles for the utterance "Die zweite Juliwoche (the second week of July)".*

6. Discussion and Conclusions

In this paper, we have presented a first step towards the development of a syllable-prominence based prosody generation functionality for unit selection speech synthesis. The prominence labeling algorithm we applied has proven reasonably successful in predicting perceived prominence patterns on data from the BOSS speech corpus. A comparison with an existing annotation of lexical stress levels has also yielded encouraging results. The automatically-assigned prominence labels have been obtained on the basis of acoustic analyses alone. No additional information was exploited with the exception of the syllable nucleus segmentation, which could be easily obtained from the existing corpus annotation on phone level. This is a particularly interesting perspective for speech synthesis corpora, as information necessary for prominence identification algorithms that rely on linguistic criteria might not be available with speech synthesis resources. It has been hypothesized that the normalization procedure on utterance level that is performed by the tagging software might lead to problems in synthesis. Although in the light of our preliminary results it seems unlikely that severe problems will be caused, this will have to be tested in perception experiments once the implementation is completed.

Further work includes the implementation of a prominence prediction functionality within the linguistic analysis component of the system and the inclusion of the prominence labels as a cost factor into the unit selection algorithm. Possible interactions with existing functionalities, such as the duration prediction, will have to be taken into account. As has been pointed out, the implementation of the prediction functionality could be realized largely within the pronunciation dictionary of the synthesis system. As for the modeling of syllable prominence as a cost factor, it is an interesting question whether to model it in terms of target or transition costs. Syllable prominence values could be modeled as target costs, in requiring a candidate unit to match a certain prominence target defined be the prominence prediction algorithm. Alternatively, prominence could be modeled in terms of transition costs, imposing that a unit to be selected exhibits a certain prominence value in relation to the neighboring units in the synthesized speech output. In either case, an important consideration would be how much relative weight the prominence value of a unit should be assigned in the computation of costs. As [5] state, "control of prosody comes at the potential cost of lower segmental quality". It is planned to conduct a series of perception experiments once the prosody generation component is implemented in BOSS in order to determine whether speech output with an improved prosody will be preferred by listeners even if it goes along with a reduction in segmental quality.

7. References

[1] Campbell, N., 1996, CHATR: A High-Definition Speech-Re-Sequencing System. In *Acoustical Society of America and Acoustical Society of Japan, Third Joint Meeting.*

[2] Terken, J., 1991, Fundamental frequency and perceived prominence. In *J. Acoust. Soc. Amer*. 89(4): 1768-1776.

[3] Wagner, P. & Fischenbeck, E., 2002, Stress perception and production in German Stress Clash Environments. In *Proc. Speech Prosody 2002*, Aix-en-Provence, France, 687-690.

[4] Portele, T. & Heuft, B., 1997, Towards a prominence-based synthesis system. In *Speech Communication 21(1997)*, 61-72.

[5] Strom, V., Nenkova, A., Clark, R., Vazquez-Alvarez, Y., Brenier, J., King, S. & Jurafsky, D., 2007, Modelling Prominence and Emphasis Improves Unit-Selection Synthesis. In *Proc. Interspeech 2007*, Antwerp, Belgium, 1282-1285.

[6] Portele, T., Heuft, B., Widera, C., Wagner, P. & Wolters, M., 2000, Perceptual Prominence. In Sendlmeier, W. (Ed.), Speech and Signals, Frankfurt a.M., Hektor, 97-115.

[7] Wagner, P., Breuer, S. & Stöber, K., 2000, Automatische Prominenzetikettierung einer Datenbank für die korpusbasierte Sprachsynthese In *Fortschritte der Akustik, DAGA 2000*, Oldenburg, Germany.

[8] Avanzi, M., Lacheret-Dujour, A. & Victorri, B., 2010, A Corpus-based Learning Method for Prominence Detection in Spontaneous Speech. In *Proc. Speech Prosody 2010*, Chicago, Illinois, W1.04.

[9] Al Moubayed, S., Anantakrishnan, G. & Enflo, L., 2010, Automatic Prominence Classification in Swedish. In *Proc. Speech Prosody 2010*, Chicago, Illinois, W1.10.

[10] Fry, D.B., 1958, Experiments on the Perception of Stress. In *Language and Speech 1*, 126-152.

[11] Fant, G. & Kruckenberg, A., 1989, Preliminaries to the study of Swedish prose reading and reading style. In *STR-QPSR 2/1989*, KTH Stockholm, 1-83.

[12] Tamburini, F. & Wagner, P., 2007, On Automatic Prominence Detection for German. In *Proc. Interspeech 2007*, Antwerp, Belgium, 1809-1812.

[13] Tamburini F. & Caini C., 2005., An automatic system for detecting prosodic prominence in American English continuous speech. In International Journal of Speech Technology 8, 33-44.

[14] Wang, D. & Narayanan, S., 2007, An Acoustic Measure for Word Prominence in Spontaneous Speech. In *IEEE Transactions on Audio, Speech and Language Processing 15(2)*, February 2007, 690-701.

[15] Martin, P., 2010, Prominence detection without syllabic segmentation. In *Proc. Speech Prosody 2010*, Chicago, Illinois, W1.08.

[16] Tamburini F., 2003, Prosodic prominence detection in speech. In *Proc. 7th International Symposium on Signal Processing and its Applications - ISSPA2003*, Paris, 385-388.

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

[18] Streefkerk, B., 2002, *Prominence – Acoustic and lexical/syntactic correlates*. PhD Thesis, University of Amsterdam.

[19] Wagner, P., 2005, Great Expectations – introspective vs. perceptual prominence ratings and their acoustic correlates. In *Proc. Interspeech 2005*, Lisbon, Portugal, 2381-2384.

[20] Goldman, J.P., Auchlin, A., Roekhaut, S., Simon, A.C. & Avanzi, M., 2010, Prominence perception and accent detection in French. A corpus-based account. In *Proc. Speech Prosody 2010*, Chicago, Illinois, P3a.22.

[21] Arnold, D. & Wagner, P., 2008, The influence of top-down expectations on the perception of syllable prominence. In *Proc. ISCA Workshop on Experimental Linguistics,* Athens, Greece, 25-28.

[22] Arnold, D., Wagner, P. & Möbius, B., 2010, The effect of priming on the correlation between prominence ratings and acoustic features. In *Proc. Speech Prosody 2010*, *Satellite Workshop on Prosodic Prominence: Perceptual and Automatic Identification*, Chicago, Illinois, W1.02.

[23] Klabbers, E., Stöber, K., Veldhuis, R., Wagner, P., & Breuer, S., 2001, Speech synthesis development made easy: The Bonn Open Synthesis System. In *Proc. Eurospeech 2001,* Aalborg, Denmark, 521-525.

[24] Stöber, K., 2002, *Bestimmung und Auswahl von Zeitbereichseinheiten für die konkatenative Sprachsynthese*. PhD Thesis, University of Bonn.

[25] Klabbers, E. & Stöber, K., 2001, Creation of speech corpora for the multilingual Bonn Open Synthesis System. In *Proc. ESCA SSW4*, Pitlochry, Scotland, 23-28.

[26] Breuer, S., 2008, *Multifunktionale und multilinguale Unit-Selection-Sprachsynthese.* Designprinzipien für Architektur und Sprachbausteine. PhD Thesis, University of Bonn.

[27] Portele, T., Krämer, J. & Stock, D., 1995, Symbolverarbeitung im Sprachsynthesesystem Hadifix. In *Proc. 6. Konferenz Elektronische Sprachsignalverarbeitung*, Wolfenbüttel, Germany, 97-104.

[28] Fant, G., Kruckenberg, A. & Liljencrants, J., 2000, Acoustic-phonetic Analysis of Prominence in Swedish. In Botinis, A. (Ed.), *Intonation*, Dordrecht, Kluwer, 55-86.

[29] Taylor, P.A., 2000, Analysis and Synthesis of Intonation using the Tilt Model. In *J. Acoust. Soc. Amer. 107(3)*, 1697-1714.

[30] Jensen, C., & Tondering, J., 2005, Choosing a Scale for Measuring Perceived Prominence. In *Proc. Interspeech 2005*, Lisbon, Portugal, 2385-2388.

[31] Eriksson, A., Grabe, E. & Traunmüller, H., 2002, Perception of syllable prominence by listeners with and without competence in the tested language. In *Proc. Speech Prosody 2002*, Aix-en-Provence, France, 275-278.