Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 21 (1997), 131-142

ACOUSTICAL CORRELATES OF PROMINENCE: A DESIGN FOR RESEARCH

Barbertje Streefkerk

Abstract

In this project, all known acoustical correlates for prominence at the sentence level, such as F_0, intensity, duration and spectral quality will be investigated. The speech material will be selected from the Dutch Polyphone corpus as recorded by SPEX and KPN Research Leidschendam. The sentences from this corpus are suitable to investigate systematically (with high speaker variability) the acoustical correlates of prominent words. 5000 Dutch speakers each read aloud five phonetically rich sentences, which were recorded over the telephone. The prominent words are detected through a perception experiment, in which naive listeners have to mark emphasized spoken words (this is an operational definition of prominence). Then the acoustical differences between the prominent and non-prominent words will be investigated. The aim of this research project is not only to describe the acoustical correlates but also to develop a predictive model. This model, most probably in the form of an artificial neural network, will have to predict prominence using all relevant acoustical information. Furthermore, some perception experiments with manipulated speech are planned, in order to investigate the effect of F_0, intensity, duration and spectral quality on the perception of prominence.

1. Introduction

A major question in several recent research projects concerns the automatic classification of sentence accent (Bagshaw, 1993; Kießling, 1996; Kompe, 1997; Taylor, 1993; Ten Bosch, 1993; Storm, 1995; Vaissière, 1989; Waibel, 1988; Wightman and Ostendorf, 1994).

There are several confusing terms used in these studies: sentence accent, pitch accent and prominence. Literature provides no unique definition of sentence accent, but it is obvious that it must refer to some major accent in a sentence. Pitch accent is an accent-lending pitch movement, whose realization has consequences for the duration and intensity (Gussenhoven, 1984). A term as pitch accent is implemented in a linguistical and theoretical concept: an intonation grammar or an intonation system, such as the IPO intonation grammar ('t Hart et al., 1990), the TOBI intonation system, (Silverman et al., 1992), or the Rise/Fall/Connection model (Taylor, 1992). Prominence refers to a greater perceived strength of words in a sentence, or put in another way, such words are perceived as standing out from their environment (Ladd, 1996; Lehiste, 1970; Terken, 1991). Lexicalstress is defined in the lexicon. Realized syllable stress makes a syllable more prominent than the surrounding syllables (see table below).

Phenomena like pitch accent and sentence accent lead to perceived prominence. In the case of sentence accent and pitch accent, words are compared with adjacent words

in the sentence. In the case of realized syllable stress, syllables are compared with adjacent syllables. Realized syllable stress is perceived as the most prominent syllable in a word, whereas pitch accents are perceived as the most prominent words in a sentence.

	domain	definition	perception of naive listeners
prominence	sentence	words perceived as standing out from their environment	emphasized words
sentence accent	sentence	major accents	emphasized words
pitch accent	sentence	an accent-lending pitch movement perceived by an expert	emphasized word
realized syllable stress	word	syllable perceived as standing out from its environment	emphasized syllable
lexical stress	word	defined in the lexicon	could be perceived as an emphasized syllable

Several attempts have been made to classify accented and non-accented words (Kießling, 1996; Kompe, 1997; Taylor, 1993; Ten Bosch, 1993; Storm, 1995; Wightman and Ostendorf, 1994). Literature provides several approaches for initial labeling of spoken utterances for accent and non-accent for training and testing purposes.

One approach is to label the pitch contour according to the IPO intonation grammar (Ten Bosch, 1993; Taylor, 1993). In the research of Ten Bosch, four intonation experts transcribed a speech corpus. The experts were asked to transcribe utterances by using the IPO intonation categories (the labels "1" to "5" for pitch rises and the labels "A" to "E" for falls, plus the "P" for a peak realized in one syllable). In the IPO intonation grammar pitch movements such as "A", "C" and "1" and "3" are accent-lending. With this labeled speech material a predictive model is developed and tested.

Taylor used four elements to describe tune (pitch contour); type "H" (high) or "L" (low) describe the pitch accents, "C" is used to describe a phonologically significant connection elements and "B" is used to describe the rise that may occur at phrase boundaries. In the research of Storm (1995) the speech material is labeled according to TOBI intonation system (Silverman et al., 1992).

The disadvantage of these approaches is, that only the pitch contour is taken into account, although in case of the TOBI intonation system there is also attention for the break indices. For rule-synthesis purposes it is sufficient to have intonation systems such as the TOBI intonation system, or the IPO intonation grammar. However, for speakers of the Dutch languageit is not mandatory to realize an accent with a pitch movement alone, there are other acoustical features such as intensity, duration and spectral quality to mark accents. If one has the aim to improve speech recognition, it is not wise to limit oneself to accent-lending pitch movements either. Rather the variability between speakers in realizing accents and the use of different acoustical cues should also be taken into account (Kraayeveld et al., 1991).

Another approach is to label the utterances for accent or non-accent based on linguistic, semantic and phonological information (Batliner et al., 1997). In the research of Kompe (1997) and Kießling (1996), the initial labeling of accent versus non-accent is done automatically for the ERBA corpus (Erlanger Bahn Anfragen). They assume that in each prosodic phrase *one* word is more prominent than all other words. Following this line they apply such rules as the right most content word of a phrase being a good candidate for sentence accent. They use these rules to label their speech

material. With the help of this labeled data base they build and test a predictive model based on acoustical information. Through this initial labeling, certain words are marked as accented while the speaker has not necessarily realized them as such. The predictive model, which classifies accent or non-accent with the help of acoustical features, then gets the wrong features, because this accent might not be realized in the spoken utterance.

In our present approach, prominence is initially marked via perceptive judgments. Naive listeners will be asked to mark the words which are spoken with emphasis (this is an operational definition of prominence). The words, which are perceived by the majority of the listeners as prominent are defined as being the prominent words. With these prominent and non-prominent words a predictive model will be trained and tested.

The sentences presented in the perception experiment are not delexicalizedand the listener will, next to the acoustic information, also have an expectation about which words are the prominent ones based on top-down information. Beside the influence of top-down information, we can be certain that there is something in the speech signal that makes words prominent. We must assume this because acoustical features are extracted from the speech signal to predict the prominence of a word and not linguistical or semantical features. De Pijper and Sanderman (1994) found no effect of top-down information when the listeners had to mark boundaries in delexicalized speech. However, if it turns out that there is still a strong effect of top-down information and if this is a disturbing factor, a pen and paper experiment is a possible option to test this effect.

2. Speech material

The speech material, which is used in this research project, is taken from the Polyphone corpus. This corpus is recorded by KPN Research and SPEX and is available on CD-ROM. Care is taken to have a proper distribution over the 5000 speakers with respect to age, regional background and sex. This corpus contains, among other recordings, 5 phonetically rich sentences per speaker. These 5 sentences, which differ per speaker, 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 5 sentences aloud from paper. Their speech is recorded via the telephone and digitized with a sampling frequency of 8000 Hz. For more details see Damhuis et al. (1994). The speech material used in this project can thus be characterized as read aloud telephone speech, spoken by many different speakers (male and female), who have different regional backgrounds, and different ages. This speech material is pre-eminently suited to investigate acoustical correlates of prominence for many different speakers. A possible disadvantage of this corpus might be that all sentences are spoken independently, out of context. This might have increased the variability of the prominence realizations, but on the other hand this speech material is rather characteristic for various speech technology applications.

3. Marking prominence with perception experiments

Our initial approach is, to mark the prominence of words via listener judgments. The advantage is that not only the pitch contour is taken into account. A pilot perception experiment was run, in which 8 naive listeners had to mark the emphasized spoken words in a subset of 81 sentences. It turns out that, first of all, naive listeners are

indeed able to mark consistently prominent words. The cumulative prominence judgment is an indication of how prominent a word is. In the pilot perception experiment the majority of the listeners judges as prominent 104 of the 853 words. On average this is 1.3 words per sentence. These words are defined as the prominent words.

There are some listeners who have a tendency to mark more words per sentence as prominent than other listeners do (see for more details Streefkerk et al., 1997). Therefore the individual prominence judgments will be corrected per listener, before the cumulative judgments are used as a prominence indicator. Dividing each score of the listener by the sum of the total number of prominence judgments of that listener, is a possible correction for individual listener behavior.

In further research a subset of some 500 sentences will be randomly selected from the Polyphone corpus. With this subset a perception experiment will be done in which naive listeners have to mark the words spoken in an emphasized way. The results of this perception experiment will give us the prominent and non-prominent words. With these prominent and non-prominent words we can train and test an artificial neural network for an automatic prominence classification task. We must assume that there are acoustical cues in the speech signal, which lead to the perception of prominence, and that not only the top-down information is responsible for the prominence judgments. The automatic classification task (classify prominent and non-prominent words) will be done with acoustical features.

4. Perception experiments

4.1. Expert perception experiment

With the help of both an expert perception experiment (judgingpitch accent), and a naive-listener perception experiment (judging prominence), we want to investigate the relation between pitch accent and prominence. With an expert perception experiment

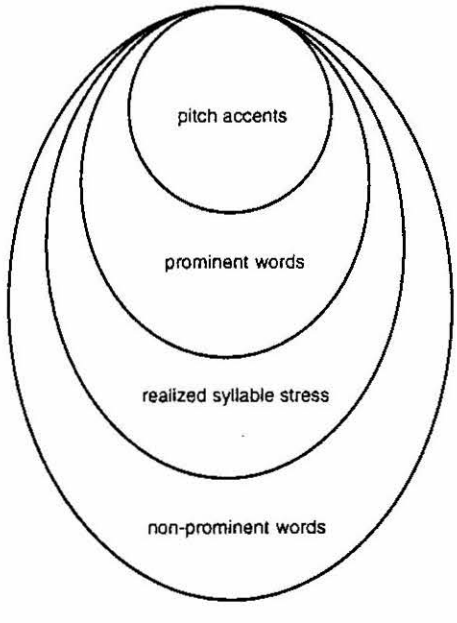

Figure 1: The relation between realized syllable stress, perceived prominence of words and pitch accent is displayed. With an expert perception experiment it should be tested if it is true that pitch-accented words are a subset of the prominent words.

and the prominence judgments of naive listeners it could be tested if words with a pitch accent are a subset of the prominent words. In figure 1 the relation between realized syllable stress, prominent words, and pitch accent is displayed. Realized syllable stress is the specification of a syllable and its domain is a word, whereas the other terms have the sentence as their domain (for further explanations, see section 1).

We expect that all pitch accents correspond to a prominence judgment, but that not all prominent words correspond to pitch accents. Such an expert perception experiment will be done with a subset of about 80 sentences from the Polyphone corpus. Independent of each other, 5 experts will label these 80 sentences for pitch accent. These data can then be compared with the prominence judgments of the perception experiment. We expect that the pitch-accented words are a subset of the prominent words. Or to say it in a different way, each pitch accented word must correspond with a prominent word but not each prominent word must correspond with a pitch accent (see figure 1).

4.2. Perception experiments with manipulated speech signals

With the help of perception experiments with manipulated speech, the effect of the acoustical correlates on the perception of prominence can be investigated. For the acoustical analyses of the prominent and the non-prominent words (see section 5), as well as for the perception experiments with manipulated speech, we need reference points. These reference points are the prominence judgments done on the normal speech signal.

A small pilot experiment with monotone pitch with 30 sentences was already done. In this experiment the perception of prominent words in sentences with monotone pitch is studied (see for more details, Streefkerk et al., 1997). The results of this perception experiment with monotone pitch show that listeners are still able to mark consistently some of the prominent words. From the 45 words perceived as prominent in the perception experiment under normal conditions, the majority of the listeners still perceive 6 words as prominent in the monotonized speech. This is about 13.3 % of the prominent words. In order to get a better overview of the effect of the perception of prominence with a monotone pitch this pilot experiment will be repeated with more sentences.

A perception experiment with monotone pitch with about 100 sentences is suggested. The subset of 100 sentences could be selected from the larger subset of 500 sentences for which listeners had to mark prominence under normal conditions. It might be better to select those sentences for which the subjects mark quite unanimously words as prominent. For this selected subset of sentences, the pitch will be made monotonous. Naive listeners have to mark the emphasized words under this condition. The result of this perception experiment with monotone pitch will be compared with the results of the original perception experiment without any speech signal manipulations. We expect that a lot of originally prominent words are not perceived as prominent anymore. But as in the pilot perception experiment with monotone pitch there will still be some words which are perceived as prominent by the majority of the listeners (see for more details Streefkerk et al., 1997).

It would be an interesting question to see which acoustical cues in these words are responsible for the perception of prominence. It could be the vowel duration, the intensity or the spectral quality. To figure this out, another perception experiment is suggested. Only the monotonized sentences with perceived word prominence will be used in these perception experiments. We suggest to make 3 subsets of manipulated sentences, one set of duration manipulations, one set of intensity manipulations and one

set of spectral quality manipulations. Phrasing and pauses could also have an effect on the perception of prominence, but as a start we suggest to manipulate only intensity, duration and spectral quality. If it turns out that also phrasing and pauses have such a strong effect on the perception of word prominence in future research we can manipulate these cues too.

The following manipulations are suggested:

•Duration:
The duration of all short and long vowels is made equal to the mean duration of the short and long vowels in all sentences. The effect of final lengthening of the vowels at the end of the sentence as well as the lengthening of the vowels in lexically stressed syllables will be ignored. The results of the perception experiment will indicate whether it will be necessary to correct for these effects also.

•Intensity:
The intensity of the vowels will be made the same as the mean intensity of the vowels in all sentences. Maybe it is useful to distinguish between open and closed vowels. Then the intensity of the open vowels is made the same as the mean intensity of open vowels and the intensity of the closed vowels is made as intense as the mean intensity of the closed vowels.

•Spectral quality:
All vowels will be reduced to schwa, so that the spectral quality is the same as the schwa spoken by that same speaker. The duration and the intensity of the original vowel must be kept, so only the effect of the spectral quality is taken away from the speech material. The other acoustical features such as intensity and duration are still available to the listener.

The sentences, consisting of one set with manipulated duration, one set with manipulated intensity and one set with manipulated spectral quality, together will form one perception experiment. These manipulated sentences will be mixed and presented to naive listeners, with the task to mark the prominent words.

The results can be put in a correspondence matrices as done in the pilot perception experiment with monotone pitch (for more details see Streefkerk et al., 1997). The 3 manipulation sets will be compared with the results from the perception experiment with monotone pitch. This gives 3 correspondence matrices. The listener judgments of the 3 manipulation sets will also be compared with each other. This gives 3 more correspondence matrices. With the help of these 6 correspondence matrices, the influence of the duration, energy, and spectral information on the perception of prominence are studied and these correlates can be ordered in terms of efficacy.

5. Analyzing the speech signal and classification: Literature survey

5.1. Acoustical analyzes and extraction of features

There are several studies dealing with the automatic classification of accent. In the research of Wightman and Ostendorf (1994), a Markov model is trained to label prosodic patterns. The training and testing data consist of speech material that is hand labeled for prominence and tone boundaries. For the automatic prominent versus non-

prominent detection the overall accuracy is up to 86%.

Kompe (1997) and Kießling (1996) train several predictive models. In their research the initial accent versus non-accent labeling was done automatically based on linguistic information. The predictive model (artificial neural nets, HMM's and hybrid models) are then trained and tested to recognize the accented and non-accented words based on many acoustical features. The number of input features is very high (up to 256 features). The recognition rate is up to 82%. A disadvantage is that nothing is known about the importance of the features. 78% recognition rate for the accented versus unaccented syllables is reached in the research of Storm (1995). In the research of Ten Bosch (1993) a classification was done based on F_0 information only. The recognition rate was up to 81%.

Table 3: A summary of the feature extraction of different studies.

Research of	Time interval	Labeling	Pitch features	Intensity features	Duration features	Lexical features
Wightman and Ostendorf	•syllables	•hand labeled prominente	•max s / mean s_{+1} •max s / max s_{-1} •max s / mean -s •min s / mean s ratio of the final F_0 to the mean F_0 within a sentence	•mean energy in the syllable	•pre-boundary lengthening •$s_{mean\ norm} - s$ •pause duration	•lexical stress •word-final position
Kießling and Kompe	•syllables and words	•automatic labeling on semantic and linguistic information	•mean and /or median •min, max of the onset and offset •the position of these values relative to the end of the syllable •regression coefficient •root mean squared differences between F_0 and the regression line	•mean or median energy •max energy •position of the max energy relative to the end •regression coefficient of energy contour •root mean squared differences between energy and the regression line	•speaking rate •average of the phoneme duration	•class of the phoneme •lexical stress •word-final position
Strom	•10 ms frame	•tone labels similar to TOBI	•interpolated F_0, •3 components of F_0 using different bandpass filters. •derivatives of the 3 functions	•nasal band (30-300 Hz) •sonorant band (300-2300 Hz) •fricative band (2300-6000 Hz)		
Ten Bosch	•vowel onset	•IPO intonation grammar	•+ 60 ms •t_0 – 60 ms •t_0 + 60 ms •t_1 – 60 ms •t_1 + 60 ms			

s = syllable onset time.
s_{-1} = previous syllable onset time.
s_{+1} = next syllable onset time.
t = vowel onset time.
t_{-1} = previous vowel onset time.
t_{+1} = next vowel onset time.
fr = frame.
max = maximum.
min = minimum.
$s_{mean\ norm}$ = the mean normalized duration of syllable duration.
pre-boundary lengthening = pre-boundary lengthening measured by the mean normalized duration of the syllable rhyme.

In the table 3, a list of input features in the studies of Kießling (1996), Kompe (1997), Taylor (1993), Ten Bosch (1993), Storm (1995), and Wightman and Ostendorf (1994) is given. It is described what kind of initial labeling is used to test and train the predictive models. Furthermore a description of the extraction of the pitch features, the intensity features, the duration features, and in some studies the lexical features, is given.

– ***Pitch features:*** In the research of Wightman and Ostendorf (1994) the features of the F_0 contour are calculated in a different way as in the other studies. Per syllable, ratios of the max F_0 to the mean F_0 of the next syllable (max s / mean s_{+1}) and the previous syllable are calculated (max s / max s_{-1}). Further the ratios of the minimum F_0 and the maximum F_0 to the mean F_0 (max s / mean s, min s / mean s) within a syllable are calculated. An additional feature is the ratio of the final F_0 and the mean F_0 within a sentence.

In the research of Kompe (1997) and Kießling (1996) the acoustical features of the F_0 contour are defined in the following way. They use both the syllable and the word as time intervals for the calculation of the acoustical features. The mean or the median, the maximum, the minimum, as well as onset and the offset of F_0 are calculated for each time interval. Furthermore, the regression coefficient of the F_0 contour, and the root mean squared differences between the F_0 values, and the respective values of the regression line are used as features in this research (see also table 3).

Strom (1995) extracts 8 features for the F_0 per 10 ms frame. First of all the interpolated F_0 contour was smoothened in three different degrees using bandpass filters. These 3 components of the interpolated F_0 and their time derivatives are used as acoustical features for detecting accents. The 3 components describe the global or the more local behavior depending on the bandpass filter. The time derivatives give some information about the increase of the interpolated F_0 and its 3 components (see for more detail table 3).

A disadvantage of this approach is that the features are calculated per 10 ms frame so the measurements are independent from the onset of the syllable or the onset of the vowel. It is shown in the research of 't Hart et al. (1990) that the position of the onset of the pitch movement influences the perception of an accent. In the research of Strom (1995) 3 feature for the energy are also calculated (see further intensity features).

In the research of Ten Bosch (1993) the aim was to classify the pitch movements according to the IPO intonation grammar. In terms of the IPO intonation grammar, experts labeled the pitch contour. In Ten Bosch's research the 5 features he uses consist of 5 pitch measurements at different times. The measurements are anchored on the vowel onset (t). The pitch is determined for the following points: t_{-1}+60 ms, t_0–60 ms, t_0+60 ms, t_1–60 and t_1+60 ms, where t_{-1}, t_0 and t_1 denote the vowel onset in the previous, current and next syllable, respectively. In this research the pitch measurements are dependent on the vowel onset, and not only pitch measurements per frame are used, as in Strom (1995), as acoustical features for the classification task. But in Ten Bosch's research the features are only calculated for the pitch, other acoustical features such as energy and duration are ignored.

– ***Intensity features:*** Wightman and Ostendorf (1994) use the mean energy in the syllable as the intensity feature.

In the research of Kompe (1997) and Kießling (1996) the mean or median energy, the maximum energy and the position of the maximum energy relative to the end of the time interval, are calculated. Also the regression coefficient and the root mean squared difference between the energy and the regression line are used as features.

In the research of Storm (1995), 3 energy features (the nasal band 30-300 Hz, the sonorant band 300-2300 Hz and the fricative band 2300-6000 Hz) are calculated per 10 ms frame.

– ***Duration features***: In the research of Wightman and Ostendorf (1994) the pre-boundary lengthening is measured via the mean normalized duration of the syllable. The difference between the mean normalized duration of the syllable and the syllable onset is also determined and used as a input feature. Also the pause duration are measured and used as a feature.
Kompe (1997) and Kießling (1996) use the average normalized speaking rate for one utterance as a feature. The pauses in the utterance are neglected. The average phone duration is also an additional duration feature.

– ***Lexical features***: Wightman and Ostendorf (1994), Kompe (1997) and Kießling (1996) use lexical features as flags, indicating whether a given syllable has lexical stress and whether it occurs in word-final position or not. In the research of Kompe (1997) and Kießling (1996) also use flags to identify the class of the phone being in a syllable nucleus position.

6. Acoustical analysis and feature extraction

6.1. Preprocessing of the sentences

In order to investigate those acoustical features, which lead to the perception of prominence, the word boundaries, the syllable boundaries, and the segment boundaries will be marked in each sentence. For all sentences in the Polyphone corpus the sheet text and a transliteration of the spoken sentences are available. In the transliteration for example mouth noises and breath noises are transcribed by hand. With this information, and the standard pronunciation of the Dutch language in SAMPA notation, an HMM recognizer was trained to localize the segment boundaries. With the help of Xue Wang (Wang, 1997) about 4500 sentences (a subset of 3 CD-ROMS with a total of 900 speakers), were automatically segmented at the phoneme level. This implies that word and syllable boundaries are in principle available. With the help of the written text and the standard pronunciation of the words, word boundaries can be marked in the speech. A set of sonorant-rules will be implemented in a program to mark the syllable boundaries. The sonorant-rules say that each syllable consists of one vowel, and that the consonants around this vowel are ordered with decreasing sonority. The farther a consonant stands away from the vowel the lesser the sonority is. Because there are words which do not behave according to these rules, the syllable boundaries will have to be corrected by hand.

The vowels in the syllables with lexical stress will be specially marked in the label file (see figure 2). Also the position of each phoneme in a word, the positions of each syllable in a word (such as word-final) and the position of each word in a sentence can be estimated from this segmentation file.

6.2. Feature extraction

In our research we intend to give more attention to intensity and duration features, since much is known about pitch movements already. The problem of intensity and duration is that these acoustical correlates are more dependent of intrinsic properties

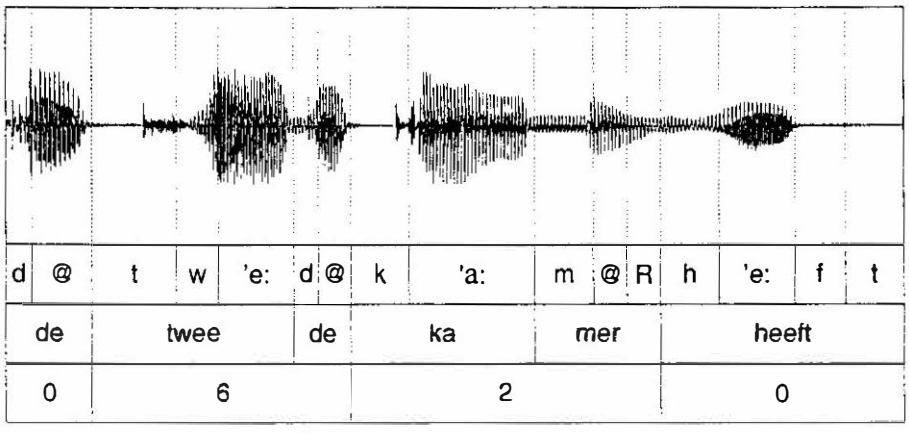

Figure 2: An example of a segmentation file "de Tweede Kamer heeft" (the (Dutch) Lower House has, /də twe:də ka:mər he:ft/). First the segment boundaries, second the syllable boundaries, and third the word boundaries with the prominent score are shown.

such as vowel type, final position and the composition of the syllable. Therefore a normalization is suggested to correct for these effects (Campbell, 1992; Campbell, 1995). All acoustical analyse and feature extraction will be done with the software package "Praat" (Boersma and Weenink, 1996).

In our research we intend to choose the following features:

•Pitch features:
For the pitch per syllable ratios of the max F_0 to the mean F_0 of the next syllable and the previous syllable (max s / mean s_{+1} and max s / max s_{-1}) will be calculated and used as features. The ratio of the minimum F_0 and the maximum F_0 to the mean F_0 within a syllable (max s / mean s and min s / mean s) will be calculated, and presented to an artificial neural network for classification.

•Intensity features:
The mean intensity of the lexically stressed vowel normalized for the vowel type is a good feature to calculate and to use for a predictive model. It might be useful to use the ratios to the next and previous lexically stressed syllables as well. The normalization could not just be done for vowel type but also for lexical stress. For that, the mean intensity for all vowel types in stressed and in unstressed position will be calculated.

•Duration features:
Pause duration between words is a possible duration feature. Also the duration of the syllable will be calculated and used as a feature. The duration of the vowel, corrected by the mean duration of the vowel type is an optional feature. Also for the duration features it is maybe useful to use such features as ratios of the next and previous syllables. The normalization of the duration could not just be done for vowel type (long versus short vowels) but also for position in the sentence. Then a mean duration for all vowel types in final or non-final position is calculated.

6.3. Classification with an artificial neural network

In this project we intend to use an artificial neural network for the automatic prominence classification task. Feedforward nets are already implemented in the software package "Praat" (Boersma and Weenink, 1996). In a pilot study we already trained and tested some feedforward nets to classify prominent and non-prominent words (see Streefkerk et al., 1997). The preliminary results are quite promising.
In future research, the various input features will selectively be added in such a way that they will introduce as much relevant information as possible. We expect that this will increase the recognition rate more than just by introducing a great variety of input features. The relation between the acoustical input features can be studied with the help of an artificial neural network. The trained weights of the artificial neural network can be analyzed and interpreted (see Streefkerk et al., 1997).

7. Discussion and conclusion

In this research, the acoustical correlates of prominence will be investigated. The role of pitch movements on the perception of prominence has so far received much more attention than that of the other acoustical correlates. The relation between pitch movements, duration and intensity has not yet been investigated thoroughly. Acoustical correlates such as energy, duration and spectral information must lead, or at least support, the perception of prominence. Acoustical correlates such as intensity, duration and spectral quality are more influenced by intrinsic segmental properties, such as vowel type, or position in the sentence than the pitch. The loudness of open vowels is different from that of closed vowels. Syllables in final position are generally longer than syllables in non-final position. This effect is known as final lengthening. Spectral quality too is not yet used in the automatic classification of prominent words. The speech material consisting of phonetically rich sentences of many different speakers, is considered to be useful to determine the various acoustical correlates for prominence. This knowledge about the location of the prominent words then can be used in various speech technology applications.

8. Acknowledgment

I would like to thank Louis Pols and Louis ten Bosch for numerous suggestions an comments.

9. References

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

Batliner, A., Kießling, A., Kompe, R., Nieman, H. en E. Nöth (1997). "Can we tell apart intonation from prosody (if we look at accents and boundaries)?", *Proceedings of the ESCA Intonation Workshop*, Athens, 39-42.

Boersma, P. (1993). "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of sampled sound", *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, **17**: 97-100.

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).

Campbell, N. (**1992**). "Prosodic Encoding of English Speech", *Proceedings ICSLP-92*, Banff, Vol. **1**: 663-666.

Campbell, N. (**1995**). "Prosodic Influence on Segmental Quality", *Proceedings Eurospeech'95*, Madrid Vol. **2**: 1011-1014.

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, 1803-1803.

De Pijper, J. R. and Sanderman, A., (**1994**). "On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues", *J. Acoust. Soc. Am.*, **96**: 2037-2047.

Gussenhoven, C. (**1984**). *On the grammar and semantics of sentence accents*, Ph.D. Thesis, Dordrecht: Foris.

'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, 1996.

Kompe, R.. (**1997**). *Prosody in Speech Understanding Systems*, Ph.D. Thesis, *Lecture Notes in Computer Science*, Springer, Berlin, New York 1997.

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

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

Lehiste, I. (**1970**). *Suprasegmentals*, Cambridge, Mass: MIT. Press.

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.

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", *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, **21**: 101-116.

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

Taylor, P. (**1992**). *A phonetic model of English intonation*, Ph.D. Thesis, University of Edinburgh.

Taylor, P. (**1993**). "Automatic recognition of intonation from F_0 contours using the Rise/Fall/ Connection model", *Proceedings Eurospeech'93*, Berlin, 789-792.

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

Terken, J. (**1991**). "Fundamental frequency and perceived prominence of accented syllables", *J. Acoust. Soc. Am.*, **89**: 1768-1776.

Vaissière, J. (**1989**). "On the automatic extraction of prosodic information for automatic speech recognition system", *Proceedings Eurospeech'89*, Paris, Vol. **1**: 202-205.

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

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

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