AUTOMATIC DETECTION OF PROMINENCE (AS DEFINED BY LISTENERS' JUDGEMENTS) IN READ ALOUD DUTCH SENTENCES

Barbertje M. Streefkerk, Louis C.W. Pols, and Louis F.M. Ten Bosch[1]

Institute of Phonetic Sciences, University of Amsterdam, Herengracht 338, 1016 CG Amsterdam,

The Netherlands, barber@fon.hum.uva.nl

[1]Lernout & Hauspie Speech Products N.V., Belgium

ABSTRACT

This paper describes a first step towards the automatic classification of prominence (as defined by naive listeners). As a result of a listening experiment each word in 500 sentences was marked with a rating scale between '0' (non-prominent) and '10' (very prominent). These prominence labels are compared with the following acoustical features: loudness of each vowel, and F_0 range and duration of each syllable. A linear relationship between the rating scale of prominence and these acoustical features is found. These acoustical features then are used for a preliminary automatic classification to predict prominence.

1. INTRODUCTION

For various speech technology applications it is necessary to know which acoustical features play a role in the perception of prominence. For speech synthesis the application of prominence is demonstrated in the research of Portele and Heuft [1]. This prominence-based approach turns out to be a useful interface between linguistics and acoustics. Mayer [2] suggests using prominence of words to disambiguate sentences in which the pronominal reference is unclear. In this kind of ambiguous sentences the notion of pitch accents is not enough for disambiguation. This underlines that prominence and its realization in the speech signal can be useful in speech synthesis and speech recognition, especially in applications where ambiguous sentences occur.

In natural speech the relationship between prominence and certain acoustical correlates, such as F_0, duration and intensity, is complex. Much is known about the acoustical correlate F_0 and its close relation to pitch accents, much less is known about other acoustical correlates such as intensity and duration. Also less is known about the variability within and between speakers to emphasize words in fluent speech. In this paper we present, next to F_0, some acoustical measurements on duration and intensity and its relation to prominence.

Despite the fact that prominence can be useful as an interface between acoustics and linguistics, prominence is not a very well defined term in literature. However, a common definition of prominence is that it refers to those words or syllables that are perceived as standing out from their environment. Or to put it in another way: prominence refers to a greater perceived strength of words in a sentence [3, 4]. Therefore, in this study prominence was defined through judgments of naive listeners, who were instructed to mark all those words they perceived to be spoken with emphasis.

In this paper we first describe the speech material used, followed by the design, procedure and results of the listening experiment to define prominence. Next we outline the preprocessing of the sentences and go in to the acoustical measurements, and discuss them as well as their relation to perceived prominence. Finally, we present some initial results of the automatic classification to predict prominence, by using the acoustical measurements as input features to a neural net classifier.

2. THE SPEECH MATERIAL

The 500 read aloud Dutch sentences used in this study were taken from the Dutch Polyphone Corpus [5], which was recorded by SPEX and KPN (Leidschendam). This large speech corpus contains the speech of 5000 Dutch speakers who had to read aloud, among other things, 5 phonetically rich sentences, which were recorded over the telephone. This speech material, with its high speaker variability, is characteristic of many speech technology applications. For the listening experiment 500 different sentences spoken by 100 different speakers, 50 male and 50 female speakers, were selected. All 5 phonetically rich sentences per speaker were included. On average the 500 sentences contain 10.4 words per sentence. Because the sentences were read aloud without any specific context the words which stand in focus were not retrievable. This could be a complicating factor for further research.

3. LISTENING EXPERIMENT FOR INITIAL LABELING

In our approach we deliberately use naive listeners (10) with the aim to get for each word a label of prominence. Each listener has to mark for all 500 sentences those word(s) which are spoken with emphasis. This instruction is used as an operational definition of prominence. The cumulative score over all 10 listeners is an indication how prominent a given word is. As a first step the words with a prominent score of (8, 9, or 10) are defined as the prominent words and the words which were never marked as being spoken with emphasis as the non-prominent words. Another possibility is, to treat the cumulative score of the 10 listeners as a rating scale of prominence where '0' means non-prominent and on the other end of the scale '10' means very prominent.

3.1. Procedure and Design

500 phonetically rich sentences spoken by 50 male and 50 female speakers are presented to 10 listeners To test how consistent the listeners were, the first 50 sentences were presented to each listener twice. Space does not permit us to discuss the within and between listener differences, but for more details see [6]. The 550 sentences (500 + 50) were randomly presented in 4 sessions, which differed per listener. The listeners listened through closed headphones. The first two sessions contained 150, and the last two sessions contained 125 sentences. The 10 listeners were all students from the Faculty of Humanities at the University of Amsterdam. The perception experiment was performed on a UNIX workstation. The printed words of each sentence were displayed on the monitor with a button underneath each word. The subjects could click on the button when a given word was perceived as being spoken with emphasis. The scores of each listener were automatically stored.

3.2. Resulting Labels from the Listening experiment

In table 1 the absolute and relative judgements over all 10 listeners are presented. Each listener judged the first 50 sentences twice, but in this table we only included the 50 sentences which were judged the second time, because in the first 50 the learning effect may still prevail. In the experiment 621 words (303+212+106) were marked as prominent by 80% or more of the listeners. This is 11.9% of the total number of words. Because there are, on average, 10.4 words per sentence, this results in 1.24 prominent word per sentence. Furthermore, it must be mentioned that about half of the words (50.6%) are never judged as prominent.

Value	Freq. words	%	Freq syllables Lexical stress	No Lexical stress	total
0	2631	50.6	516	2585	3101
1	357	6.9	226	417	643
2	246	4.7	202	309	511
3	221	4.2	195	306	501
4	242	4.7	215	354	569
5	266	5.1	244	415	659
6	273	5.2	260	425	685
7	346	6.6	326	573	899
8	303	5.8	277	454	731
9	212	4.1	183	284	467
10	106	2.0	94	148	242
total	5203	100	2738	6270	9008

Table 1: In this table the cumulative prominence judgments over all 10 listeners are shown. For example the number 266 in the second column means that this is the number of times that 5 of the 10 listeners judge a given word as prominent. Furthermore the numbers of syllables with and without lexical stress are shown.

The acoustical features are measured on syllables and on each vowel of that syllable, so the prominence values must be assigned to the syllables. (For more details see section 4.) The resulting numbers of syllables specified for lexical stress are also shown in table 1. Lexical stress is defined as primary stress on content words (as looked up in the standard pronunciation lexicon (CELEX)) and no-lexical stress implies non-primary stress including no stress at all. In the set of 2631 words, which are never judged prominent, only 516 of the 3101 syllables are lexically stressed. The relative low number of syllables in this set of words (3101 syllables versus 2631 words) shows that most of these words are monosyllabic.

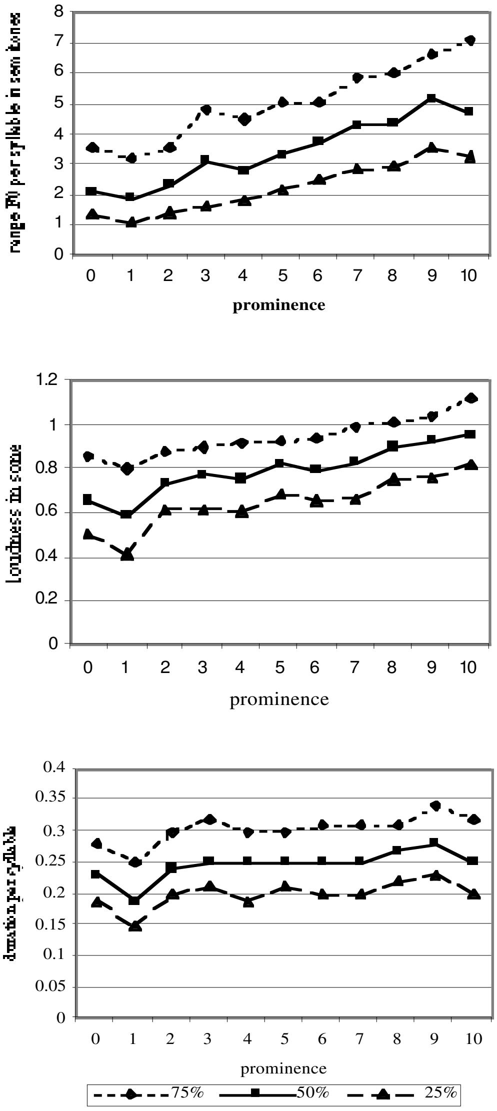
4. PREPROCESSING AND ACOUSTICAL MEASUREMENTS

Before the acoustical features can be measured, the phoneme and syllable boundaries of each sentence must be determined. Because the transliteration of each sentence was available it was possible to look up each word in a standard pronunciation lexicon (CELEX). For each sentence an array of all phonemes that occur in that sentence was used to train an HMM-model on a subset of 4553 sentences from 978 different speakers (this are not round numbers because 447 sentences were excluded because of bad quality). The trained HMM-model was used to find the boundaries of each phoneme in our 500 spoken sentences. Sonorant-rules say that each syllable consists of one vowel and that the consonants following that vowel are ordered with decreasing sonority. The farther a consonant stands away from the vowels the lesser the sonority. These sonorant-rules were implemented in a program to mark the syllable boundaries. Because there were words which did not behave according to these rules, the syllable boundaries were also compared with the boundaries in the CELEX lexicon and hand corrected. With the help of the phoneme label files a syllable label file with syllable boundaries was created. Since we used a lexicon the lexically stressed syllables were also known, and for the content words these lexically stressed syllables were marked and added to the label file. A next and final step in preprocessing the sentences was to connect the cumulative prominence judgments of the 10 listeners with the phoneme and syllable labeling. In summary the identity and boundaries of the phonemes, the syllables with lexical stress markers on content words and boundaries of the syllables, as well as the prominence labels were available for further acoustical analyses.

As a first step we decided to measure the following acoustical features.

- F_0 range per syllable in semitones
- Duration per syllable in seconds
- Loudness of the vowel in sone corrected for the average loudness per sentence

Because the loudness of a vowel is generally responsible for the loudness of the whole syllable, using only the loudness of the vowel works better than that of the whole syllable. The perceived loudness was measured in sone units. This method takes into account the filtering by the basilar membrane by



226 195 244 326 183
516 202 215 260 277 94

Figure 1: The prominence labels and the median, 25 and 75 percentiles of the range F_0, loudness and duration are plotted in these graphs. The loudness measurements were corrected for the averages per sentence. Only the data of the lexically stressed syllables are presented. The numbers (N) of syllables or vowels on which the median and the percentiles are calculated are given in the bottom line.

using a frequency function expressed in Bark units. Loudness per vowel is not corrected for the intrinsic loudness as done in Kießling [7], this must be a next step in further research. The F_0 range is measured per syllable. In future we also intend to use relative acoustic features, by comparison with neighbouring syllables [7]. More specifically we could compare the F_0 range with the adjacent syllables and calculate the ratio as done in the research of Wightman and Ostendorf [8]. As a third feature the syllable rather than the vowel duration is taken, since no effect was found for vowel duration corrected for the intrinsic duration of each vowel type.

4.1. Prominence and Acoustical Features

In order to see the relation between the prominence judgments of the listeners and the acoustical measurements, three graphs are presented in figure 1. Because the major effect of prominence is in the lexically stressed syllables (see table 2) only those data are presented in the graphs in figure 1. The three graphs show the prominence labels (0-10) and the median, the 25 percentile and the 75 percentile values of the acoustical features, namely F_0 range, duration, and loudness. The perceived prominence versus the F_0 range per syllable and the loudness per vowel show a higher correlation than the perceived prominence versus the duration of the syllable (see table 2). This is not surprising if one realizes that final lengthening and speaking rate also influence duration. It must be mentioned that Portele and Heuft [1] found a stronger effect of syllable duration versus prominence, but their speech material is only from 3 speakers.

To test if there is a linear relation between various acoustical measurements and perceived prominence, Spearman's correlation coefficients were calculated not just for the stressed but also for the unstressed condition. The resulting correlations are presented in table 2. Only for the loudness in unstressed condition no significant correlation could be found.

The significant correlations show that there is a positive linear relation between prominence as marked by listeners and loudness per vowel, F_0 range and duration per syllable. In case of duration of lexically stressed syllables the relation is not as strong as in case of loudness and F_0 range.

	Prominence		
Lexical stress	**Yes**	**No**	**All**
F_0 range	0.389	0.124	0.245
Loudness	0.317	0.017	0.151
Duration	0.151	0.159	0.228

Table 2: Spearman's correlation coefficients between prominence and acoustical features are presented in this table. Only for the loudness in unstressed condition there is no significant correlation. For the other acoustical measurements there is a significant correlation at the 0.01 level (2-tailed),

Despite the fact that there is a linear relation the graphs in figure 1 show that the variability cannot be ignored. The difference between the 25 and 75 percentile values is so large that the automatic classification of non-prominent (0) and prominent (8,9,10) will not be an easy task. For example in the upper graph where the F_0 range is plotted, the 25 percentile value at prominence 10 still lies between the median and the 75 percentile values of prominence 0. The same is the case for duration and loudness.

5. PREDICTION OF PROMINENCE

The measurements described above are not only used to analyze various relations, but are also used as training and testing data for predicting prominence. Can a simple net predict prominence and if so to what extent? The prominence can be classified at different information levels. The higher the level the more information is added to the classifier for the prediction of prominence. The first level is to classify only with acoustic information such as intensity, duration and F_0. On the second level the speech signal is divided in meaningful parts and boundaries of the phonemes are added as a feature. On the third level the syllable boundaries are also added. On the fourth level the phoneme identity is added, and on the final level also lexical features such as lexical stress are available for the classification task. Of course it would be ideal for various speech technology applications if one could classify on acoustical information only. In this paper as a first step the prominence is classified with such information as lexical stress, and phoneme and syllable boundaries, but the identity of each vowel is not jet used for classification. We use range of F_0 per syllable in semitones, duration per syllable, and loudness of each vowel corrected for the average loudness per sentence. Additional information such as syllable boundaries, phoneme boundaries and lexical information is not presented to the net as an extra input feature, but incorporated in the acoustical measurements and in the pre-selection of the input features.

As a first step 4 simple artificial neural nets (ANN) without hidden layer were trained to classify prominent and non-prominent syllables. An ANN without hidden layer can be as good as a discriminant analysis. For this classification with neural networks we present some preliminary results. Because of the variability in the speech material the classification was done between non-prominent (0) and prominent (8,9,10) lexically stressed syllables. This resulted in a data set of 516 non-prominent and 554 (277+183+94) prominent lexically stressed syllables, see table 1. An independent test set of 140 prominent and 140 non-prominent lexically stressed syllables was randomly selected. The remaining 376 non-prominent and the 414 prominent syllables were used for training. The percentages correct for the test set are presented in table 3.

ANN without hidden layer

F_0 range	**Loudness**	**Duration**	**Test set**
x	x	x	81.07
x			72.86
	x		70.71
		x	63.21

Table 3: Percentages correct prominence classification of different ANN's with different acoustical input features.

With all 3 acoustical features the recognition rate came up to 81.07 percent correct. The classification with only loudness or F_0 range as input feature reached 72.86 and 70.71 percentage correct, respectively Using the duration as only input feature lowered the recognition rate to 63.21 percentage correct, as expected, because the correlation of prominence and syllable duration was lower than the correlation of F_0 range and loudness.

6. CONCLUDING REMARKS

In conclusion, it can be said that prominence, as defined by naive listeners' judgements, can function as an interface between acoustics and linguistics. As shown in this paper the complex relation between prominence and acoustical correlates can be estimated. It turns out that not only the F_0 range per syllable has a high correlation with prominence (in lexically stressed syllables the correlation is 0.389), but also the loudness per vowel (in lexically stressed syllables the correlation is 0.317). In case of the syllable duration the relation towards prominence is not that strong. Not surprisingly, the automatic classification of non-prominent (0) and prominent lexically stressed syllables (8,9,10) with only syllable duration as input feature is not as good as the automatic classification with only F_0 range or loudness. The low correlation between syllable duration and prominence and the corresponding low recognition rate can be explained by the fact that duration is also influenced for example by speaking rate, by final lengthening and by the intrinsic duration of each phoneme. A combination of the three features leads to a recognition rate of 81,07% correct. We will study whether normalizations at these levels will be possible. Furthermore it is worth mentioning again that the speech material is complex, because of high speaker variability. However, this high speaker variability is the reality for most speech technology applications. For a further automatic classification a thorough analysis and more data are needed.

7. REFERENCES

1. Portele, T., and Heuft, B., "Towards a prominence-based synthesis system", *Speech Communication,* 21: 61-71, 1997.

2. Mayer, J., *Intonation und Bedeutung*, Ph.D. Thesis, Phonetik AIMS, Stuttgart, Vol. 3, 1997.

3. Ladd, D. J., *Intonational Phonology,* Cambridge University Press, 1996.

4. Terken, J., "Fundamental frequency and perceived prominence of accented syllables", *Journal of the Acoustical Society of America* 89: 1768-1776, 1991.

5. Damhuis, M., Boogaart, T., in 't Veld, C., Versteijlen, M., Schelvis, W., Bos, L., Boves, L.,"Creation and analysis of the Dutch Polyphone corpus", *Proceedings ICSLP 94* Yokohama, 1803 –1806, 1994.

6. Streefkerk, B.M. and Pols, L.C.W., "Prominence in read aloud Dutch sentences as marked by naive listeners" *Tagungsband KONVENS-98,* Frankfurt a.M., 201-205, 1998.

7. Kießling, A., *Extraktion und Klassifikation prosodischer Merkmale in der automatischen*

Sprachverarbeitung, Berichte aus der Informatik, Shaker Verlag, Aachen, 1996.

8. Wightman, C.W. and Ostendorf, M., "Automatic labeling of prosodic patterns", *IEEE Transactions on Speech and Audio Processing*, 2: 469-481, 1994.