PREDICTION OF WORD PROMINENCE

Christina Widera, Thomas Portele, and Maria Wolters

Institut für Kommunikationsforschung und Phonetik (IKP), University of Bonn,
Poppelsdorfer Allee 47, 53115 Bonn, Germany
{cwi, tpo, mwo}@ ikp.uni-bonn.de

ABSTRACT

Control of prosody is essential for the synthesis of natural sounding speech. Text-to-speech systems tend to accent too many words when taking into account only the distinction between open-class and closed-class words. In the prominence-based approach [1], the degree of accentuation of a syllable is described in terms of a gradual prominence parameter. This paper presents the calculation of the prominence level of words based on their word class, the classes of the surrounding words, and their position in a clause. Rules predicting word prominence are derived from statistical analysis of a prosodic database. The hand-crafted rules are compared with the results of several machine learning algorithms on the same material. Furthermore, a perceptual test and an analysis of the resulting speech signals are carried out.

1. INTRODUCTION

Good prosody control improves the naturalness of synthesised speech. Moreover, it aids comprehension. In practice, abstract prosodic labels are derived from the text and then used to control the acoustic parameters of text-to-speech (TTS) systems. This paper focusses on the prediction of the degree of accentuation of words, their prominence. Numerous factors influence the prominence of words. When distinguishing solely between open-class and closed-class words, TTS systems tend to accent too many words (see also [2]). For the London-Lund corpus, Altenberg [3] found that a more precise subclassification of open- and closed-class words brings out more clearly their prosodic potential. Ross & Ostendorf [2] used regression trees to predict the prominence of syllables. After establishing the pitch accent location (accented vs. unaccented) and the pitch accent type (high, downstepped, low) with two different Markov models, the prominence levels defined by F0 levels of pitch accented syllables and normalized energy peaks were predicted by regression trees. In contrast to their approach where prominence is defined by acoustic parameters, prominence is regarded here as a perceptual parameter. In our TTS system prominence operates as an intermediate gradual parameter between linguistics and acoustics [1]. In this paper, word prominence is investigated depending on word classes and position in a clause. Rules predicting word prominence are derived from statistical analysis of a prosodic database. These rules are evaluated by a comparison with the predicted prominence values of four machine learning (ML) algorithms and by a perceptual test.

2. DATABASE

The database [4] consists of 6434 words. It was built from a corpus recorded by three German speakers, two female and one male. The corpus is composed of isolated sentences, question-answer pairs, and short stories. Every syllable of a word has been labelled by three subjects with perceptual prominence values scaled from 0 to 31. Between subjects, the labelled prominences correlate strongly (rho > 0.8; [5]). The prominence of a syllable is taken to be the median of the three labellers' judgements. The prominence of a word is defined by the maximal prominence of its syllables. There are 21 word classes (for a detailed list, see Figures 1 and 2). Each word is assigned information about its word class, the word class of the three preceding words, and the word class of the following word. Furthermore, the position of the word in the clause is taken into account. Five positions are distinguished: first, second, third, medial, and last.

3. RULES

As expected, closed-class words are less prominent than open-class words. Figures 1 and 2 show the prominence of the word classes. There is no significant difference between the prominence value of the two auxiliaries 'will' and 'to have'. The prominence values of words with the same word class differ according to their clause position. Prominence values tend to be increased by about 4 points in clause initial and final positions. If two words of the same word class occur one after the other, one of them will be less prominent. Furthermore, the results indicate that pronouns and co-ordinated

conjunctions have to be subdivided. Personal and reflexive pronouns are less prominent than the other pronouns (e.g. relative pronouns, possessive pronouns). The subdivision of conjunctions is due to their semantics. Contrastive conjunctions like 'also' and 'but' are more prominent than the conjunctions 'and' and 'or'. Although prominence values also depend on the word classes of the preceding and following word, it seems that clause position is more relevant than the surrounding word classes.

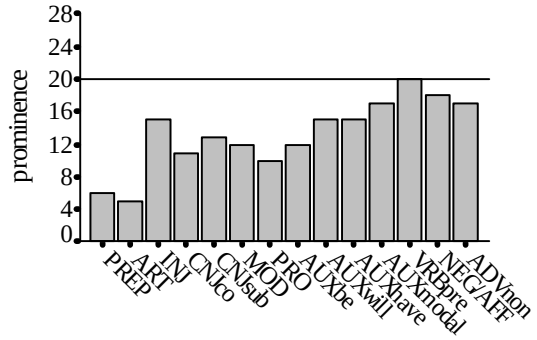

Figure 1: Prominence of subcategorized closed-class words (PREP: prepositions, ART: articles, INJ: interjections, CNJco: conjunctions (co-ordinated), CNJsub: conjunctions (sub-ordinated), MOD: modal particles, PRO: pronouns, AUXbe: forms of the auxiliary 'to be', AUXwill: forms of the auxiliary 'will', AUXhave: forms of the auxiliary 'to have', AUXmodal: modal auxiliaries, VRBpre: detachable prefixes of verbs, NEG/AFF: negations/affirmations, ADVnon: adverbs (non-flectional)).

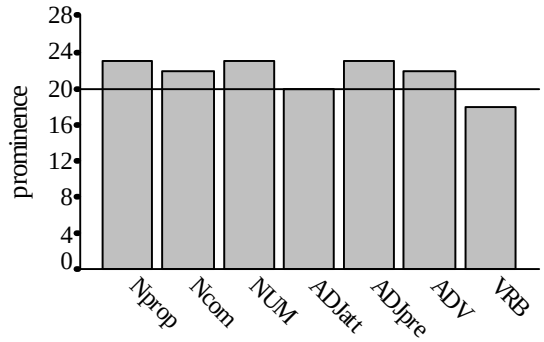

Figure 2: Prominence of subcategorized open-class words (Nprop: proper nouns, Ncom: common nouns, NUM: numerals, ADJatt: adjectives (attributive use), ADJpre: adjectives (predicative use), ADV: adverbs (adverbial use of adjectives), VRB: main verbs).

4. AUTOMATIC PREDICTION OF PROMINENCE

Four ML algorithms were used for the automatic prediction of prominence: IGTree (information gain tree; [6]), SCT (semantic classification tree; [7]), T2 (2 level decision tree; [8]), and two artifical neural networks. The features used for classification were word class, surrounding word classes, and clause position. Both networks had 130 input units and two hidden layers (90-40). One network (NN 1) had 1 output unit, the other 32 (NN 32). In the case of NN 1, prominence is regarded as a continuous parameter. Before training IGTree, the features had to be ordered by their relevance for the classification. Two arrangements were chosen: in both the word class was the most important feature, the second one was either the preceding word class ($IGTree_{pWC}$) followed by the following, the other two preceding word classes, and the clause position or the clause position ($IGTree_{CP}$) followed by the directly preceding, following, and the other two preceding word classes.

5. RESULTS

5.1. Performance

Learning was complicated by a great dispersion of the prominence values within a word class. This is caused not only by lexical and syntactic factors, but also by differences between the three speakers, e.g. different interpretation of the discourse structure, speaking style, etc. (c.f. [2]). We use the mean deviation (md) calculated on the confusion matrix and the correlation between the predicted prominence and the prominence values of the database to judge the ability of the algorithms and of the hand-crafted rules to generalize. The mean deviation is defined by:

$$md = \frac{1}{n}\sum_{n}\left|P_{(D)} - P_{(P)}\right|$$

with n = number of cases; $P_{(D)}$ = labelled prominence in the database; $P_{(P)}$ = predicted prominence. All algorithms were tested on the whole training set, since the hand-crafted rules had been written using the complete database.

The recogniton rates[1] are very low for all algorithms and the hand-crafted rules (<= 41%, see Figure 3). The recognition rates, the mean deviations and the correlations of IGTree and SCT with multiple levels are superior to T2 with only 2 levels (Figures 3-5). The results of SCT, which has no information about the relevance of the features for classification, are lower than those of IGTree. Furthermore, the tree of $IGTree_{CP}$ is smaller than the one of $IGTree_{pWC}$. This supports the hypothesis that clause position is more important than the classes of surrounding words. The fact that the mean deviation for NN 1 is smaller than that of NN 32 (Figure 4) suggests that prominence is a gradual

[1] Strictly speaking, the recogniton rate of NN 1 is a hit rate defined by the rounded prominence values.

parameter as proposed in [5]. The mean deviation and correlation of the hand-crafted rules and of T2 differ only slightly (Figures 4 and 5). Furthermore, the decision trees generated by IGTree, T2, and SCT have been transformed into rule sets. We found that the hand-crafted rules are far less complex than the automatically generated rules.

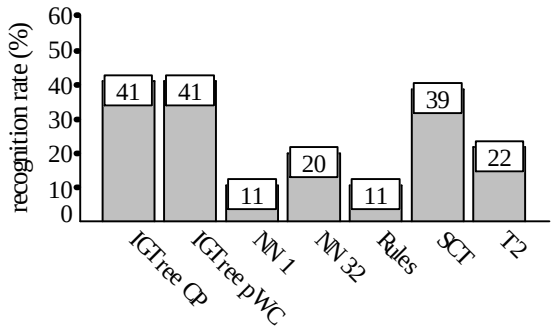

Figure 3: Recognition rate of the prediction of prominence of the automatic classificators and of the hand-crafted rules.

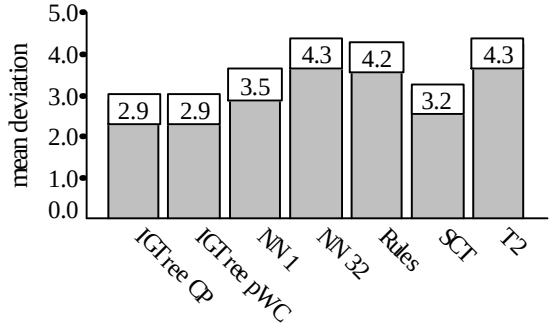

Figure 4: Mean deviation of the prediction of prominence of the automatic classificators and of the hand-crafted rules.

5.2. Perceptual and acoustic evaluation

The hand-crafted rules, IGTree, SCT, T2, and NN 1 were evaluated perceptively. The following five short sentences were synthesized by the TTS system HADIFIX [9] with the prominence values predicted by the five algorithms:

1. *Heute ist es bitterkalt.*
   (Today it is bitterly cold.)
2. *Bitte sei doch bis um 15.30 Uhr in diesem Gasthaus.*
   (Please, be in this guesthouse by 3.30 pm.)

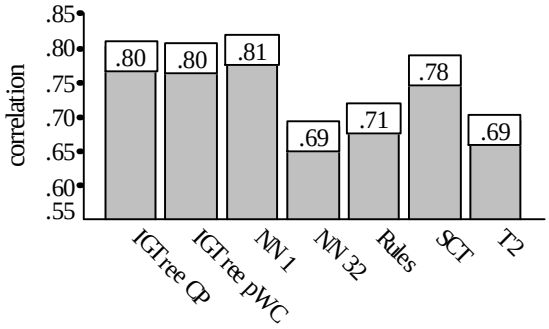

Figure 5: Correlation of the prediction of prominence of the automatic classificators and of the hand-crafted rules (α<.01).

3. *Hastig eilen die Leute vorbei.*
   (Quickly the people hurry past.)
4. *Susanne kann leider nicht zu Frank kommen.*
   (Unfortunately, Susan cannot come to Frank.)
5. *Bis zum 1. Mai werden wir dieses gelbe Fahrzeug ausleihen.*
   (Until the first of May, we are going to rent this yellow car.)

The sentences were presented pairwise (AB and BA) to 11 subjects. The subjects were asked to listen to 100 pairs (5*5*4*2/2) and to decide which sentence of a pair sounds more natural.

There are no significant differences in the subjects' overall judgement. Only one sentence (sentence 1) shows significant differences between the judgements (χ^2=18.09, n=220; f=4, α<.01; Figure 6). Informal discussions with some of the subjects indicate that the naturalness of sentences was difficult to judge because all versions appeared to be very similar. Additionally, the decisions were complicated by a partially defective duration control in the version of the synthesis system used for the test.

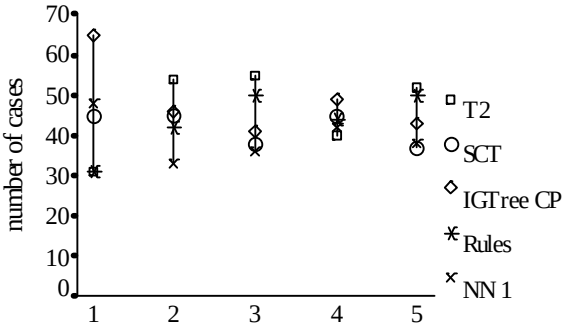

Figure 6: Preferred version for each sentence (IG-Tree$_{CP}$, SCT, T2, NN 1, hand-crafted rules).

For each pair, the resulting speech signals were compared using dynamic time warping and the pairwise correlation between the predicted prominence values was calculated. The perceptual ability of the subjects to

differentiate was defined by the χ^2-value of the judgements of the sentences of each pair. The comparison of the three results (acoustical, perceptual and predicted prominence values) shows a negative correlation between speech signal differences and predicted prominence values (ρ=-.515, n=50, α<.01). This means that a strong correlation of the predicted prominence values of the sentences of a pair corresponds to small acoustical distances of resulting speech signals; but probably the acoustical distances between prominence values are small. No correlation is found between the predicted prominence values and the perceptual evaluation and between acoustical values und perceptual evaluation.
From these results we can conclude that the differences between the predictated prominence values of the ML algorithms and the hand-crafted rules are negligible. This matches the results of the quantitative evaluation. Although the prominence values have acoustic correlates, their effect appears to be masked by problems in the duration control in the version of the synthesis system used for this experiment [10].

6. CONCLUSION

The object was to derive rules from an analysis of a prosodic database for predicting the prominence of words by their word classes, and their position in a clause, the classes of the preceding words, the following word. These rules were evaluated by a comparison with the predicted prominence values of four ML algorithms and by a perceptual test.
The results indicate that for predicting the prominence of words, word class and the position in a clause are most relevant. The prominence values of the ML algorithms and of the hand-crafted rules do not differ significantly. Despite their simplicity, the hand-crafted rules are adequate.
Since part-of-speech information is insufficient for predicting prominence (c.f. [11]), further work is required to examine the influence of other factors such as semantics and discourse structure on word prominence.

REFERENCES

[1] T. Portele and B. Heuft, „Towards a prominence-based synthesis system“, *Speech Communication*, Vol. 21, pp. 61-72, 1997.

[2] K. Ross and M. Ostendorf, „Prediction of abstract prosodic labels for speech synthesis“, *Computer Speech and Language*, Vol. 10, pp. 155-185, 1996.

[3] B. Altenberg, „*Prosodic patterns in spoken English. Studies in the correlation between Prosody and Grammar for text-to-speech conversation*“, Lund Studies in English 76, Lund University Press, Lund, 1987.

[4] B. Heuft, T. Portele, F. Höfer, J. Krämer, H. Meyer, M. Rauth, and G. Sonntag, „*Parametric description of F0-contours in a prosodic database*“, Proc. ICPhS'95, pp. 378-381, Stockholm, 1995.

[5] B. Heuft and T. Portele, „*Synthesizing prosody: A prominence-based approach*“, Proc. ICSLP'96, pp. 1361-1364, Philadelphia, 1996.

[6] W. Daelemans, A. van den Bosch, and T. Weijters, „IG-Tree: Using trees for compression and classification in lazy learning algorithms“, *AI Review* (to appear).

[7] R. Kuhn and R. De Mori, „The application of Semantic Classification Trees to natural language understanding“, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 17 (5), pp. 449-460, 1995.

[8] P. Auer, R. C. Holte, and W. Maass, „Theory and applications of agnostic PAC-Learning with small decision trees“, in: A. Prieditis and S. Russell (eds.), „*Machine Learning: Proceedings of the 12th International Conference*“, Morgan Kaufmann Publishers, 1995.

[9] T. Portele, F. Höfer, and W. Hess, „*Structure and representation of an inventory for German speech synthesis*“, Proc. ICSLP'94, pp. 2182-2185, Yokohama, 1994.

[10] H. Meyer, „*Silben- oder Lautdauersteuerung?*“, 8. Konferenz Elektronische Sprachsignalverarbeitung, Cottbus (submitted).

[11] J. Hirschberg, „Using discourse context to guide pitch accent decisions in synthetic speech“, in: G.Bailly, C. Benoît, and T. R. Sawallis (eds.), „*Talking Machines: Theories, Models, and Designs*“. Elsevier Science Publishers, North-Holland, Amsterdam, 1992.