ACOUSTICAL FEATURES AS PREDICTORS FOR PROMINENCE IN READ ALOUD DUTCH SENTENCES USED IN ANN'S

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

*Institute of Phonetic Sciences Amsterdam (IFA) / IFOTT

**Lernout & Hauspie Speech Products N. V., Belgium

ABSTRACT

In this paper we present several acoustical features, which are used as predictors for prominence. A set of 1244 sentences from 273 different speakers is selected from the Dutch Polyphone Corpus. Via listening experiments the subjective prominence markers are obtained. Several acoustical features concerning F_0, energy and duration are derived and used as predictors for prominence. The sentences are divided in a test and a training set, to test and train neural networks with different topologies and different input features. The first results show that a classification of prominent and non-prominent words is possible with 82.1% correct for an independent test set.

1. INTRODUCTION

Knowing the relevant features for perceived prominence can be useful in several speech technology applications. For example in speech synthesis, where words can be realized with an accent-lending pitch movement, it should be the next step to introduce different degrees of prominence. Knowing more about the realization of various degrees of prominence can be helpful to realize more natural speech. Also in speech recognition the optimal features of prominence can help to disambiguate sentences.

The concept of prominence is not very clearly defined, but it is obvious that it refers to those words or syllables, which are perceived as standing out from their environment [14]. Or to say it in other words, it refers to the perceptual salience of a unit. In the research of Fant and Kruckenberg [4] and Portele and Heuft [9] prominence is defined via detailed listening: each syllable is marked with a degree of prominence on a scale from 0 to 30. So the listener must distinguish 31 degrees of prominence, which is a very difficult, and maybe even impossible task. We also define prominence via a listening experiment, but just by asking naive listeners to mark the emphasized words. The sum of all marks per word results in a prominence scale, where words with a low degree of prominence could perhaps be interpreted as words with realized lexical stress and words with a high degree of prominence be interpreted as carrying sentence accent.

In our opinion the acoustical features for prominence will not be limited to the pitch movements only, but will also include aspects of intensity, duration and spectral quality. The prominent words coincide with F_0 changes, and they are longer and louder than their non-prominent counterparts. F_0 change is a very important feature to detect the most prominent words, and for this measure most of the time speaker or context normalization is not necessary. Although some effect of vowel identity may exist, F_0 change expressed in semitones is mostly independent of intrinsic properties. However, in the case of duration for example the vowel identity and final lengthening are properties that substantially influence the actual duration of the vowel and the syllable.

Research of Sluijter [10] shows that duration and energy are mainly features of realized word stress and that features concerning F_0 are used for pitch accent. In their research it turns out that also spectral tilt is a strong correlate of realized word stress. In contrast, van Kuijk's work [8], which concerns acoustic differences between lexically stressed and unstressed vowels, shows that spectral tilt is not a useful feature.

Much is known about pitch movements for the use of speech synthesis. For example with the IPO intonation system [5] one can evoke a very good pitch accent for one speaker. However, the extraction in natural speech of useful features concerning the pitch movement is a problem. In present-day speech recognition the problem to recognize accents is an unsolved problem, yet. A first good attempt is made by Kompe [7] and by Wightman and Ostendorf [15]. However, these approaches are using as many features as possible to recognize sentence accent, not yet leading to any specific knowledge about which are the most optimal features.

2. SPEECH MATERIAL

The speech material we used was selected from the Dutch Polyphone corpus. This corpus contains 12500 different sentences, each sentence was used twice, so 5 different sentences from 5000 different speakers resulting in 25000 sentences are recorded over the telephone. All sentences were taken from Dutch newspapers and were constructed in such a way that each set contains all phonemes of the Dutch language at least once. The speakers were instructed to read the sentences aloud from paper via the telephone (for more details see [3]). Since the sentences were isolated and read aloud without context, focus could not be determined anymore. This speech material contains speech of a lot of speakers from different regions, different ages and different socio-economic statuses and therefore the variability of this speech material is huge. On the one hand this makes analyses very complex, but on the other hand this only reflects the actual situation in several speech technology applications, particularly in speech recognition. The acoustical analyses, as well as the prediction of prominence based on the acoustical features, are done on a subset of 1244 sentences from this Polyphone corpus. This subset contains speech of 273 different speakers, with up to 5 different sentences spoken by one speaker. Only few sentences are spoken more than once, by different speakers. This subset contains 1196 different sentences, implying that 48 sentences are spoken by more than one speaker. The total set consists of 13092 words, resulting on average in 10.5 words per sentence. In terms of syllables the set consists of 22496 syllables, which is on average 18.1 syllables per sentence.

3. PRE-PROCESSING

Before the 1244 sentences can be analysed in more detail, the prominence labels must be derived from a listening experiment, which is described in more detail in section 4. In order to derive various acoustical features we first of all have to determine the phoneme and syllable boundaries within each sentence. Because the transliteration of each sentence was available it was possible to look up most words in a standard pronunciation lexicon (CELEX) [1]. For each sentence an array of all phonemes that occur in that sentence was used to train an HMM-model, with the help of my colleague Xue Wang [16], on a different subset of 4553 sentences from 978 different speakers also taken from the Dutch Polyphone corpus (these are not round numbers because 447 sentences were excluded due to bad quality). The trained HMM-model was used to find the boundaries of each phoneme in our 1244 spoken sentences. The syllable boundaries were derived from the CELEX lexicon, but since this lexicon did not contain all words, some syllable boundaries had to be added by hand. With the help of the phoneme label file, a syllable label file with syllable boundaries was created. Since we used a lexicon, the lexically stressed syllables of most content words were also known, the lexically stressed syllables of non-content words were also added by hand. A next and final step in pre-processing the sentences consisted of connecting the cumulative prominence markers of the 10 listeners with the phoneme and syllable labelling. In summary the identity and boundaries of the phonemes, the syllables with lexical stress markers, and the boundaries of the syllables, as well as the prominence labels were available for further acoustical analyses.

4. PROMINENCE MARKING

In order to derive the prominence judgements, two listening experiments were constructed one with a subset of 500 phonetically rich sentences spoken by 100 speakers and one with another subset of 744 sentences spoken by 173 speakers. To test how consistent the listeners were we presented the first 50 sentences of the first subset to each listener twice (for more detail see [12]). The 550 sentences (500 + 50), of the first listening experiment as well as the 744 sentences of the second listening were presented in 4 sessions. For both listening experiments the sentences are presented to a partly overlapping group of 10 listeners. The sentences are presented individually in random order to compensate for possible learning effects. In both listening experiments the 10 listeners were students from the Humanities Faculty of the University of Amsterdam. The experiments were constructed in such a way that the written words of each sentence were displayed on the monitor with a button underneath each word. The listeners could click on the buttons corresponding to words perceived as being spoken with emphasis.

In table 1 the results of both listening experiments combined are presented. In the second column the absolute, and in the third column the relative, numbers of the cumulative prominence judgements of all 1244 sentences are given. In case of the first listening experiment (with 500 sentences) the second prominence judgements of the 50 sentences which are presented twice are included in this table, because in the first 50 sentences a learning effect may still prevail. The variability of the listeners is described in more detail in [12]. It is worth pointing out that about half of the words (45.45%) are never marked as prominent, and that 2059 (802 + 739 + 518) words, which is 15.73% of all words, are marked as prominent by 8 or more of the listeners. Because there are on average 10.5 words per sentence this results in 1.49 prominent words per sentence.

The acquired prominence markers can be used in several ways. First, the prominent and the non-prominent words can be selected: the non-prominent words are the words which are never marked as prominent, and the prominent words are the words which are judged by 8 or more of the listeners. The prominence judgements can also interpreted on a prominence scale from 0 to 10. Since there is variability between listeners, it cannot be expected that the scale from 0 to 10 represent 11 separate classes. From a lexical analysis [6] of a subset of these sentences it can be expected that 4 classes might be a good compromise.

Words with a low degree of prominence could perhaps be interpreted as words with realized lexical stress. The question then is, do words with prominence judgements at the word level receive also markers when the task implies marking the prominent *syllables*. Such a comparison is presented in Streefkerk et. al [11]. It is indeed true that if there is a marker at the word level this always coincides with a marker at the syllable level.

Value	Freq. words	Percentage
0	5950	45.45
1	1161	8.87
2	679	5.19
3	595	4.54
4	586	4.48
5	603	4.61
6	645	4.93
7	813	6.21
8	802	6.13
9	739	5.64
10	518	3.96
Total	13092	100

Table 1: In this table the combined cumulative prominence judgements of both listening experiments are shown. For example in column 'Freq. words' the number 802 means that so many words have been marked by 8 listeners as being prominent.

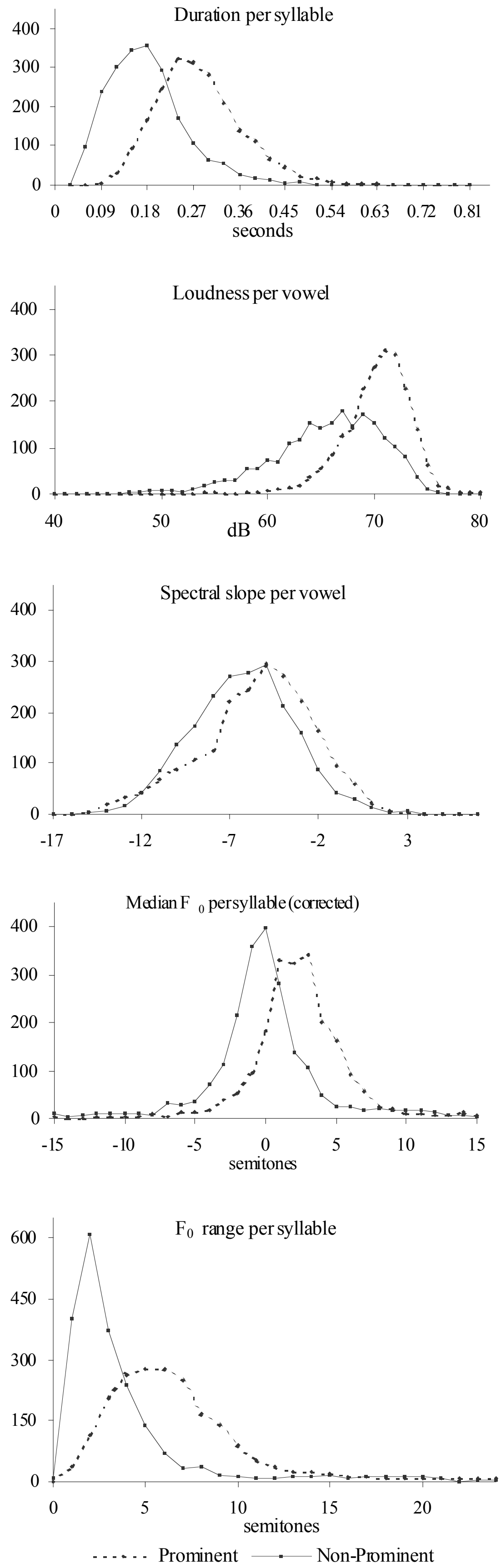

Figure 1: Distributions of prominent and non-prominent syllables.

5. ACOUSTICAL FEATURES

The set of 1244 sentences is involved in various acoustical analyses as well as in prominence classifications of the neural networks. As described in the pre-processing section the prominence labels of each word are connected to the speech material, as well as to the phoneme and syllable boundaries. As a first step towards finding optimal features for the prediction of prominence, the following raw features are derived:

- Duration per syllable in seconds
- Intensity per vowel in dB
- Spectral slope per vowel
- Median F_0 per syllable in semitones
- F_0 range per syllable in semitones

Since the phonemes of each sentence are automatically segmented, the duration of each segment is determined by the HMM-recogniser. The minimum duration of each vowel is 30 ms, because accuracy of each frame was fixed to 10 ms and minimally three states per vowel must be visited. The intensity, spectral slope, and F_0 measurements were performed by using the software package PRAAT [2]. The intensity was calculated for each vowel. The spectral slope of each vowel was estimated as the slope of the regression line of lower and higher octave bands of a long-term average spectrum. In case of the features concerning pitch, first of all the F_0-contour of the whole sentence was determined and automatically smoothed. On the resulting pitch curve the F_0 range per syllable, the median F_0-value per syllable and the median F_0-value of the whole sentence was measured.

All features, except median F_0, are raw features. So in this paper the duration per syllable is not yet corrected for intrinsic vowel length or final lengthening. Also the local speaking rate is not yet taken into account. In case of intensity, this is not yet corrected for the different vowel strengths. Since the distribution of the intensity of all vowels corrected for the overall intensity per sentence shows no noticeable difference with the non-corrected distribution, only the uncorrected data are presented in this paper. However, concerning the median F_0-value per syllable, the overall median F_0-value of each sentence is subtracted. Because otherwise all syllables coming from a sentence spoken by a female, would be classified as prominent, since women have a higher overall voice.

In figure 1 the histograms of the acoustical features of a subset of the non-prominent (no prominent marks) and the prominent (8 or more prominent marks) words are plotted. In the case of more-than-one-syllable-words these distributions show the values of the lexically stressed syllable. The distributions of the prominent syllables are plotted in the graphs as *dotted lines*. An equal number of non-prominent words are randomly selected, and the distributions of the acoustical features for these words are plotted as *continuous lines* in the graphs. For all acoustical features it is true that the prominent syllables show distributions which are more shifted to the right. So the tendency is indeed that prominent syllables are longer, are realized with more intensity, have a slightly flatter spectral slope, have a higher median F_0, and are realized with a greater range in F_0 than their non-prominent counterparts. In Streefkerk et al [13] it is shown that there is a positive relation between the prominence scale from 0 to 10 and the acoustical features duration and intensity per vowel, and the F_0 range values of syllables. From the distributions presented in figure 1 it can be deducted that the spectral slope is the worst predictor for prominence, because there is a substantial area of overlap between the prominent and the non-prominent distributions.

6. CLASSIFICATION WITH FEED FOREWARD NETWORKS

The raw acoustical features, described in section 4 are not only used to analyse various distributions concerning prominent and non-prominent words, but are also used as input features for several ANN's (artificial neural network) to predict prominence. As a first step features belonging to the prominent words and to the same amount of randomly selected non-prominent words are taken from the 1244 sentences, and divided in a test and atraining sets. To get to know more about the most optimal topology we need, and about the input features which are most optimal for the prediction of prominence, we trained and tested all input combinations and all numbers of nodes in the hidden layer from 3 to 6. In total 31 combinations of input features and 124 feed forward nets were trained with 125 iterations for those nets with 3 and 4 nodes and with 200 iterations with 5 and 6 nodes in the hidden layer. The best result (84.2% correct on the training set and 82.1% correct for the independent test set) was reached with all input features, duration, F_0 range, and median F_0 per syllable, and intensity and spectral slope per vowel. Leaving out spectral slope as input feature, lowered the percentage correct classification not substantially (83.5% correct classification for the training set and 81.8% for the test set). So in further research we will not concentrate on the spectral slope of each vowel.

At the conference more data will be presented also on different output units, namely a classification in terms of the prominence scale from 0 to 10, as well as the re-scaled prominence values from 0 to 3.

7. CONCLUSION

Concluding it can be said that duration per syllable, intensity per vowel, median F_0 per syllable corrected for the median F_0 per sentence, and F_0 range per syllable are promising features as predictors for prominence. The spectral slope per vowel thus not seem to be a predictive feature for prominence. The same was found in the research of van Kuijk and Boves [8]. The plotted distributions of spectral slope for prominent and non-prominent words confirm the bad distinctive quality of this feature. A combination of all 5 features leads to a recognition rate of 84.2% correct for the test set and 82.1% for the training set. We will study in future research if possible normalisation of duration and intensity leads to higher recognition rates. Also F_0 features relative to their environment will be tested in future, and more sophisticated ANN's will be trained.

8. REFERENCES

[1] Baayen, R. H., Piepenbock, R. and van Rijn, H., 1993. The CELEX lexical database (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

[2] Boersma, P. and Weenink, D., 1996. Praat, a system for doing phonetics by the computer. Version 3.4., *Report 132 of the Institute of Phonetic Sciences Amsterdam*, http://fonsg3.hum.uva.nl/praat/.

[3] Damhuis, M., Boogaart, T., in 't Veld, C., Versteijlen, M., Schelvis, W., Bos, L. and Boves, L., 1994. Creation and analysis of the Dutch Polyphone corpus. *Proc. ICSLP-94*, Yokohama, 1803- 1806.

[4] Fant G. and Kruckenberg A., 1989. Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR* 2, 1-83.

[5] 't Hart J., Collier R. and Cohen A., 1990. *A perceptual study of intonation*. Cambridge, University Press.

[6] Helsloot K. and Streefkerk B. M., 1998. Perceived prominence and the metrical-prosodic structure of Dutch sentences. *Proc. of the Institute of Phonetic Sciences of the University of Amsterdam*, 22, 97-111.

[7] Kompe, R., 1997. *Prosody in Speech Understanding Systems*, Lecture Notes for Artificial Intelligence. Springer-Verlag, Berlin.

[8] Kuijk, van D. and Boves, L. 1999. Acoustic characteristics of lexical stress in continuous telephone speech. *Speech Communication,* 27, 95-111.

[9] Portele, T. and Heuft, B., 1997. Towards a prominence-based synthesis system. *Speech Communication*, 21, 61-71,

[10] Sluijter, A. M. C. and Heuven van , V. J., 1996. Spectral balance as an acoustic correlate of linguistic stress. *J. Acoust. Soc. Amer*. 100, 2471-2485.

[11] Streefkerk, B. M., Pols L.C.W. and ten Bosch L. F. M., 1997. Prominence in read aloud sentences, as marked by listeners and classified automatically. *Proc. of the Institute of Phonetic Sciences of the University of Amsterdam*, 21, 101-116.

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

[13] Streefkerk, B. M., Pols, L. C. W. and ten Bosch, L. F. M., 1998. Automatic detection of Prominence (as defined by listeners' judgments) in read aloud Dutch sentences. *Proc. of ICSLP-98*, Sydney, Vol. 3, 683-687.

[14] Terken, J., 1991. Fundamental frequency and perceived prominence of accented syllables. *J. Acoust. Soc. Am.*, 89, 1768-1776.

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

[16] Wang, X., 1997. Incorporating knowledge on segmental duration in HMM-based continuous speech recognition, Ph.D. Thesis, University of Amsterdam.