Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 24 (2001), 155-166.

ACOUSTICAL AND LEXICAL/SYNTACTIC FEATURES TO PREDICT PROMINENCE[1]

Barbertje M. Streefkerk,
Louis C.W. Pols and Louis F.M. ten Bosch

Abstract

In this study acoustical as well as lexical/syntactic correlates of prominence are analyzed and discussed. Prominence is defined at the word level and is based on listener judgments. Spoken sentences from many different speakers, taken from the Dutch Polyphone corpus of telephone speech, are analyzed. A selection of useful acoustical input features is chosen for classification of word prominence, by means of Feed Forward Nets. For an independent test set of 1,000 sentences about 79% of the words are correctly classified as prominent or not. We also developed an algorithm, based on text input, using linguistic/syntactical features derived from text only, to predict prominence. The prediction agrees with the perceived prominence in 81% of the cases for the independent test set.

The results of this project show that certain acoustical and linguistic correlates of prominence can be extracted automatically and can be used to accurately predict prominence with a consistency similar to the pominence assignment by naive listeners.

1 Introduction

In current speech-technology applications the use of prosodic information is not without difficulties.

In *text-to-speech synthesis*, accentuation (most of the time in terms of pitch accents) and phrasing can only rely on textual information (Hirschberg, 1990). Variation in strength of the boundaries and in prominence of the pitch accents is rarely taken into account in current speech synthesis systems. However, the simple distinction between accented and non-accented is generally not sufficient to make the intonation natural sounding (see for example http://www.fluency.nl, Fluent Dutch Text-To-Speech Version 1.0). Certain attempts are made to introduce the notion of prominence (Portele & Heuft, 1997; Fant et al., 2001) for speech synthesis purposes as well. The first step to make synthetic speech more natural is to predict the prominence based on available textual information. The second step is to transform the predicted prominence into acoustical features in order to implement the various prominence levels and to make them perceivable.

[1] Parts are published in Streefkerk et al. (2001).

In most current *speech recognition* systems, prosodic information is hardly used at all. However, a device that would indicate the degree of prominence of every word or of specific words would be helpful in several speech recognition applications. Prominence could then serve as an indication for islands of reliability, that can carry important and/or new information, such as a negation. Proper recognition of such words can be very advantageous. It could also be used in various ways in *dialogue systems*. Dialogue management frequently requires disambiguation of two or more possible interpretations of an utterance, for instance of the type "in CAPable hands" versus "INcapable hands". The present project might help to incorporate prosodic features into speech recognition and into speech synthesis applications.

For a set of 2,224 Dutch sentences, word prominence is marked by listeners and the acoustical as well as the textual correlates are investigated and tested for their predictability for prominence. More details are presented below.

2 Speech material and initial prominence labeling

In this research project the speech material is taken from the Dutch Polyphone Corpus (Damhuis et al., 1994). Phonetically rich newspaper sentences are spoken by people from all over the Netherlands. Each individual reads aloud five of these sentences which are then recorded over the telephone. From this material of over 5,000 speakers, we have randomly selected 1,244 sentences for training and an additional 1,000 sentences for testing. Altogether 497 different speakers are involved. These sentences generally have a rather simple grammatical structure, with on average 10 words per sentence. However, the variability in this material is substantial, both in terms of channel conditions and home-environment recording conditions, as well as in terms of speakers (age, sex, education, speaking style), which makes this a very challenging corpus.

2.1 Design of the training set

10 Naive Dutch listeners were asked to mark all prominent words of the 1,244 sentences of the training set. A subset of 50 sentences was presented twice, in order to get information about the within-listener consistency as well (Streefkerk & Pols, 1998). The individual marks of the listeners are binary (either 0 or 1), but the summed marks per word result in a prominence scale (per word) from 0 to 10. The agreement between the 10 listeners can be expressed by Cohen's Kappa (κ), which appeared to be on average 0.5 with a standard deviation of 0.16.

Because this 10-point scale only suggests a high accuracy and furthermore depends on the actual number of listeners used, we simplified it, by means of a hierarchical cluster analysis, to a 4-point scale (0, I, II, III). Zero indicates no prominence at all and III is the category of highest word prominence. This scale is further simplified to a binary scale putting 0 and I together (non-prominent) as well as II and III (prominent). This merging of categories makes the scores more comparable with those from the test set. See Table 1 for the distribution of these two combined classes in the training data.

Table 1. Absolute and relative number of words in the training set of 1,244 sentences belonging to the prominent (II and III) and non-prominent (0 and I) class.

Prominence	Number	Percentage
0 and I	7,818	59.6
II and III	5,301	40.4
Total	13,119	100

2.2 Design of the test set

For efficiency reasons, the test set of 1,000 sentences was processed in a slightly different way. From the initial group of 10 listeners only the one with the highest between (mean $\kappa = 0.55$) and within ($\kappa = 0.8$) agreement was chosen to mark the prominence of the words in this test set. So, for these words only a binary score of 0 (indicating no prominence) or 1 (indicating prominence) was available. Of the total of 10,330 words, 3,998 words were marked by this listener as prominent, this is 39%, which is comparable to the 40.4 % for classes II and III in the training data, see Table 1.

3 Prominence classification based on acoustical features

An HTK speech recognizer (Wang, 1997) automatically segmented the speech material at the phoneme level by using a forced alignment. Since we suppose the text to be known, word boundaries, syllables boundaries and segment boundaries are then also available and measurements can be done at all these levels. Because the prominence marks are at the word level and acoustical features are generally not extracted at this word level but most of the time at the level of vowels and syllables, we limit ourselves for acoustic feature extraction in the case of polysyllabic words to the lexically stressed syllables only.

It is to be expected that the stressed syllables in prominent words, and thus also the vowels, are louder, longer and show more pitch variation than in the non-prominent words. Furthermore, it is certainly worthwhile to explore how much discriminative power can be gained by normalizing, among other things, for intrinsic vowel duration and for speaking rate.

3.1 Acoustical features

In general the following features are supposed to have an influence on word prominence: F_0, intensity, duration, and perhaps spectral quality (see also Batliner et al., 1999). The basic features that we use are: vowel and syllable duration, vowel

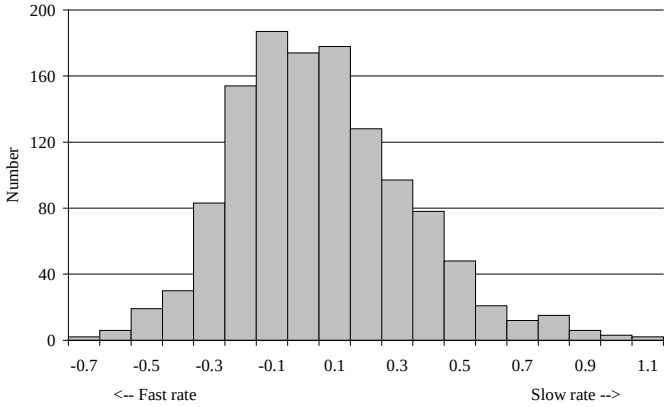

Fig. 1. Distribution of the sentences over the sentence speaking rate.

intensity, F_0 range per word and per syllable, and F_0 median value per word and per syllable. Beyond that, additional features are used, namely the sentence speaking rate, the overall intensity of a sentence, and the median F_0 of a sentence, plus some normalized features, namely vowel duration normalized for the intrinsic duration and vowel intensity normalized for intrinsic intensity. Without going into detail for all individual acoustical features, some details for the sentence speaking rate are given below.

The sentences show overall variation in sentence speaking rate. This speaking rate (r) is defined as the average normalized phoneme duration (τ) per sentence, see (1).

$$r = \frac{1}{N}\sum_{i=1}^{N}\tau_i \qquad \tau = \frac{d-\mu}{\sigma} \tag{1}$$

Zero thus implies an average sentence speaking rate, a positive value implies a slow rate and a negative value a fast rate. The actual variation in speaking rate over all sentences can be seen in Fig. 1. The mean vowel duration as a function of the sentence speaking rate in these 1,244 sentences is shown in Fig. 2. Since the vowels belonging to the most prominent words (scale III) are displayed separately from those belonging to the least prominent words, it can easily be seen that the former are substantially longer.

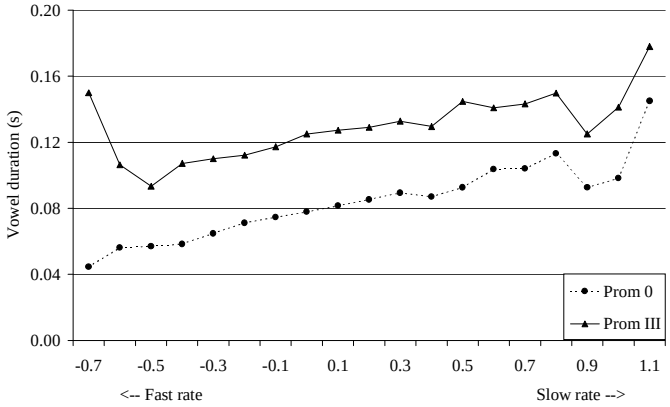

Fig. 2. Average vowel duration in seconds as a function of the sentence speaking rate. The two curves represent the most (scale III) and the least (scale 0) prominent vowels.

3.2 Classification with feed forward nets

In total a set of 12 features is used to train feed forward nets for prominence classification. By using a hidden layer with a variable number of nodes, we trained a feed forward net that had a binary output (prominence 0 or 1). The original training set was divided into two parts; one part is used as training and the other part is used as development test set. Several networks were trained, both with equal numbers (randomly selected) and with unequal numbers of non-prominent (scale 0 + I) and prominent (scale II + III) data.

3.3 Results

Under the unbiased condition the best overall performance of prominence classification is 82.01% on the Development test set (see Table 2). These results are achieved with 14 nodes in the only hidden layer. The best performance for a biased trained network (10 hidden nodes in the only hidden layer) results in an overall performance of 80.11% on the Development test sentences (see Table 2).

The between-listener agreement, expressed in Cohen's Kappa (κ), could be calculated for the results of the neural network, and the 'optimal' listener from the training set who marked all 1000 test sentences. Kappa values are 0.53 (unbiased training) and 0.57 (biased training), see Table 3, expressing the agreement between the neural network and the 'optimal' listener. Similar values (on average $\kappa = 0.50$; Std. Dev. = 0.16) were measured for the between-listener agreements for the test set.

This means that the neural network behaves similarly to any listener, or in other words, the differences in prominence classification are as accurate as the prominence classification of any listener.

Table 2. This table presents the recognition rates of prominence classification on the Training set and a Development test set. The total numbers as well as the percentages are given for the networks trained under unbiased and biased conditions, the networks with the topology of 12-14-2 and 12-10-2 were optimal.

	Equal numbers (unbiased)		%
All	Training set		**86.66**
Non-prom	2232	418	84.67
Prom	289	2361	89.54
All	Dev. test set		**82.01**
Non-prom	2114	537	79.73
Prom	417	2234	84.28
	Unequal numbers (biased)		%
All	Training set		**83.34**
Non-prom	3279	597	84.60
Prom	496	2187	81.51
All	Dev. test set		**80.11**
Non-prom	3228	714	81.89
Prom	591	2027	77.43

Table 3. This table presents the recognition rates of prominence classification on the independent Test set of 1000 sentences. The total numbers as well as the percentages are given for the networks trained under unbiased and biased conditions, the networks with the topology of 12-14-2 and 12-10-2 were optimal.

	Unbiased training			Biased training		
	Non-prom	Prom	%	Non-prom	Prom	%
	Test set (biased)		**77.1**	Test set (biased)		**78.9**
Non-prom	4907	1425	77.5	5232	1100	82.6
Prom	942	3056	76.4	1079	2919	73.0
Measure of agreement (κ)			0.53			0.57

4 Prominence prediction based on textual information

In synthesis applications it would improve the speech quality and the communicative function of the speech material generated, if we were able to properly predict the word prominence from textual information only. In most present-day commercial text-to-speech systems one does not get much further than giving all content words a pitch accent. In the present paper we explore the possibilities of using several textual correlates to properly predict prominence. Since it is rather ambitious to extract meaning from a given sentence in a given context, we limit ourselves to correlates that can be derived automatically from text, such as: POS, number of syllables, position of words in the sentence and co-occurring word classes, such as the Adjective-Noun combination. Word classes were assigned automatically for the test and training sentences by a memory-based Part-of-Speech (POS) tagger (Daelemans et al., 1996), which compares a particular word in a particular context with a most similar case in memory. Comparing its performance with hand-derived POS labels for the training set of 1244 sentences shows 92% correspondence in labels.

4.1 Textual features

11 Word categories are distinguished: Articles, Conjunctions, Prepositions, Pronouns, Auxiliary Verbs, Verbs, Numerals, Adverbs, Adjectives, Nouns, and Negations. Table 4 shows the frequency of occurrence of these word categories in the training data, as well as the mean prominence score (between 0 and 10) and its standard deviation per word class. The word classes are ordered from least to highest average prominence. It is clear that, next to Nouns, also Numerals, Negations, and Adjectives receive high prominence.

Table 4. The frequency of occurrence of the 11 word categories in the training sentences. The mean perceived prominence score per category and its standard deviation are also given. A dashed line separates the function words from the content words.

Word Class	Number of words	Prominence Mean	Prominence Stand. Dev.
Article	1,912	0.1	0.8
Auxiliary verbs	708	0.3	1.2
Preposition	1,795	0.4	1.4
Conjunction	434	0.4	1.2
Pronoun	1,121	1.5	2.8
Verb	1,734	2.6	3.1
Adverb	765	3.8	3.5
Noun	3,173	5.6	2.8
Numeral	327	5.7	3.1
Negation	173	6.2	3.1
Adjective	977	6.3	2.8
Total	13,119		

Table 5. Mean perceived prominence as a function of the number of syllables in the words, as well as its standard deviation.

Num Syllables	Total occurrence	Prominence Mean	Prominence Stand. Dev.
1	7,751	1.4	2.7
2	2,769	4.4	3.4
3	1,571	5.5	3.0
4	729	5.9	2.8
5	241	6.1	2.3
6	57	6.4	2.4
7	1	7.0	-
Total	13,119		

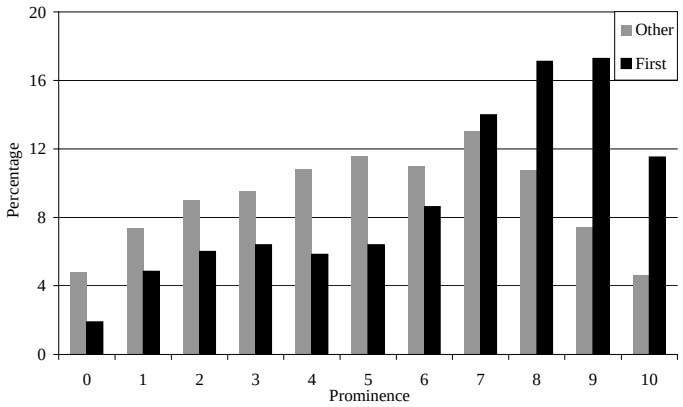

Fig. 3. Distribution of the prominence of the word classes Noun, Adjective, Numeral, and Negation either at the initial position of the sentence, or elsewhere.

In Table 5 we show the distribution of prominence as a function of the number of syllables in the words. Again it is clear that the longer words generally have a higher probability of being prominent.

From Fig. 3 it can be seen that generally the prominence of the word classes Noun, Adjective, Numeral and Negation is much higher if these words occur in the first position of the sentence rather than elsewhere in the sentence. This is another element that is taken care of in developing an algorithm to predict prominence on textual information only.

The same is true for the phenomenon illustrated in Fig 4. This shows that Nouns preceded by an Adjective generally have a substantially lower prominence than all other Nouns. These and other characteristics in our sentence material have been used to optimize our algorithm for predicting prominence.

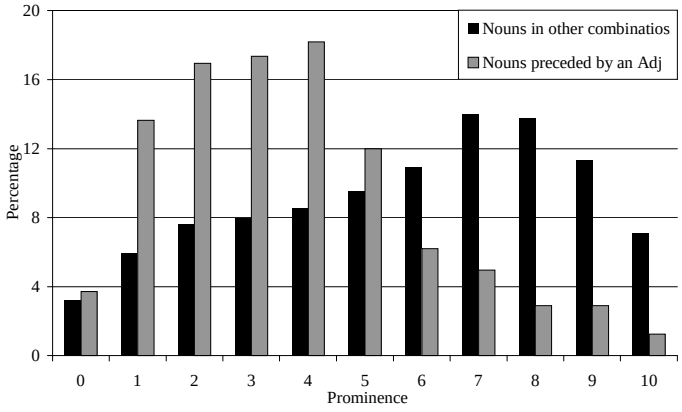

Fig. 4. This is a comparison of the distribution of the prominence over the Nouns, when they are either preceded by an Adjective or in all other positions.

4.2 Algorithm to predict prominence

From the regularities found in our data, we have so far derived the following rule set:

- **rule I:** every content word receives one mark;
- **rule II:** every word from the classes {Noun, Adjective, Numeral, Negation} receives an additional mark;
- **rule III:** every polysyllabic word from the classes {Pronoun, Verb, Adverb} receives an additional mark, and every word from the classes {Noun, Adjective, Numeral, Negation} receives once more an additional mark;
- **rule IV:** the first content word in the sentence receives an additional mark;
- **rule V:** every Noun preceded by an Adjective is decreased by one mark;

4.3 Results

Table 6 shows the results of applying these rules to the 1244 training sentences. The second column gives the frequency of occurrence of each prominence mark. The next columns indicate the related mean values and standard deviations of the actually

perceived prominence. Generally there is a good overall relationship between perceived and predicted prominence.

Table 6. Predicted prominence marks compared to the actually perceived mean prominence marks (scale from 0-10) for the 1244 *training sentences*.

Predicted Prominence	Total occurrence	Prominence Mean	Std. Dev.
0	5,796	0.40	1.36
1	1,480	3.15	3.40
2	2,430	3.84	3.09
3	2,673	5.85	2.77
4	740	7.50	2.25
Total	13,119		

The actual test of course should be done with the independent test set of 1000 sentences, which however has the drawback that the perceived prominence level is only defined in a binary way. The same parser used for the training data automatically produced POS tags for these sentences of the test set as well. With the help of the automatically derived word class labels, the number of syllables in each word, and the position of the word in the sentence, the various prominence levels were predicted according to the algorithm described in the former section. The prominence prediction on a 5-point scale and the prominence judgments of the one listener had to be matched, as presented in table 7. The middle section of the predicted prominence scale is distributed over prominence and non-prominence.

Table 7. Perceived prominence and predicted prominence marks for the independent *test set*.

Perceived prominence	Predicted prominence marks					
	0	1	2	3	4	Total
0	4001	841	930	516	44	6332
1	180	272	1284	1709	553	3998
Total	4181	1113	2214	2225	597	10330

Table 7 shows that prominence prediction with mark 4 is rare, but if it occurs then the word is almost always perceived as prominent. For these independent test data we reduce the number of predicted categories from 5 to 2 categories (prominence marks 0 and 1, just as prominence marks 2, 3 and 4 are put together). A direct comparison with the perceived binary scores then becomes possible. We observe to our satisfaction that the overall performance, even on this independent test set, can reach 81.2% correct classification. The exact data are given in Table 8.

Table 8. This table presents the predicted prominence marks and the perceived marks and an overall result in percentage of the performance.

Perceived	Predicted prominence		Total	%
prominence	0 +1	2+3+4	-	**81.2**
0	4842	1490	6332	76.5
1	452	3546	3998	88.7
Total	5294	5036	10330	-
Measure of agreement (κ)				0.62

Since the perceived prominence scores for the test set are derived from one listener, it is possible to calculate what the amount of agreement is between this single listener and the prediction achieved through the use of rules based on textual input. The resulting Cohen's Kappa of $\kappa = 0.62$ is even higher than the between-listener agreement, with an average of $\kappa = 0.50$ and a standard deviation of 0.16.

The analyses in this study show that the automatic classification of prominence can achieve a performance rate that is indistinguishable from that achieved by a group of naive transcribers. This result has been obtained by using sentences with a relatively simple grammatical structure. More research is required to optimize the predicted prominence marks also for other speech material and to transform these marks into adequate acoustical features and then into an appropriate synthetic speech quality.

5 Concluding remarks

This study underlines that prominence is reflected in the acoustic and the linguistic domain, and that a binary prominence prediction with a selected set of relatively simple lexical/syntactic features can lead to a similar performance as that of naive listeners.

This also allowed us to run the opposite test, namely to predict the word prominence, based on acoustical features extracted from the speech signal only. In section 3 we have shown that a simple feed forward net with one hidden layer, can predict this word prominence rather well, if fed with the appropriate acoustical features.

This study has shown that it is possible to predict with reasonable accuracy the word prominence from linguistic/syntactical features based on the isolated sentence text input only. We were able to test the accuracy of this prediction since these sentences were also available in spoken form in an annotated speech corpus.

References

Batliner, A., Buckow, J., Huber, R., Warnke, V., Nöth, E. & Nieman, H. (1999): "Prosodic feature evaluation: Brute force or well designed?", *Proceedings ICPhS'99*, San Fransisco, 3: 2315-2318.

Daelemans, W., Zavrel, J., Berck, P. & Gillis, S. (1996): "MBT: A memory based part of speech tagger generator", *Proceedings of the 4th Workshop on Very Large Corpora*, Copenhagen, 14-27.

Damhuis, M., Boogaart, T., in 't Veld, C., Versteijlen, M., Schelvis, W., Bos, L. & Boves, L. (1994): "Creation and analysis of the Dutch Polyphone Corpus", *Proc. ICSLP'94*, Yokohama, Vol. 4: 1803-1806.

Fant, G., Kruckenberg A., Liljencrants, J. & Botinis, A.(2001): "Prominence correlates. A study of Swedish" *Proceedings of the Eurospeech'01*, Aalborg, 657-660.

Hirschberg, J. (1990): "Accent and discourse context: Assigning pitch accent in synthetic speech", *Proc. of the 8th National Conference on AI*, Menlo Park, 952-957.

Portele, T. & Heuft, B. (1997): "Towards a prominence-based speech synthesis system", *Speech Communication*, 21: 61-72.

Streefkerk, B.M. & Pols, L.C.W. (1998): "Prominence in read aloud Dutch sentences as marked by naive listeners", *Proceedings of the 4th Conf. on Language Processing KONVENS-98*, Bonn, 201-205.

Streefkerk, B.M., Pols, L.C.W. & ten Bosch, L.F.M. (2001): "Up to what level can acoustical and textual features predict prominence", *Proceedings of Eurospeech 2001*, Aalborg, 811-814.

Wang, X. (1997): *Incorporating knowledge on segmental duration in HMM-based continuous speech recognition*, Ph.D. thesis, University of Amsterdam, SLLU 29: 190 pp.