Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 21 (1997), 101-116

PROMINENCE IN READ ALOUD SENTENCES, AS MARKED BY LISTENERS AND CLASSIFIED AUTOMATICALLY

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

Abstract

Two perception experiments with different instructions and different presentations were used to locate prominence in 81 read aloud sentences. Results show that, depending on the instruction and presentation (mark prominent words or prominent syllables), the subjects can listen more analytically or more globally. The results indicate that a word perception experiment is a good method to detect prominence at sentence level, coming closer to sentence accent than in a syllable perception experiment. Another pilot perception experiment was run to investigate prominence marking in monotonized sentences. This pilot experiment shows that listeners are indeed able to mark prominent words even in sentences with monotone pitch. Furthermore various acoustical measurements were done both on the most prominent as well as on the non-prominent words. With the help of an artificial neural network, prominence is classified on the base of acoustical information only. The initial results of this classification are promising.

1. Introduction

The automatic classification of accented and non-accented words in speech is a major question in recent research (e.g., Bagshaw, 1993; Kießling, 1996; Kompe, 1997; Storm, 1995; Taylor, 1993; Ten Bosch, 1993; Wightman and Ostendorf, 1994). There are different approaches to label the speech required for training and testing automatic classification. In the research of Kießling and Kompe the initial labeling of accented and non-accented words was also done automatically based only on linguistic, semantic and phonologic information. With this initial labeling, a classifier that used acoustical information was trained and tested. A disadvantage of this approach is that the labeled accented and non-accented words are not necessarily realized as such. Ten Bosch and Taylor labeled the pitch contour according to the IPO intonation grammar ('t Hart et al., 1990) or the Rise/Fall/Connection model (Taylor, 1992), respectively. In the research of Storm, the speech material was labeled according to TOBI (Silverman et al., 1992). In the research of Ten Bosch, Taylor, and Storm the pitch contour was labeled by hand to test and train a classifier. Pitch has, of course, a direct accent signaling function, but there are multiple cues for accent of which pitch is only one. Wightman and Ostendorf (1994) choose to use hand-labeled prominences to train and test an automatic classifier.

Our present approach will be to mark the prominence of words or syllables by perception experiments. Two perception experiments are run, a word perception experiment and a syllable perception experiment (see section 2). Naive listeners are asked to mark those words or syllables, which are spoken with emphasis (this is an operational definition of prominence). The cumulative score per word or per syllable is an indication of how prominent words or syllable are. The words, which are perceived as emphasized by a majority of listeners are considered to be the prominent words.

In the word perception experiment, naive listeners have to mark the prominent words (word perception experiment). In the second perception experiment naive listeners have to mark the prominent syllables (syllable perception experiment). The question is which type of prominence detection, the detection based on words or on syllables, came nearest to sentence accent.

The importance of the pitch movements is investigated in a third perception experiment. In this perception experiment the subjects hear re-synthesized sentences with monotone pitch. The subjects had to mark the perceived prominent words under this condition (see section 3).

Both the most prominent and least prominent words (according to the results of the word perception experiment), were selected for various acoustical analyses. The mean and the range of the pitch movement per word were calculated. Also the mean intensity per word was calculated. These acoustical features are used to classify a given word as prominent or non-prominent (see section 4). For the classification task an artificial neural networks with different topologies and different input vectors was used. The output was always discrete, either accented (Accent) or non-accented (NAccent).

2. Perception experiments to mark prominence

Two perception experiments, using different instructions and a different layout of the sentences, will be presented. In both experiments the acoustical presentation of the sentences is identical, but the text is displayed in a different way. For the word perception experiment the listeners see the normally written text on the monitor and the instruction is to mark all emphasized spoken words. For the syllable perception experiment a white space between the syllables additional to the '-' sign is displayed. The task is to mark all emphasized spoken syllables (see for more detail section 2.2). In the syllable perception experiment the subjects use a more analytic perception mode, while comparing each syllable with its neighbors, than in the word perception experiment. A word perception experiment leads the attention to a higher level, the sentence or phrase level. The listeners than compare each word with other words in the sentence.

> Research hypothesis:
> A word perception experiment is a better instrument to detect prominence
> at the sentence level than a syllable perception experiment.

Therefore it is expected that in the word perception experiment a lower number of prominence judgments per sentences will be given than in the syllable perception experiment, because in the syllable perception experiment also realized words stress regularly be perceived as prominent.

2.1. Speech material

The speech material, 81 phonetically rich sentences, was selected from the Polyphone corpus. This corpus, that is available on CD-ROM, contains 5 phonetically rich but varying sentences from 5000 different speakers. In total, 12500 different phonetically rich sentences were constructed. So each sentence was spoken twice, each time by a different speaker. These 5 sentences are constructed in such a way that each set contains all phonemes of the Dutch language at least once. The speakers are instructed to read the

sentences aloud from paper via the telephone. This material was digitized with a sampling frequency of 8000 Hz (for more details see Damhuis et al., 1994).
As far as possible, we included all 5 utterances of the 19 speakers selected. However some of the sentences had to be discarded due to bad sound quality, resulting in 81 sentences spoken by 19 different speakers. 9 male and 10 female speakers speak 40 or 41 of these sentences, respectively. The grammatical structure of these sentences varied.

2.2. Design of word perception experiment and syllable perception experiment

An UNIX workstation controls the perception experiments. Eight subjects per perception experiment hear each sentence repeated three times and see the sentence in written form. The sentences are presented in random order. Under each word a button is displayed on which the subject can click whenever such a word is judged as being prominent. The instruction was to mark all those words, which were spoken with emphasis ("met nadruk zijn uìtgesproken").

In the second perception experiment the sentences were displayed on the monitor with spaces between the syllables. Similar to experiment 1, the subjects hear each sentence repeated three times via headphones. Below each syllable a button is displayed on which the subjects click if that syllable is spoken with emphasis ("welke lettergrepen beklemtoond zijn uitgesproken").

Once the subject has clicked on one or more of these words or syllables, another button with the word "klaar" ("ready") can be touched. Then the next sentence is displayed on the monitor and is made audible.

The 16 subjects (8 subjects for each perception experiment) were all employees or students of the Institute of Phonetic Sciences. The listeners were not trained and were not paid for this task. All listeners were native speakers of the Dutch language and did not report any hearing loss. The experiments took about 35 minutes each, and the acoustical signals were presented via closed headphones. The responses of the subjects were automatically recorded.

2.3. Results of the word and the syllable perception experiment

The 81 sentences used in both perception experiments consist of 853 words and 1461 syllables. The average number of words is 10.5 per sentence the average number of syllables per sentence is 18.0.

The total number of prominence judgments in the word perception experiment for all 853 words in the 81 sentences, added over all 8 listeners, was 1890. The average number of prominence judgments for all 853 words per listener was 236.3 ± 60.6 (sd). The average prominence judgment per listener per sentence for the word perception experiment is 2.9 ± 0.7 (sd).

The total number of prominence judgments in the syllable perception experiment for all 8 listeners is 3315. The average number of prominence judgments per listener is 414 ± 137 (sd). This results in an average score per sentence of 5.1 ± 1.7 (sd) of prominent syllables.
An example of how the results of the word perception experiment are stored in a matrix is given in table 1. In table 2 a part of the raw data of the syllable perception experiment is given. This matrix has the same structure as the matrix presented in table 1, but now not the words but the syllables are given in the third column. The sums of the scores per word or per syllable are shown in the last column of both tables.

Table 1: In this table a part of the raw data matrix of the word perception experiment is represented. The individuals as well as the cumulative word prominence scores over all 8 listeners are given.

			listeners								
sent. num.	word num.	words	1	2	3	4	5	6	7	8	sum
1	1	Er	0	0	0	0	0	0	0	0	0
	2	gaat	1	1	0	1	1	1	1	1	7
	3	om	0	0	0	0	0	0	0	0	0
	4	half	0	1	1	0	0	1	0	0	3
	5	drie	0	0	1	0	0	0	0	0	1
	6	een	0	0	0	0	0	0	0	0	0
	7	bus	1	0	0	0	1	1	1	1	5
	8	uit	0	0	0	0	0	0	0	0	0
	9	Amsterdam	1	1	1	1	1	1	0	1	7
	10	naar	0	0	0	0	0	0	0	0	0
	11	Utrecht.	1	1	1	1	1	1	1	1	8

Table 2: In this table part of the raw data matrix of the results of the syllable perception experiment is represented. The individuals as well as the cumulative syllable prominence scores over all 8 listeners are given.

			listeners								
sent. num.	syll. num.	syllables	1	2	3	4	5	6	7	8	sum
1	1	Er	0	0	0	0	0	0	0	1	1
	2	gaat	1	1	1	1	1	1	1	1	8
	3	om	0	0	0	0	0	0	0	0	0
	4	half	0	1	0	1	1	0	0	0	3
	5	drie	0	0	1	1	1	1	0	0	4
	6	een	0	0	0	0	0	0	0	0	0
	7	bus	1	1	1	1	1	1	1	0	7
	8	uit	0	0	0	0	0	0	0	0	0
	9	Am-	1	1	1	1	1	1	0	0	6
	10	ster-	0	0	0	0	0	0	0	0	0
	11	dam	0	0	0	1	0	1	0	0	2
	12	naar	0	0	0	0	0	0	0	0	0
	13	U-	1	1	1	1	1	1	1	1	8
	14	trecht.	0	0	0	0	0	0	0	0	0

It is difficult to compare the two perception experiments because words would somehow have to be compared with syllables. To solve this problem we took only the prominence scores of the lexically stressed syllables and compared these with the scores of the word experiment. Table 3 presents a matrix in which the scores of the word experiment and the scores of the syllable experiment are compared.

For example in the word 'notehout' (walnut), where the first syllable has lexical stress, only the cumulative score '7' on 'no' is compared with the score '7' on the word 'notehout', the scores '3' and '4' are neglected (see below). This adds one point to the 7 : 7 cell entry in table 3.

No- te- hout
7 3 4
Notehout
7

Table 3: This correspondence matrix shows the cumulative score of the lexically stressed syllables of the syllable experiment and the cumulative score of the word experiment. This matrix contains the judgments of the 81 sentences, which are used in the two perception experiments. See for further explanation the text.

score of the syllable experiment

		0	1	2	3	4	5	6	7	8	
	0	253	95	55	16	7	2	5	0	0	433
	1	1	8	15	11	11	12	6	1	0	65
	2	1	3	1	6	7	8	9	5	1	41
score of word	3	0	1	2	7	2	15	10	11	5	53
experiment	4	0	0	0	6	5	7	13	7	6	44
	5	0	0	0	1	2	3	20	15	9	50
	6	0	0	0	1	3	7	17	19	16	63
	7	0	0	1	2	3	3	2	22	16	49
	8	0	0	0	0	1	2	7	20	25	55
		255	107	74	50	41	59	89	100	78	853

The matrix in table 3 shows the correspondence between the syllable and the word perception experiment. For example cell 0 : 0 contains the number 253, this means that 253 of the 853 words altogether received the score 0 in both experiments. The number 12 in cell 1 : 5 indicates the number of times that in the word experiment only one listener marked a word as being emphasized whereas in the syllable experiment 5 of the 8 listeners marked the lexically stressed syllable in the same word as prominent. There are in the word perception experiment 104 words judged as prominent by the majority of the listeners. This results in an average prominent word per sentence of 1.3. For the syllable perception experiment there are 178 syllables of the lexically stressed word judged as prominent by the majority of the listeners. This would result on average in 2.2 prominent words per sentence.

The cells above the diagonal are more filled than the cells below the diagonal. This fact is reflected even better in figure 1, in which the total scores in all cells parallel to the diagonal are shown.

If the listeners had listened with the same listening mode (see section 2), we would expect the scores of the two perception experiments to be symmetrically distributed around the diagonal. That is not the case (see figure 1) the distribution is skewed to the right. The cases above the diagonal (443) and the cases beneath the diagonal (69) are summed up and it is tested with a sign test if these values are equally likely. It turns out that these two values are significantly different from each other (n+ = 443, n- = 69, p ≤

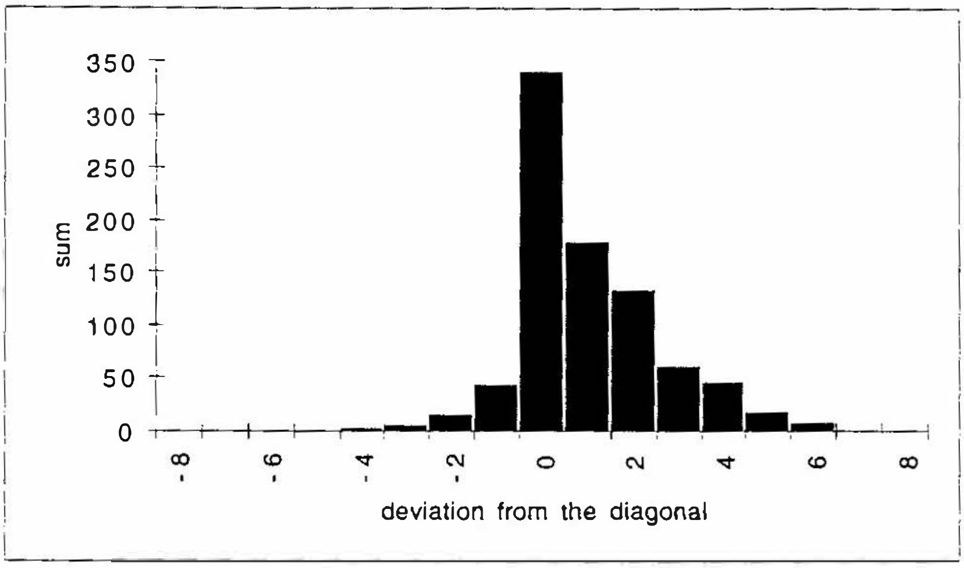


Figure 1: In this figure the total scores parallel to the diagonal are shown. The column indicated with 0 gives the sum of the scores in the diagonal of the matrix in table 3. Positive numbers in the figure mark the sum of the score one or more rows above the diagonal, negative numbers show the sum of the cells below the diagonal.

0.001). First, this indicates that more syllables per sentences are marked prominent than words. Second it can be concluded that in most cases, if a word is judged to be prominent this corresponds to a prominence judgment in the lexically stressed syllable in that word.

An example of a high cumulative score in the syllable perception experiment and a low cumulative score in the word perception experiment is the sentence shown in table 4. Such a disagreement between the cumulative scores of both perception experiments, is common as shown table 3 and figure 1. In this sentence (see table 4) there is one word 'notehout' (walnut) that is the best candidate for realized sentence accent, the results of both perception experiments show that this word is indeed perceived as prominent by the listeners. Furthermore the scores for the words 'dure' (expensive) and 'houtsoort' (wood type) in the word-perception experiment are not very high, '4' and '3' respectively. However the score for the lexically stressed syllables of these words are very high, '8' and '7', respectively (see table 4). An explanation for this could be that the lexically stressed syllables in the words 'dure' (expensive) and 'houtsoort' (wood type) are realized with syllable stress. In these words there is realized syllable stress but not a sentence accent. From the word perception experiment it turns out that the word 'notehout' (walnut) is the most prominent one in this sentence and not the words 'dure' and 'houtsoort'.

Table 4. This table shows the results of the two experiments for sentence "Notehout is een dure houtsoort" (walnut is an expensive wood type). In the first row the sentence is written down, in the second row the cumulative score per word and in the fourth row the cumulative score per syllable are presented.

Notehout			is	een	dure		houtsoort	
7			0	0	4		3	
No-	te-	hout	is	een	du-	re	hout-	soort
7	3	4	0	1	8	1	7	0

An example of a high score in the word and a low score in the syllable perception experiment is shown in table 5. Such cases are rare. In the word 'Amsterdam' lexical stress is normally realized on the third syllable. However in this case the speaker realized word stress on the first syllable. This adds one point to the 7 : 2 cell entry in table 3 and not to the 7 : 6 cell entry.

Table 5. This table shows the results of the two experiments for sentence "Er gaat om half drie een bus uit Amsterdam naar Utrecht" (A bus is going at half past two from Amsterdam to Utrecht). In the first row the sentence is written down, in the second row the cumulative score per word and in the fourth row the cumulative score per syllable are presented.

Er	gaat	om	half	drie	een	bus	uit	Amsterdam			naar	Utrecht.	
0	7	0	3	1	0	5	0	7			0	8	
Er	gaat	om	half	drie	een	bus	uit	Am-	ster-	dam	naar	U-	trecht.
1	8	0	3	4	0	7	0	6	0	2	0	8	0

In future research we want to perform the acoustical measurements of the prominent words automatically. The search for cues should be done in the lexically stressed syllable because these are the only syllables that we can easily identify. We want to know in how many cases there is a syllable, which although having the highest cumulative score, is not the lexically stressed syllable. In this pilot study with 81 sentences this was only found in the word "Amsterdam", were the third syllable has lexical stress but the prominent syllable is the first syllable (see table 5). In such a case probably caused by stress clash or rhythmic requirements, the search for cues, which lead to the perception of prominence, will be done on the wrong syllable.

2.4. Conclusion of the perception experiments

First of all it can be said that the perception experiment, in which the subject had to mark emphasized spoken words, is an easy task to do for listeners. The average number of prominence judgments per sentence is in case of the word perception experiment (2.9 ± 0.7), lower than in the case of the syllable perception experiment (5.1 ± 1.7). From a t-test for two samples it turns out that the two mean values differ significant from each other ($t = -3.356$, $\nu = 14$, $p \leq 0.005$). In the syllable perception experiment, realized word stress are also judged as prominent. Therefore the word perception experiment came closer to sentence accent than the syllable perception experiment. This could be explained as follows: There are two different perception modes used in the two perception experiments. In the syllable perception experiment, it is most likely that listeners use an analytical perception mode in which they compare each syllable with the surrounding syllables. If this is the case, the listeners have more syllables to compare and the result is a higher average score per sentence (5.1). In the case of the word perception experiment the listeners compare a word with the surrounding words this results in a lower average score per sentence (2.9).

3. Prominence marking without pitch movements

According to most prosody models, pitch movement is the most important feature to mark prominence. This raises the question whether in monotone sentences any prominence can be marked. With the help of a pilot perception experiment we want to investigate, if and how well naive listeners can mark the prominent words even when the pitch is monotonous. The speech material consisted of 30 sentences, which were randomly selected from the subset of the Polyphone corpus used in the two perception experiments described above. These 30 sentences are re-synthesized with a monotone pitch using PSOLA without duration manipulations. This pilot perception experiment is done in two runs simply because there were to groups of students available. In each run 15 monotone sentences are presented via headphones to the listeners. A total of 16 naive listeners (for each run 8 naive listeners) were instructed to mark one or more emphasized words in the sentences. Each sentence was repeated 3 times, the written text was displayed on the computer screen and the listener had to click a button below those words, which the subject judged to be spoken emphasized. The prominence judgments of these two subsets of 15 sentences are compared with the prominence judgments for the same sentences from the word perception experiment described in section 2.3.1.

3.1. Results

Since we had 2 subgroups of sentences we had to treat their results separately. The 15 sentences of the first subset contained 147 words, so the mean number of words per sentence is 9.8. The total number of prominence judgments for all 8 listeners is 224. The average number of prominence judgments per listener is 28 ± 9 (sd). The average number of prominence judgments per sentence per listener used in the first run is 1.9 ± 0.6 (sd).

In the second run the 15 sentences consist of 165 words, resulting in 11.0 words per sentence on average. All 8 listeners of the second run judge a word as prominent 311 times. The average number of prominence judgments per listener is 38.9 with a standard deviation of 17.9. The average number of prominence judgments in the 15 sentences used in the second run is 2.6 ± 1.2, this is substantially higher than that for set 1. Maybe the two sets are not comparable.

The results of both sets of this monotone perception experiment and the results of the previous word perception experiment are compared in a correspondence matrix (table 6).

Table 6: A correspondence matrix for the scores of the perception experiment without pitch movements and the results for the same 30 sentences from the regular word perception experiment.

		word experiment without pitch movements									
		0	1	2	3	4	5	6	7	8	
	0	118	25	10	0	0	0	0	0	0	153
	1	6	8	5	5	2	0	0	0	0	26
	2	2	4	2	2	1	0	0	0	0	11
word experiment	3	2	7	2	5	2	2	0	0	0	20
	4	2	3	4	4	2	0	1	0	0	16
	5	1	1	9	3	7	3	1	0	0	25
	6	0	1	4	3	4	4	0	0	0	16
	7	0	2	2	3	6	3	4	0	0	20
	8	0	0	1	1	3	7	7	4	2	25
		131	51	39	26	27	19	13	4	2	312

The results in table 6 show, that from the 45 words uniformly judged to be prominent the regular word perception experiment (if we take the words with score 7 or 8) the majority of the listeners still mark 6 words as being prominent even if there is a monotone pitch. All listeners uniformly marked two words as prominent. The number of the non-prominent words is about the same as in the word perception experiment under normal conditions. 179 for the word perception experiment and 182 for the perception experiment with monotone pitch. The number of words for which about half of the listeners (scores 2, 3, 4, 5, and 6) mark given words as prominent, increases: 88 words from the word perception experiment versus 124 words from the word perception experiment without pitch movements (see table 6). It is tested with a χ^2 test if 6 versus 45, 124 versus 88 and 182 versus 179 are from the same distribution (χ^2 = 34.10, v = 2, p = 0.001). The frequency distributions of the judgments of the two listening experiments are significantly different. The results show that the individual behavior of the listeners is quite different. It furthermore indicates that the task of marking prominence in sentences with a monotone pitch is difficult but still possible.

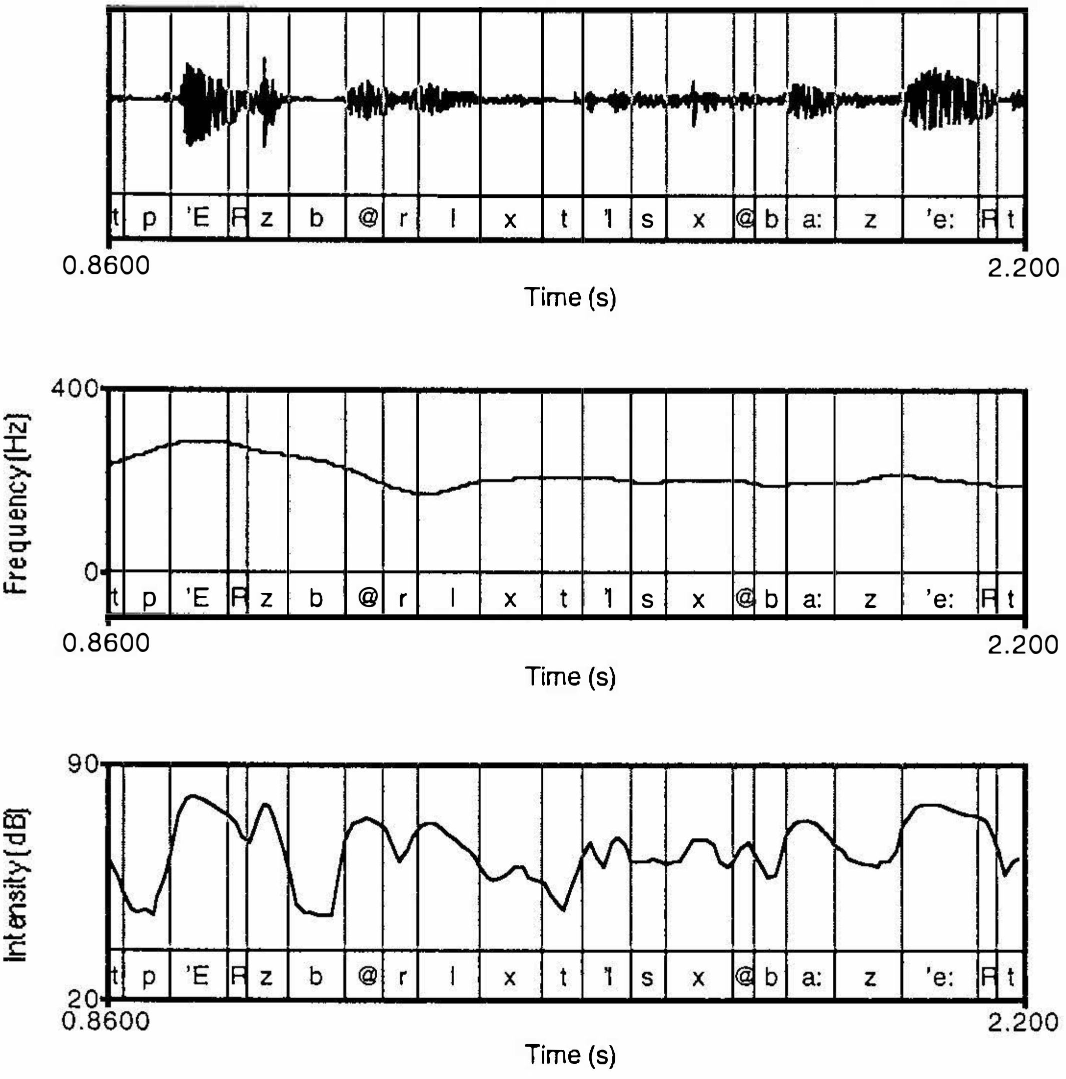


Figure 2: Acoustical realization of the words 'persbericht is gebaseerd' (press report is based, /pɛrzbəʀɪxtɪsxəba:ze:ʀt/) and the segmentation in SAMPA is shown in this figure. The listeners unifonnly mark the word 'persbericht' as prominent when the sentence was presented with a monotone pitch.

3.2. Discussion and conclusion

We can conclude from this pilot perception experiment that listeners are still occasionally able to mark prominence in monotone sentences. From the 45 prominent words from the perception experiment under normal conditions the majority of the listeners under monotone pitch perceives still 6 words as prominent. Top-down information and the expectation of the listener about the prominent words in the sentence could explain the decrease of the score 2, 3, 4, 5, 6 (see table 6). The implication of this perception experiment is that prominence at sentence level is not only evoked by pitch movements, but that other acoustical correlates, such as duration and intensity, can be additional features for the listener to perceive prominence.

In figure 2 the acoustical realization of the word (/p'ɛʀzbənxt/, press report) is shown, which is marked unanimously as prominent by all 8 naive listeners in the monotone pitch experiment. In the original utterance there is a clear pitch movement realized on the word (/p'ɛʀzbəʀɪxt/), but despite the absence of this cue the perception of prominence is very clear.

The intensity, or the rèlative long duration of the vowel /ɛ/ could be an explanation. The short vowel /ɛ/ is relative long with regard to the long vowel /aː/ in the word (/xəbaːz'eːʀt/, based). In case of the other 5 words with a prominence score of 7, the intensity or in the duration cues make these words so prominent. But still further research is needed for a more sophisticated explanation of the uniform judgment of the listeners.

4. Acoustical analyses

Acoustical analyses have been carried out on the set of 81 sentences from the Polyphone corpus. First of all the speech material is automatically labeled at the segment level with a Hidden Markov Model with the help of Xue Wang (Wang, 1997). We analyze only those words for which the majority of listeners mark a given word as prominent (7, 8) or the majority of the listeners don't mark words as prominent (0, 1). This results in a total of 104 words with a score of 7 or 8 (Accent) (see last 2 rows in table 3) and 498 words with a score 0 or 1 (NAccent) (see first 2 rows in table 3). To make sure that there is an honest testing and training situation for the classification with the neural net we randomly selected 104 from the 498 non-accented words (NAccent). This corresponds to the same number as all available accented words (Accent).

Because we could not yet assign the syllable boundaries automatically, these are not available. The segment boundaries from the HHM segmentation and the word boundaries, as well as the prominence score of the word perception experiment are available, therefore we choose to use the word boundaries for these pilot measurements. However, even if we had had the syllable boundaries, it would still be unclear at what time instances to perform the measurements (e.g., vowel onset plus or minus 60 ms as done by Ten Bosch, 1993; or the whole syllable as done by Wightman and Ostendorf, 1994). So, in order to test the automatic acoustical measurement procedure and to see how far we can get with these acoustical measurements the acoustical measurements are done on whole words. In future (see Streefkerk, 1997) the acoustical measurements will also be done on other segments such as syllables or vowels of the prominent and non-prominent words.

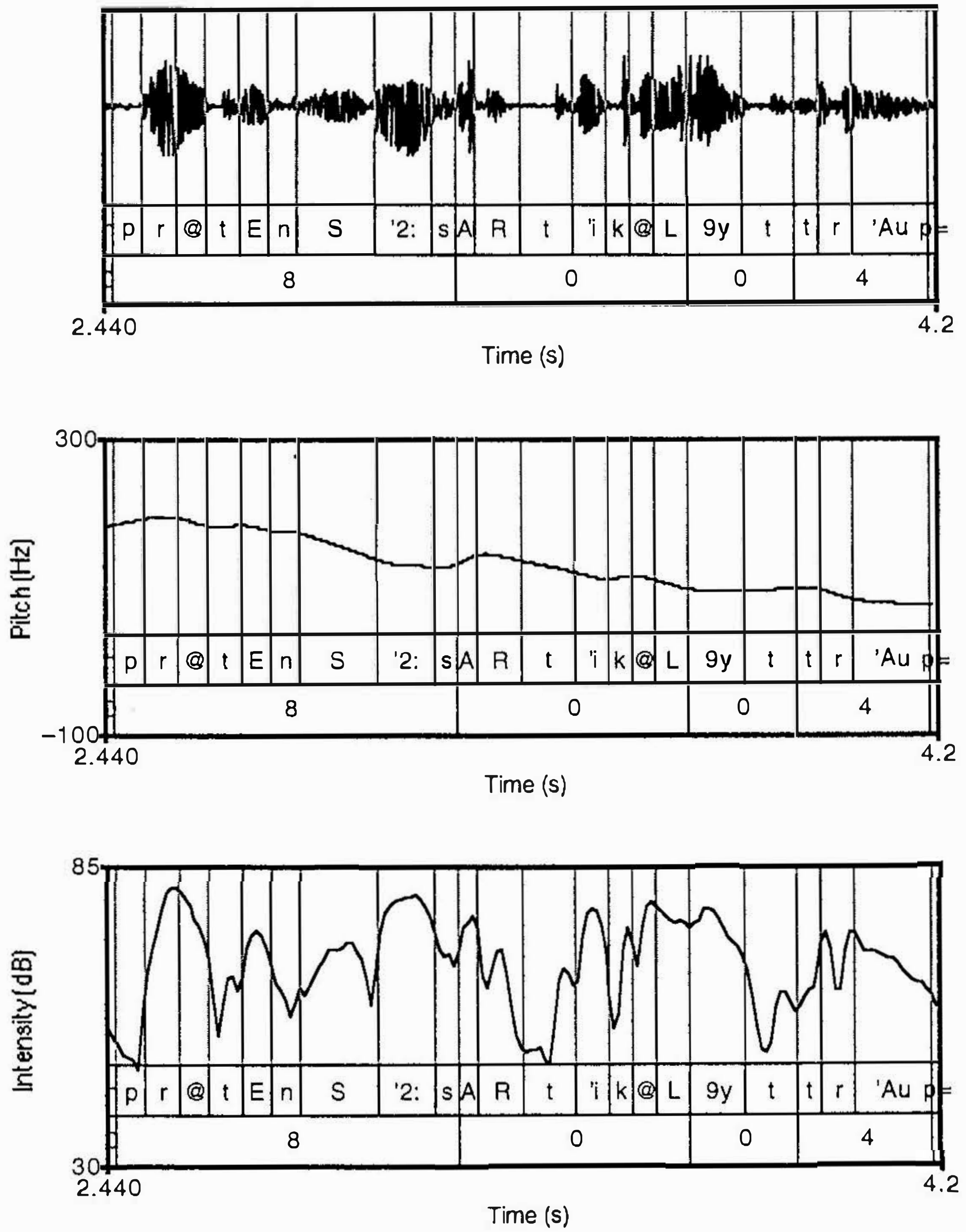


Figure 3: An example of a sentence and its acoustical measurements ('pretentieus artikel uit Trouw', /prətɛnʃøsɑrtikəlœyttrɑu/, pretentious article from Trouw, i.e., a newspaper). The top panel shows the oscillogram with the automatic segmentation and the cumulative listener word prominence judgments. The middle and bottom panel display F_0 and intensity, respectively.

In this pilot the selected words are acoustically analyzed with the help of the software package 'Praat' (Boersma and Weenink 1996). The following values were determined per word:

- Mean F_0 (semitones)
- Range of the F_0 (semitones)
- Mean intensity (dB)

The F_0 range is defined as the difference between the maximum and the minimum. The pitch contour is measured with the autocorrelation method. The pitch contour is corrected for octave jumps, whereas we also interpolate the pitch movements and smooth them, so that there is a continuos pitch contour without gaps for the voiceless parts of the signal. The mean and range values of the F_0 and the mean intensity measurements per word are used as input features for a neural network classifier. The input features for the neural network must be scaled between 1 and 0, so the acoustical measurements are divided by the highest value over all used words to make sure that the input data is scaled between 1 and 0.

5. Classification with an artificial neural net (ANN)

The total of 208 selected words is divided into two groups: a test set (60 words) and a training set (148 words) with equal numbers of prominent and non-prominent words and with equal numbers of male and female speakers. We used a feedforward net with a conjugate gradient learning algorithm for the training. Training and testing of Feed Forward Nets is implemented in the software package 'Praat' (Boersma and Weenink, 1996).

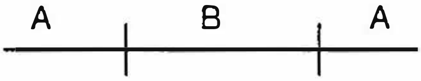
Table 6: Percentages correct score of different ANN's with different input features and different topologies.

ANN with 2 nodes in hidden layer	mean F_0	range F_0	mean int.	RR test set	RR training set
	x	x	x	75%	85%
	x			60%	57%
		x		72%	83%
			x	48%	53%

ANN without hidden layer	mean F0	range F0	mean int.	RR test set	RR training set
	x	x	x	75%	84%
	x			62%	51%
		x		72%	83%
			x	47%	53%
classification on range F_0 if > 3.5 sem ¬A if < 3.5 sem A				75%	83%

Altogether 8 neural nets (ANN) with different topologies are trained; 4 with and 4 without a hidden layer. The output layer for all ANN's was the same (Accent or NAccent). The input features differ: there are ANN's trained with all 3 features (mean F_0, range F_0 and the mean intensity) and with each feature separately.

There is a large difference in recognition rates between the training and the test set: the scores of the training data are always better. One reason is that the test set is very small (only 60 words) and a single recognition error already has a large effect on the percentage correct score. This could be an indication that the net is over-trained. A comparison of the recognition rates of the ANN's with and without hidden layer shows that there is no positive effect of adding a hidden layer to the topology of the neural network for this type of data. Adding a hidden layer makes it possible to separate a class when the data of a class are lying in between the other class. For example when class 'B' lies in between class 'A'.



The recognition rates of the classification with a neural net show that the range of F_0 per word is a very important feature. If in the two ANN's with and without a hidden layer, only the range F_0 of each word is used as input feature to discriminate between Accent or NAccent, we come to a recognition rate of 83% for the training data set and 72% for the test data. The ANN without a hidden layer and with the F_0 range per word as input, is an important and interesting network. Below we discuss this ANN in more detail.

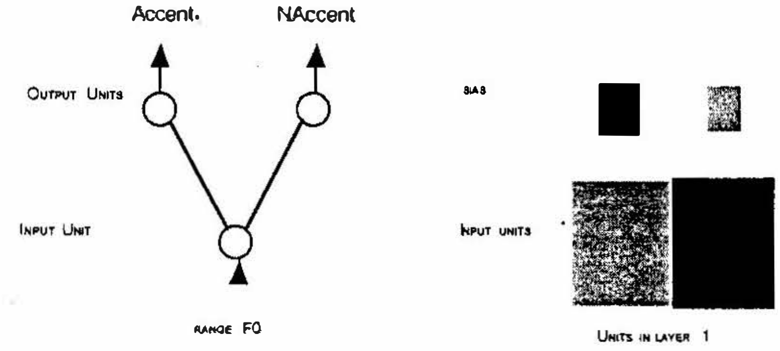


Figure 4: An ANN with the topology of 1 input node and two output nodes (1:2) and the weights of the ANN drawn in squares and the bias after training. Black squares indicate negative values and white squares indicate positive values.

In general each node of an artificial neural net can be calculated with the activation function (Kompe, 1997). In case of our simple ANN with two output nodes the finally resulting activation threshold of each output node can be expressed as follows, where O_1 and O_2 are the output functions of the two output units (see figure 4):

$$O_1 = \frac{1}{1+e^{-(+17.1\cdot\text{Range}-2.6)}}$$

$$O_2 = \frac{1}{1+e^{-(-17.1\cdot\text{Range}+2.6)}}$$

Because of the symmetry, the sum of the two output functions (O_1 and O_2) is 1; therefore there is one crossover point for the two functions (see figure 5). The crossover point is exactly the critical point were the ANN decides if the data belong to Accent or NAccent. The value for this crossover point for the F_0 range is 0.15. This value must be multiplied by the scaling factor (23.2) as used to scale the input for the artificial neural nets. The result is a value of 3.5 semitones for the F_0 range as the critical value.

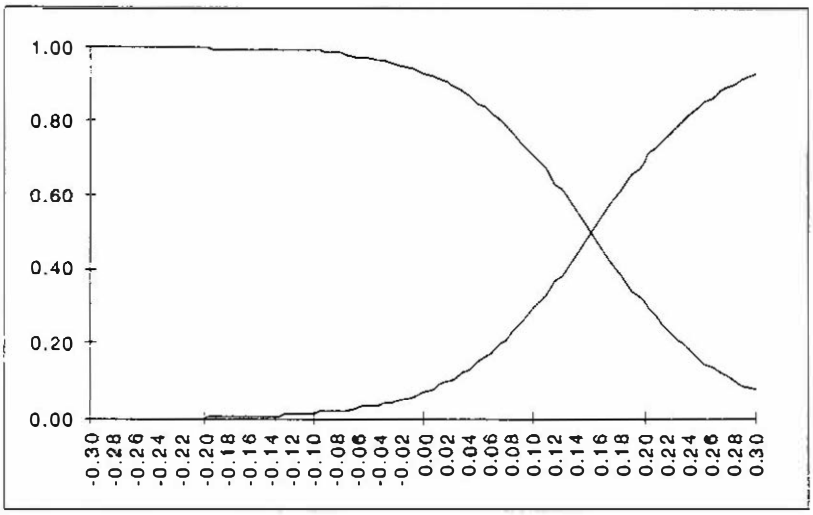


Figure 5: In this figure the graphs of the two activation functions for the output nodes are plotted. There is one crossover point near 0.15.

To verify this value (3.5 semitones), the F_0 range of the 60 words from the test set and the 148 words from the training set were sorted. The words with a range > 3.5 semitones were classified as Accent and the words < 3.5 semitones are classified as NAccent. Comparing this classification with the classification of the listeners, there is a correct recognition rate of 75% for the test set and of 83 % for the training set, which means that the range of F_0 per word could be a very important feature for the classification of Accent or NAccent. In this case with the help of the artificial neural net, a simple linear relation is found. This relation could also have been found by some statistical techniques, but in case of more complex relations ANN's are a good alternative to get to know more about acoustical correlates of prominence. The perception experiment with monotone pitch showed that other acoustical features such as duration and intensity could also be important. The classification task with a neural network with hidden layer also show this: the results with all 3 input features are always a little bit better than with one feature only. In case of additional features such as duration and intensity features such measurements are more dependent on intrinsic properties such as long versus short vowel and closed versus open vowels, than the F_0 feature.

6. Discussion and conclusion

First of all it can be said that perception experiments are useful tools to detect prominence. Further research must be done to investigate the relationship between prominence and the acoustical cues.

A simple ANN, using as acoustical input features mean F_0, range F_0 and the mean intensity per word, is already able to classify prominence with a 75% recognition rate on a test set of 30 most prominent words and 30 non-prominent words. Measuring features per word, the range of F_0 is the most important of these 3 features, and with the help of the weights of the ANN, the boundary of the range F_0 can be calculated. For this speech material, most of the time an F_0 range larger than 3.5 semitones indicated a prominent word and an F_0 range below 3.5 semitones a non-prominent word. This value is also mentioned by 't Hart et al. (1990). Duration and intensity features are more dependent on the intrinsic properties such as long versus short vowels and open versus closed vowels, therefore the measurement of these features is much more difficult than that of pitch features.

In earlier research of Streefkerk (1996a), 81.6 % of the prominent words appeared to be realized with a pitch movement in the lexically stressed syllable and 85.9 % in the whole prominent word (Streefkerk, 1996b). Pitch movements are defined as either a fall, a rise, or a peak in the pitch contour, so this implies that there is a high F_0 range for these words. These 81.6% to 85.9% pitch movements is of about the same order as the recognition rate on the range of the pitch movement of 75% on the test set and 83% on the training set. But the more interesting question, is what about the other 18.4% to 14.1%? In the perception experiment with monotone pitch, listeners are still able to mark 6 of the 45 original prominent words as being prominent. This is a strong indication that features such as energy, duration and spectral information are also useful for the listener to mark prominence next to pitch movements.

In further research (Streefkerk, 1997), acoustical features such as pitch, duration and energy and spectral quality must be investigated. The energy, duration and spectral quality features will most probably have to be corrected for vowel type, position of the syllable in the word and the position in the sentence, before they can be used as input features for a neural network. Pitch may be the primary feature for the perception of prominence but less is known about the other features. The interaction of the acoustical features for prominence is not yet investigated.

7. Acknowledgment

We thank M. Bellaar, S. Wassen, J. Rog, J. Aldenhoven and A. Olander, students of the Institute of Phonetic Sciences, who ran the perception experiment with monotone pitch and who performed the sentence manipulations.

8. References

Bagshaw, P. C. (**1993**). "An investigation of acoustic events related to sentential stress and accents, in English", *Speech Communication,* Vol. **13**: 333-342.

Boersma, P. and Weenink, D. (**1996**). *PRAAT: A system for doing phonetics by computer,* Report of the Institute of Phonetic Sciences of the University of Amsterdam **132**, (http://fonsg3.let.uva.nl/paul/praat.html).

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", *Proceedings ICSLP-94*, Yokohama, Vol. **4**: 1803 - 1806.

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

Kießling, A. (**1996**). *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung*, Ph.D. Thesis, *Berichte aus der Informatik*, Shaker Verlag, Aachen.

Kompe, R. (**1997**). *Prosody in speech understanding systems*, Ph.D. Thesis, Lecture Notes in Computer Science, Springer Berlin, New York.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J. (**1992**). "TOBI: A standard for labeling English prosody", *Proceedings ICSLP-92*, Banff, Vol. **2**: 981-984.

Strom, V. (**1995**). "Detection of accents, phrase boundaries, and sentence modality in German with prosodic features", *Proceedings Eurospeech' 95*, Madrid, Vol. **3**: 2039-2041

Sluijter, A. M. C. (**1995**). *Phonetic correlates of stress and accent*, Ph.D. Thesis, Leiden University.

Streefkerk B. M. (**1996a**). *Prominent zinsaccent en toonhoogtebewegingen*, Report of the Institute of Phonetic Sciences of the University of Amsterdam, **131**.

Streefkerk B. M. (**1996b**). "Prominent sentence accent and pitch movements", *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, **20** 111-119.

Streefkerk B. M. (**1997**). "Acoustical correlates of prominence: A design for research", *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, **21** 131-142.

Taylor, P. (**1993**). "Automatic recognition of intonation from F_0 contours using the rise/fall/connection model", *Proceedings Eurospeech'93*, Berlin, Vol. **2**: 789-792.

Ten Bosch, L. F. M. (**1993**). "On the automatic classification of pitch movements", Proceedings *Eurospeech '93*, Berlin, Vol. **2**: 781-784.

Waybill, A. (**1988**). *Prosody and speech recognition*, Ph.D. Thesis,, Carnegie-Mellon University.

Wang, X. (**1997**). *Incorporating knowledge on segmental duration in HMM-based continuous speech recognition*, Ph.D. Thesis, Amsterdam University.

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

Kraayeveld, J., Rietveld, A. C. M. and van Heuven, V. J. (**1991**). "Speaker characteriation in Dutch using prosodic parameters", *Proceedings Eurospeech'91*, Genova, Vol. **2**: 427-430.