TOWARDS FINDING OPTIMAL FEATURES OF PERCEIVED PROMINENCE

Barbertje M. Streefkerk[*], Louis C.W. Pols[*] & Louis ten Bosch[**]
[*]*Institute of Phonetic Sciences/IFOTT, University of Amsterdam, The Netherlands*
[**]*Lernout & Hauspie Speech Products N.V., Belgium*
barber@fon.hum.uva.nl

ABSTRACT

In this paper we present acoustical as well as lexical features for classification purposes of perceived prominence in read aloud Dutch sentences. Via a perception experiment with 10 naive listeners we derived prominence labels at the word level for 500 sentences. Part of these sentences are used for lexical/syntactical analyses. It turns out that most of the function words are never perceived as prominent, and that specific content words namely adverbs, nouns and adjectives are almost always perceived with some degree of prominence, whereas verbs form a middle class. So we decided to concentrate on the lexically stressed syllables of content words, because these are the words whose prominence is not uniquely classified by their lexical class.

In this paper we use F_0 range per syllable, both 'raw' and corrected for the declination line, to distinguish between the most prominent and non-prominent content words, although intensity and duration features can be used as additional features to improve the classification. As an initial result we can conclude that F_0 range is a very good feature to distinguish between prominent and non-prominent content words.

1. INTRODUCTION

Knowing the relevant features for perceived prominence can be very useful in several speech technology applications, such as speech synthesis, where nowadays accented words are realized with an accent-lending pitch movement. Much is known about pitch movements for the use of speech synthesis, see for instance the IPO intonation grammar [2]. The next step is to introduce different degrees of prominence. Knowing more about the realization of different degrees of prominence can be helpful to generate more natural speech. This prominence based approach turns out to be a useful interface between acoustics and linguistics [5]. In speech recognition the detection of prominence can help to disambiguate sentences, which is especially important in dialog systems. In case of speech recognition the problem to recognize accented words is an unsolved problem, yet. A first attempt is performed in the research of Kompe [4] and Wightman and Ostendorf [9].

The concept of prominence is not very clearly defined, but it is obvious that it refers to words or syllables, which are perceived as standing out from their environment [8]. Therefore we define word prominence in our training data via a perception experiment, where naive listeners had to mark those words perceived as being spoken with emphasis.

The optimal acoustical features to detect prominence for recognition purposes may not be limited to the pitch movements, but probably include also aspects of intensity, spectral slope, and duration [7]. However, F_0 seems to be a very important feature to detect the most prominent words. An advantage of F_0 is that it is more independent for intrinsic properties of speech sounds than duration, but the extraction of F_0 features from the pitch movements is still a problem. Therefore, we concentrate on measurements within the syllable. As for the duration, for example, the vowel identity and final lengthening are properties that substantially influence the actual duration of vowels and of syllables. This makes it especially difficult to extract optimal features for prominence detection.

In speech recognition results of prominence detection depend on what kind of lexical information is available. Of course it would be ideal for various speech technology applications if one could classify on acoustical information only, but this would be a very complicated task. In this paper we use all kinds of information up to lexical information. Also an analysis of lexical word class (such as nouns, verbs or function words) related to degree of prominence is presented.

2. GETTING PROMINENCE LABELS

2.1. Speech Material

The speech material was selected from the Dutch Polyphone corpus. This corpus contains 5 sentences from 5000 different speakers resulting in 12500 mostly different sentences recorded over the telephone. All sentences were selected from a Dutch newspaper 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 [1]). 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 this only reflects the situation in several speech technology applications, particularly in speech recognition.

2.2 Listening Experiment

In order to derive the prominence judgments, 500 phonetically rich sentences spoken by 50 male and 50 female speakers were selected for a listening experiment. To test how consistent the listeners were we presented the first 50 sentences to each listener twice. The 550 sentences (500 + 50) were presented in 4 random order sessions, which differed per listener, to compensate for possible learning effects. The first two sessions

contained 150, and the last two sessions contained 125 sentences. All 10 listeners were students from the Humanities Faculty of the University of Amsterdam. The written words of each sentence were displayed on the monitor with a button underneath each word. The subjects could click on the buttons corresponding to words perceived as being spoken with emphasis.

2.3. Resulting Labels from the Listening Experiment

In table 1 the absolute and relative judgments for all 500 sentences 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 a learning effect may still prevail. The variability of the listeners is described in more detail in [6]. 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. The words that are judged as prominent by 80% or more of the listeners are called the prominent words, and the words that are never judged as prominent are called the non-prominent words. The prominence judgments can also be seen as a scale between '0' (non-prominent) and '10' (very prominent). 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: The cumulative prominence judgments over all 10 listeners. 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. The number of syllables with and without lexical stress are shown on the right-hand side.

The acoustical features are measured on syllables, so the word prominence labels must be assigned to syllables as well. The resulting number of syllables specified for lexical stress are also shown in table 1. Lexical stress is defined as primary stress in content words (as looked up in the standard pronunciation lexicon CELEX). No lexical stress implies non-primary stress including no stress at all. In the set of 2631 words which are never judged as prominent (non-prominent), only 516 of the 3101 syllables are lexically stressed. The relatively low number of syllables in this set of words (3101 syllables versus 2631 words) shows that most of these words are monosyllabic. The number of lexically stressed prominent syllables is 554 (277+183+94). So, according to majority judgments, there are 554 prominent syllables versus 516 non-prominent syllables in 2738 stressed syllables.

3. LEXICAL FEATURES

To have a closer look at the relation between lexical properties and prominence judgments of the listeners, 50 sentences, which form a subset of the 500 sentences, are analyzed in more detail. These are the sentences, which are judged twice, resulting in a 'deviant' prominence scale ranging from '0' to '20'. As shown in table 2, these 50 newspaper sentences consist, on average, of 10.38 words per sentence, and the average number of syllables per sentence is 18.48. About half of the words are function words and half of the words are content words. There are 8 words, which are not yet classified as being a function or a content word. It is to be expected that the function words are perceived as less prominent than the content words.

	Total number	Mean per sentence
Words	519	10.38
Content words	278	5.56
Function words	233	4.66
Rest words	8	-
Syllables	924	18.48

Table 2: Overview of various numbers of word types and syllables for this selection of 50 sentences.

In table 3 the number of words for which the cumulative prominence judgments is greater than 10, smaller than or equal to 10, and 0, are given. The mean number of words per sentence which are never marked as prominent, is 4.88; this is almost equal to the amount of function words per sentence. If we sum the mean number of words, which are ever marked as prominent, these are 5.5 words per sentence. This is also almost equal to the amount of content words per sentence.

Prominence	Total number	Mean per sentence	
11-20	136	2.72	} 5.5
1-10	139	2.78	
0	244	4.88	
Total	519	10.38	

Table 3: Distributions of the prominence degrees on a scale of 0-20, (more than 10, less than 10 or equal to 10, and 0). The mean number of words with prominence degree (11-20), prominence degree (1-10), and prominence degree (0) per sentences are also given.

In table 4 the perceived prominence judgments are analyzed with respect to monosyllabic and polysyllabic function words. Of the 233 function words 216 are monosyllabic and 17 polysyllabic. The listeners perceived 205 monosyllabic words as bearing no prominence at all, and the other 11 are perceived with a very low degree of prominence (see table 4). Of the polysyllabic function words 12 are perceived as bearing no prominence, and 4 are perceived with a greater degree of

prominence (>5). In summery we can say that our data once again confirm that function words correspond to less/non prominent words.

Prominence	**Function words**			
	Monosyllabic		Polysyllabic	
0	205	95 %	12	70 %
1-5	11	5 %	1	6 %
6-20	-	-	4	24 %
Total	216	100 %	17	100 %

Table 4: Absolute number and percentage of mono- and polysyllabic words regarding prominence degrees on a scale of 0-20.

For the content words we count how many nouns, verbs, adverbs and adjectives are never marked as prominent or have some prominence degree. Indeed, listeners perceived prominence on many nouns, verbs, adjectives and adverbs (see table 5). Adverbs and adjectives are always perceived with prominence degree (1-5), and often as very prominent (6-20). In general, the nouns are also perceived with a degree of prominence, except to of them. The verbs are not as frequently perceived as prominent, only in 19 out of 50 cases, which is 38%. For more detail see [3].

Promi-	**Content words**							
nence	Nouns		Verbs		Adverbs		Adjectives	
0	2	1.3 %	19	38 %	-		-	-
1-5	34	23.7%	33	29 %	11	33 %	7	15 %
6-20	107	75 %	17	33 %	22	67 %	39	85 %
Total	143	100 %	50	100 %	33	100 %	46	100 %

Table 5: Absolute number and percentage of content words regarding prominent degrees on a scale of 0-20.

The distribution of prominence and lexical class of a word can be used for prominence prediction. Very generally, we can say that function words have a low chance to be perceived as prominent, verbs form a middle category with about 38% chance of bearing no prominence, and nouns, adverbs and adjectives have a very high chance of being perceived as prominent. The use of lexical features for the prediction of prominence assumes that lexical information is available for the recognition task. This information can be extracted from a lexicon, but in some cases a syntactical analysis is necessary. In order to find optimal features for the prediction of prominence it still has to be tested how far we can get when word class is assigned automatically with the help of a lexicon and when this information is used as input feature for a classifier.

4. ACOUSTICAL FEATURES CONCERNING F_0

500 sentences from the Dutch Polyphone Corpus, labeled for prominence on a scale between 0-10 have been used in the acoustical measurements concerning F_0, duration and intensity [7]. Because of the above-presented lexical analysis of a subset of these sentences, we concentrate in this paper on the content words only. Because we have word prominence judgments and because we want to measure on syllables we select the lexically stressed syllables of the content words. But before the acoustical features can be measured, the phoneme and syllable boundaries within 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 also 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 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 vowel 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.

It is obvious that the pitch movement plays a very important role in the perception of very prominent words in the sentence. The notion of pitch accent as formulated in the IPO intonation system [2], is exclusively based on pitch movements. However, the implementation of pitch accents in speech synthesis systems has sometimes consequences for duration and intensity and spectral quality. We belief that pitch accents, intensity, spectral slope, and duration are important for prominence, but in this paper we concentrate on the extraction of F_0 features: namely, on the F_0 range of each syllable. In order to extract features independently from gender, we concentrate on *differences in semitones* (for example the difference between the minimum and the maximum measurements within one syllable), and not on absolute measurements. Moreover, we try to extract features with respect to the local syllable environment. As a first step we corrected for the declination line [2]. We used the following method: a regression line per sentence was estimated on the automatically smoothed pitch curve, and subtracted. On this resulting pitch curve the range per syllable was measured. The means and standard deviations of the prominent (8, 9, 10) and non-prominent (0) lexically stressed syllables of content words are given in table 8. In figure 1 the distribution for the lexically stressed syllables of content words both for prominent and non-prominent ones is plotted as a histogram. Figure 2 presents the same F_0 range data as in figure 1, but this time corrected for the

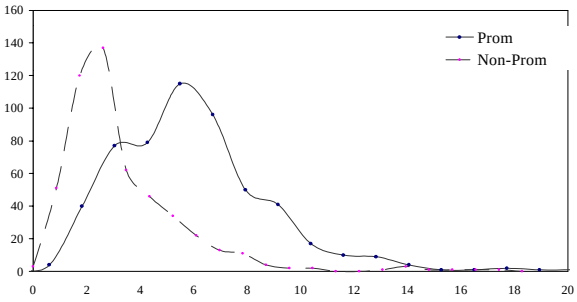

Figure 1: Histogram of the F_0 range in semitones of the lexically stressed syllable for prominent (8,9,10) and non-prominent (0) content words.

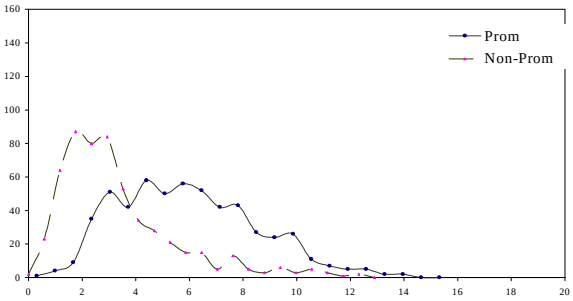

Figure 2: Histogram of the F_0 range in semitones corrected for the declination line of the lexically stressed syllable for prominent (8,9,10) and non-prominent (0) content words.

	F_0 Range	
	Prominent	Non-Prominent
Mean	5.32	2.94
Std.	3.13	2.44
	F_0 Range corrected for declination line	
	Prominent	Non-Prominent
Mean	5.65	2.90
Std.	2.65	1.90
Number	654	516

Table 8: Means and standard deviations of the F_0 range per syllable for the prominent and the non-prominent lexically stressed content words.

overall declination. It can be observed that the prominent and non-prominent words can partly be separated on the base of the F_0 range, but at first sight there is no difference between the corrected and the non-corrected data. A discriminant analysis confirms this: both ranges can discriminate between prominent and non-prominent words with a percentage correct of 72%. However, the Spearman correlation coefficient between the prominence scale (0-10), (so the words with a prominence label between 1-7 are also involved) and the acoustical features for the corrected F_0 range is 0.412 and for the non-corrected it is 0.389. This may indicate that the corrected F_0 ranges are more sensitive for smaller movements, which are perceived with a lower degree of prominence [8]. The classification of function words with the help of acoustical features such as F_0 range should be included in following analyses, in order to see how good this distrimination is.

5. CONCLUSIONS

Concluding we can say that there is a relation between word class and prominence. The information of the word class can be used as a feature to detect very prominent and non-prominent words in a sentence. As a result we can say that most of the function words are never perceived as prominent by any of the listeners, and that content words namely adverbs, nouns and adjectives are almost always marked with some degree of prominence, whereas verbs form a middle class. The role of word class features for distinguishing non-prominent and non-prominent words must be tested in future reseach.

Furthermore, the F_0 range per syllable measured in semitones is a useful feature to classify prominent and non-prominent content words. The distinctions between the lexically stressed syllables of the prominent and the non-prominent content words (72% correct with a discriminant analysis) are promising for further analysis. We did not find a difference between the declination-corrected and the non-corrected features. In future research we will also concentrate on additional features such as measurements concerning duration and intensity, and on classification with more sophisticated classifiers, most probably in the form of a feed forward neural network.

ACKNOWLEDGEMENT

We thank Xue Wang for his help with the automatic segmentation with a HMM recognizer, and especially Karijn Helsloot for her huge involvement with the lexical syntactical analyses.

REFERENCES

[1] 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.s ICSLP 94*, Yokohama, 1803- 1806.

[2] ‘t Hart J., Collier R. and Cohen A. 1990. *A perceptual study of intonation*. Cambridge, University Press.

[3] 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.

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

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

[6] 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.

[7] 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.

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

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