Prominence-Based Prosody Prediction for Unit Selection Speech Synthesis

Andreas Windmann[1], Igor Jauk[2], Fabio Tamburini[3], Petra Wagner[1]

[1] Faculty of Linguistics and Literature, Bielefeld University, Germany
[2] Faculty of Technology, Bielefeld University, Germany
[3] Department of Linguistics and Oriental Studies, University of Bologna, Italy

andreas.windmann@uni-bielefeld.de, ijauk@techfak.uni-bielefeld.de,
fabio.tamburini@unibo.it, petra.wagner@uni-bielefeld.de

Abstract

This paper describes the development and evaluation of a prosody prediction module for unit selection speech synthesis that is based on the notion of *perceptual prominence*. We outline the design principles of the module and describe its implementation in the Bonn Open Synthesis System (BOSS). Moreover, we report results of perception experiments that have been conducted in order to evaluate prominence prediction. The paper is concluded by a general discussion of the approach and a sketch of perspectives for further work.

Index Terms: speech synthesis, unit selection, perceptual prominence, prosody modeling, metrical phonology

1. Introduction

Prosody modeling in unit selection speech synthesis systems is usually realized by employing algorithms that predict acoustic-prosodic parameters of the speech output, such as F0 values or segmental durations. In this paper, we explore an alternative strategy: Keeping in mind that a speech synthesis system is essentially a tool designed for human listeners, we propose to focus on the *perceptual* rather than the *acoustic* dimension of prosody. In this view, prosodic structures are represented as patterns of *perceptual prominence*.

We define perceptual prominence as the gradually perceived prosodic markedness of a syllable or a higher-level linguistic unit relative to its environment [1]. The perception of prominence is determined by the individual acoustic-prosodic parameters of the speech signal [2],[3],[4],[5], but to a considerable extent also by linguistic intuitions of listeners [1],[4],[6],[7],[8],[9]. Applying perceptual prominence to prosody generation in unit selection synthesis requires three components: (1) a module for predicting prominence from text, (2) annotation of the system's speech corpus for perceptual prominence and (3) the modeling of prominence as a cost factor.

In a number of studies on prosody modeling, prominence has been employed as an intermediate representation that is predicted from text and then used as the basis for computing individual acoustic-prosodic parameters of the speech output [10],[11],[12]. [13] report on automatic prominence labeling of a synthesis corpus on a four-point scale and the modeling of prominence as a unit cost factor, but do not deal with prominence prediction from text. [14] describe their approach as prominence-based, but employ a definition that is quite different from the above one, basically equating the term with the probability of a word carrying a pitch accent.

This paper addresses the implementation and evaluation of prominence-based prosody prediction in the Bonn Open Synthesis System (BOSS) [15],[16]. The annotation of the BOSS speech corpus for perceptual prominence has been described in detail in a previous paper [17] and will not be discussed at length here. The rest of this paper is structured as follows. In section 2, we outline the phonological rules the prominence prediction module is based on. Its implementation in BOSS is addressed in section 3. In section 4, we discuss perception experiments that have been conducted in order to evaluate our approach. Section 5 sums up the results and addresses perspectives for further work.

2. Prominence Prediction

The prominence prediction module is based on the implementation of a set of metrical-phonological rules for German proposed in [1], drawing on earlier work by [18]. The rules are successively applied to utterances, assigning rhythmical *beats* to the individual syllables if they fulfill certain linguistic criteria as outlined below. Application of the rules to an utterance generates a *metrical grid*, which represents its prominence pattern. The prominence value of each syllable in the context of the utterance it belongs to can be read from the number of beats it has been assigned. The set consists of the following rules.

1. Assign a default beat to every syllable
2. Assign a beat to every syllable whose nucleus is not a reduced vowel or a syllabic consonant
3. According to the part-of-speech of a word, assign beats to its constituent syllable bearing primary stress:
   a) Nouns, proper names, numerals: 5 beats
   b) Adjectives, adverbs: 4 beats
   c) Full verbs, pronouns: 3 beats
   d) Auxiliary verbs, affirmative and negation particles: 2 beats
   e) Other POS: 1 beat
4. Assign an additional beat to the syllable that bears primary stress within the last noun, adjective or adverb in a prosodic phrase and within each verb that does not follow a noun, adjective or an adverb
5. Assign an additional beat to the syllable bearing primary stress within an utterance-initial function word
6. Assign an additional beat to every second syllable in sequences of three or more syllables carrying two beats

Originally, these rules are based on introspective phonological reasoning and incorporate intuitions such as the Nuclear Stress Rule (rule 4) or grid euphony (rule 6). However, the empirical adequacy of the approach has been established by [1], who showed in a large-scale corpus study that the rules predict prominence patterns that closely match human perception. The proposed prediction algorithm thus presents a simple, yet powerful solution for modeling prominence patterns of pragmatically neutral declarative utterances in German. An example of a metrical grid that comprises all of the above rules is shown in Figure 1.

									x4
									x3
x5	**x4**			**x3**					**x3**
x3	**x3**			**x3**					**x3**
x3	**x3**			**x3**					**x3**
x3	**x3**		**x3**	**x3**		**x6**			**x3**
x2	**x2**		**x2**	**x2**	**x2**	**x2**	**x2**		**x2**
x1	**x1**	**x1**	**x1**	**x1**	**x1**	**x1**	**x1**	**x1**	**x1**
ɪç	fliː	gə	am	aɪn	ʊn	tsvan	tsɪçs	tən	maɪ
Ich	flie	ge	am	ein	und	zwan	zigs	ten	Mai
PRON	V		PREP	ADJ					N

Figure 1: *Metrical grid representation of the utterance "Ich fliege am einundzwanzigsten Mai (I am flying on the 21st of May)" as predicted by the proposed set of rules. The numbers indicate the affiliation of beats to rules.*

3. Implementation in BOSS

Figure 2 depicts the system architecture of BOSS, including the prominence prediction module. Prominence prediction is preceded by two processing steps. Initially, the *BOSS Client* performs text preprocessing on the user input and determines the locations of phrase boundaries based on punctuation. An XML representation of the preprocessed text, which is hierarchically structured into sentence and word level elements, the latter including phrase boundary marks, is then passed on to the *BOSS Server*. Here, the transcription module supplies phonetic transcriptions and lexical stress information and expands the hierarchy of elements in the XML document to syllable, phone, and half-phone level. Transcription, syllabification and stress placement are carried out using a pronunciation dictionary. A German morpheme list and decision-tree-based transcription serve as fall-back mechanisms for OOV words [16].

After the transcription process is completed, prominence prediction consecutively applies rules 1 through 6 to the syllable elements in the current XML document. With the exception of part-of-speech (POS) labels, all information required for applying the rules is available at this point. Phrase-final and utterance-initial words (rules 4,5) are identified by the BOSS Client; the transcription module provides syllabification, phonetic transcription (rule 2) and primary stress locations (rules 3,4,5). The HMM-based POS tagger described in [19] is used for obtaining the POS information required for rules 3, 4, and 5. Being called by the prominence prediction module, it receives the orthographic word sequence and determines the POS labels, which are then written into the word elements of the utterance XML document. Once prominence prediction at syllable level is completed, prominence values are propagated to word and phone level elements in the XML document. Word elements receive the prominence value of the most prominent syllable they contain; phone elements are simply assigned the prominence value of the syllable they are part of.

Unit selection in BOSS starts out by applying a preselection algorithm that creates a search space of potential candidate units which match the segmental structure and possibly other features of the desired target utterance. This begins at word level and is successively passed on to syllable, phone and half-phone level if no matching units on the respective higher levels are found. The necessary information about the units in the corpus is contained in an SQL database, organized into separate tables for the individual unit levels. Once a search space of potential candidates has been established, unit selection as such is performed, finding the “cheapest” sequence of units in terms of unit and transition costs. Prominence is considered as a unit cost factor. The SQL database containing the corpus description has been enriched with perceptual prominence labels, using an automatic annotation algorithm based on analyses of acoustic correlates of prominence in the speech corpus [20]. Levels of perceived prominence of units are represented as numerical values on a continuous scale ranging from 0 to 1. For unit cost computation, the predicted prominence values are linearly scaled to the value range of the prominence labels in the corpus meta-data. The difference between the predicted and actual prominence value of every candidate unit in the search space is added to its unit cost vector. The weight of the prominence cost factor has been set so as to be approximately balanced against the other costs. Thus, the system is capable of selecting a sequence of units that matches the predicted prominence pattern of a target utterance.

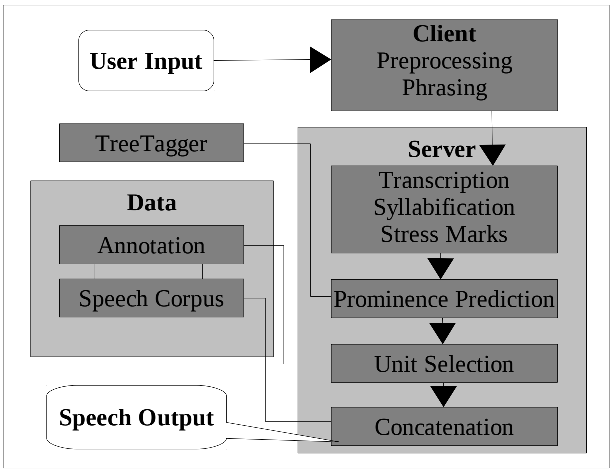

Figure 2: *BOSS system architecture including prominence prediction.*

4. Evaluation

4.1. Prosody

4.1.1. Method

We tested the impact of prominence prediction on the prosody of the BOSS speech output in a pairwise comparison task including two lists of ten pairs of sentences each. Each pair consisted of two instances of the same sentence, one of which was synthesized with the standard configuration of BOSS (*standard* condition) and the other with the BOSS version comprising the prominence prediction module (*prominence* condition). The two lists were designed to serve different purposes: the *diagnostic* list aimed at a specific diagnostic evaluation of the prominence prediction module, while the *global* list was designed to provide a more holistic assessment of overall speech quality.

For the ten stimulus pairs in the *diagnostic* list, we employed a trial-and-error strategy in order to find sentences which the BOSS standard configuration produced with inappropriate prominence patterns, mostly due to overly prominent instances of function words being selected. The members of each stimulus pair from the *diagnostic* list were designed so as to be equal in all other respects, controlling for correct transcription and POS assignment and excluding stimulus pairs if the *standard* stimulus showed peculiarities which might divert listeners' attention from the prominence pattern, such as F0 jumps or bad segmental quality. The *diagnostic* stimuli were restricted to meaningful sentences, exclusively consisting of word-level candidates. It was taken care that a substantial number of alternative candidate units, displaying a variety of different prominence values were available in the corpus at least for the problematic units. In contrast, the stimulus pairs in the *global* list consist of meaningful, but otherwise randomly chosen sentences comprising all unit levels.

The evaluation was conducted using a web interface. Stimulus pairs were presented on separate screens, with a play button and a check box for each stimulus. The individual stimulus pairs appeared in randomized order. The assignment of the individual stimuli to the left or right play button and check box was randomly varied for each stimulus pair. Subjects were instructed to listen to both stimuli up to three times and to tick the check box for the version they preferred

4.1.2. Results

The experiment was completed by 105 subjects, 12 of which were exempted because they were not native speakers of German, reported hearing impairments or background noise. The remaining 93 subjects (36 m, 57 f) were aged between 19 and 60, with a mean of 27.8 years. 75 subjects were experienced in linguistics or phonetics, 18 subjects reported experience with synthetic speech. 44 subjects used built-in speakers, 26 subjects used external loudspeakers, 13 subjects used full-size headphones and 10 subjects used ear buds.

Results of the prosody evaluation procedure are shown in Figure 3. The *prominence* stimuli were preferred in the majority of cases over their *standard* counterparts in both the *diagnostic* (χ^2(1,930)=215.81, p<0.0001) and the *global* (χ^2(1,930)=273.14, p<0.0001) list. This result is consistent for all but two stimulus pairs: in one *diagnostic* pair, there was in fact a clear preference for the *standard* stimulus, probably due to a conspicuous F0 jump in the *prominence* condition. In one *global* pair, both versions were preferred equally often. Inspection of the data did not suggest major influences of any of the mentioned control variables. It can be summarized that prominence prediction has the potential to improve synthetic prosody

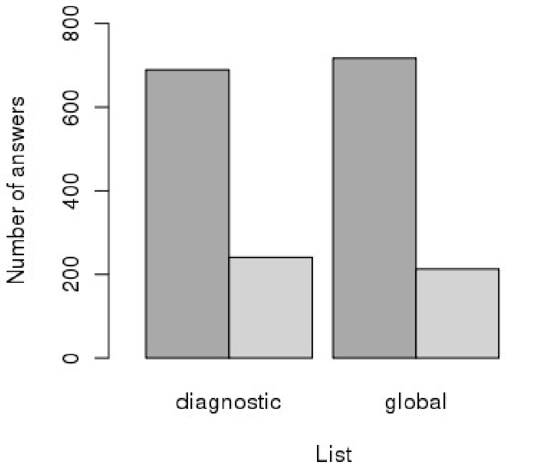

Figure 3: *Numbers of answers with preference for prominence (dark) and standard (light) condition in the diagnostic and global list.*

4.2. Intelligibility

4.2.1. Method

A transcription task was carried out in order to assess possible effects of prominence prediction on the intelligibility of the synthetic speech. We synthesized a set of 20 stimulus pairs for the intelligibility test. Each pair consisted of two instances of the same German sentence, one synthesized with the *standard* and one synthesized with the *prominence* configuration of BOSS. The stimuli were semantically unpredictable sentences, mostly composed of phone-level units. They consisted of either six or seven words and were constructed according to a syntactic template with minor variations. In order to rule out learning effects in subjects due to listening to the same sentence twice, we distributed these stimuli evenly to two subsets, only one of which was played to each participant. This was done such that each subset contained half of the stimuli from either condition and exactly one member of each stimulus pair. The intelligibility test was implemented within the same web application as the first experiment, so that the same subjects took both tests. In order to prevent learning effects, they actually went through the intelligibility test first. Subjects were randomly assigned to one of the two stimulus subsets. Stimuli were presented in randomized order. Each stimulus was presented on a separate screen, with a play button and a text field. Subjects were instructed to listen to the stimulus by clicking the play button and to write down what they had heard. Each stimulus could be played only once. Subjects were informed in advance about this and were also told that they would be listening to meaningless sentences.

4.2.2. Results

The two subsets in the intelligibility test were assigned 46 and 47 subjects, respectively. As suggested by [21], intelligibility was assessed by measuring the Levenshtein distance on word level between subjects' transcriptions and reference transcriptions of the stimuli, after normalizing case and punctuation. For statistical analysis, Levensthein distances were divided by sentence length in terms of syllables, in order to normalize for the fact that the number of syllables varied considerably across test sentences. Figure 4 shows median normalized Levenshtein distances for the *prominence* and *standard* stimuli in the two subsets.

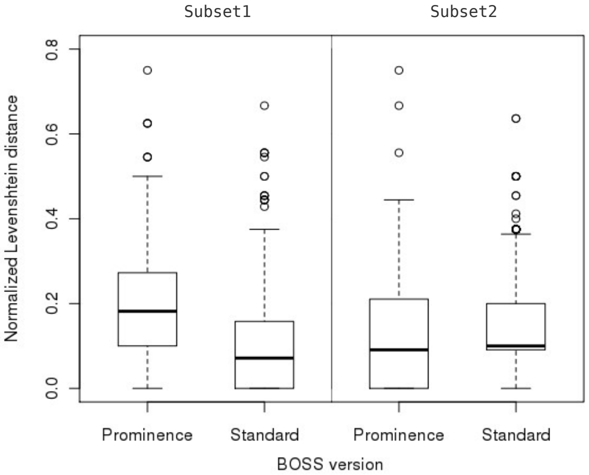

Figure 4: *Median normalized Levenshtein distances for prominence and standard stimuli in both subsets.*

The plot suggests different outcomes in both subsets. Using the R package `lme4`, we fitted linear mixed effects models to the experimental data for statistical analysis, applying stepwise model selection procedures and *p*-value estimation as described in [22]. As for subset 1, a model including *BOSS Version* (t(913)=-2.53, p<0.05,) and *Audio Equipment* (t(913) =-2.699, p<0.01) as fixed and *Stimulus* and by-subject adjustments for *Audio Equipment* as random effects turned out to provide the best fit to our data. In contrast, *BOSS Version* did not make a difference in subset 2 and was not included in the respective model, but an effect of *Audio Equipment* (t(937)=-2.529, p<0.05) was present here as well. Thus, a detrimental effect of prominence prediction on intelligibility is present in subset 1, but not in subset 2. Since there were no systematic differences between the stimuli in both subsets and subjects were randomly allocated, we take the different outcomes to stem from idiosyncratic properties of the individual stimuli.

5. Discussion and Conclusions

We have demonstrated the applicability of perceptual prominence to prosody generation in unit selection synthesis. Our prediction algorithm presents a theoretically well-motivated and empirically adequate solution that is computationally simple and did not require any major alterations to existing synthesis algorithms in our system. Rule-based prosody prediction has widely gone out of fashion, but it presents the advantage of not requiring large quantities of hand-labeled training data. More importantly, results of the pair comparison task suggest that our very simple prediction algorithm is capable of significantly enhancing synthetic prosody.

The somewhat contrary result of the second experiment may be interpreted in such a way that the success of prominence prediction depends on sufficient corpus coverage. If the baseline quality is already poor due to a high density of concatenation points, the additional cost imposed by prominence prediction may even cause further deterioration of the speech output. This was the case in the second experiment, where stimuli were semantically unpredictable and for the most part consisted of phone-level units. This is, in fact, a general problem of prosody prediction algorithms; for example, [14] have observed that “control of prosody comes at the potential cost of lower segmental quality”.

Perspectives for further work include comparison of different methods for prominence prediction and corpus annotation, as machine learning schemes present alternatives to the methods we used for both tasks. As a long-term perspective, it would be interesting to see how successful prominence prediction performs in direct comparison to conventional prosody modeling in the acoustic domain.

6. Acknowledgements

We would like to thank Hendrik Hasenbein for building the web interface and Annett Jorschick for advice on statistics. Further thanks go to three anonymous reviewers for their helpful comments on an earlier draft of this paper and to our patient test subjects.

7. References

[1] Wagner, P., 2000, Evaluating metrical phonology – a computational-empirical approach. In *Proc. KONVENS 2000*, Ilmenau, Germany, 243-248.

[2] Fry, D.B., 1958, Experiments on the perception of stress. In *Language and Speech 1*, 126-152.

[3] 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.

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

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

[6] Widera, C., Portele, T. & Wolters, M., 1997, Prediction of word prominence. In *Proc. EUROSPEECH 1997*, Rhodes, Greece, 999-1003.

[7] 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.

[8] 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.

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

[10] Ross, K. & Ostendorf, M., 1996, Prediction of abstract prosodic labels for speech synthesis. In *Computer Speech and Language (1996) 10*, 155-185.

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

[12] Fackrell, J., Vereecken, H., Martens, J.P. & van Coile, B., 1999, Multilingual prosody modelling using cascades of regression trees and neural networks. In *Proc. EUROSPEECH 1999*, Budapest, 1835-1838.

[13] Wightman, C.W., Syrdal, A.K., Stemmer, G., Conkie, A. & Beutnagel, M., 2000, Perceptually based automatic prosody labeling and prosodically enriched unit selection improve concatenative text-to-speech synthesis. In *Proc. ICSLP'00*, Bejing, 71-74.

[14] 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.

[15] 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.

[16] Breuer, S. & Hess, W., 2010, The Bonn Open Synthesis System 3. In *International Journal of Speech Technology 13*, 75-84.

[17] Windmann, A., Wagner, P., Tamburini, F., Arnold, D. & Oertel, C., 2010, Automatic prominence annotation of a German speech synthesis corpus: towards prominence-based prosody generation for unit selection synthesis. In *Proc. ISCA SSW7*, Nara, Japan, 377-382.

[18] Uhmann, S., 1991, *Fokusphonologie. Eine Analyse deutscher Intonationskonturen im Rahmen der nicht-linearen Phonologie*, Tübingen, Germany, Niemeyer (Linguistische Arbeiten, 252).

[19] Schmid, H., 1995, Improvements in part-of-speech tagging with an application to German. In *Proc. ACL SIGDAT*, Dublin, 47-50.

[20] Tamburini, F. & Wagner, P., 2007, On automatic prominence detection for German. In *Proc. Interspeech 2007*, Antwerp, Belgium, 1809-1812.

[21] Bunnell, H.T., 2010, Crafting small databases for unit selection TTS: effects on intelligibility. In: *Proc. ISCA SSW7*, Nara, Japan, 40-44.

[22] Baayen, H., 2008, *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge: CUP.