Speaker characterization in Dutch using prosodic parameters

Kraayeveld, J.; Rietveld, A.C.M.; Heuven, V.J.J.P. van

SPEAKER CHARACTERIZATION IN DUTCH USING PROSODIC PARAMETERS

J. Kraayeveld, A.C.M. Rietveld and V.J. van Heuven

Nijmegen University, Dept. of Language and Speech
P.O. Box 9103, NL-6500 HD Nijmegen, tel: +31 80 612055, fax: +31 80 515939
e-mail: kraayeveld@lett.kun.nl

abstract

In this study the speaker characterising properties of two methods of representing pitch contours were compared. The first is Atal's (1972) approach, in which the entire intonation contour is divided up into 40 segments that form the input to data reduction and analysis techniques.

The second method is a more analytical one, in which the contour is summarized by measurements that are related to 'key points' in the contour. The first approach turned out to yield superior recognition results. However, these results must probably be attributed to differences in the underlying phonological form and not to individual differences in the realisation of this representation.

Keywords: speaker recognition.

1. INTRODUCTION:

The main goal of our research project is to determine the freedom speakers have in their prosodic behavior and the degree to which individual speakers vary on the different dimensions of this behavior.

The acoustic measures of prosodic behavior can be divided into statistical and dynamic ones (O'Shaughnessy, 1986). By averaging parameters over large stretches of time and over many different segments more or less text-independent measures can be obtained, some of which are quite powerful with respect to speaker identification. Jassem et al. (1973), for instance, found that speakers can be identified rather easily on the basis of their mean fundamental frequency and its standard deviation.

Dynamic parameters are measured within their temporal context. This makes them more difficult to apply, since the behavior of different speakers is not timed in a uniform way. Also, these parameters are dependent on the exact lexical content of the utterances they occur in.

Apart from their problematic applicability, from a linguistic point of view dynamic parameters are of more interest than the statistical ones, since they can convey interpretable information on the differences in the behavior of speakers. Furthermore, in using only statistical parameters one throws away speech characteristics that might well contribute to speaker recognition.

Sambur (1975) evaluated a large group of features, both of the statistical and of the dynamic type. He obtained a rank list of measures, the best of which turned out to be spectral ones. Only one prosodic parameter, the fundamental frequency of the word "cash" in the sentence "Cash this bond, please" was among the best 10 parameters of the 92 parameters that were tested.

Atal (1972) tried to identify speakers by their overall intonation contour. To this end he made six recordings of 10 female speakers that produced the sentence "May we all learn a yellow lion roar". To reduce the data to a workable amount, he removed the unvoiced parts of the sentences and divided the voiced part of each sentence in 40 equally large segments. The fundamental frequency was sampled at all these intervals. The amount of data was further reduced by the Karhunen-Loève transformation, a technique closely related to factor analysis. This resulted in a set of 20 KL-components. These accounted for 99.5 % of the total variance. Using five utterances of each speaker to form a reference pattern and the sixth as the test pattern, only two of the 60 classifications turned out to be incorrect.

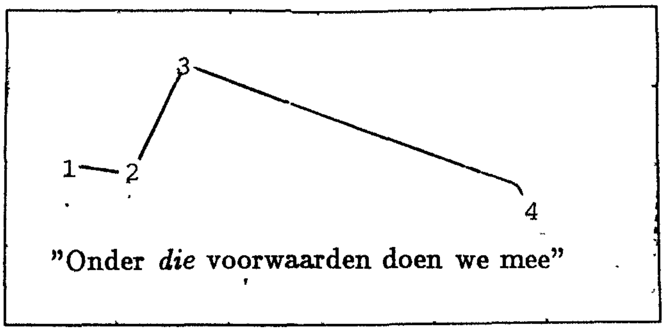
A problem with using intonation contours is that one implicitly assumes that the underlying linguistic form of all utterances is equal. For Atal's study this is not true. If we take a look at the examples of pitch contours he presents, differences can be observed that must be related to different underlying phonological forms.

To be able to use the characteristics of controlled

pitch movements, like their slopes, it is important to control the exact intonation contour of the utterance. It is to be expected that these characteristics are to a smaller extent under voluntary control than the choice of the contours, and may therefore be more speaker specific. In our research project we try to deal with the problem of different underlying phonological forms by selecting stimuli that elicit uniform prosodic behavior (as far as this is possible). This will enable us to use measures of individual pitch movements in a meaningful way.
An advantage of Dutch over other languages is that for Dutch a manageable and fairly simple description of the possible pitch contours is available: the intonation grammar of Collier and 't Hart (1981).
In a pilot experiment (Kraayeveld, Rietveld and Van Heuven, 1990) we found that measures derived from individual pitch contours, like the steepness of F0 rises and falls and declination can contribute to a correct assignment of utterances to speakers.
The aim of the present study is to compare the speaker characterising properties of prosodic parameters that are related to pitch properties in well-controlled utterances to Atal's more wholistic approach of the contours. This approach has proved to be a succesful method of using pitch contour measures for speaker recognition.

2. METHOD:

2.1 *Speech recordings.*
Like in Atal's research, 10 female speakers were selected to read six cards on which the same five sentences were written in different orders. Only one of the sentences was used in the subsequent analyses: "Onder *die* voorwaarden doen we mee" (Under those conditions we go along).
The mean duration of the sentence was somewhat shorter than in Atal's study (about 2 seconds). It lasted between 1.4 and 2.3 seconds. The average duration of pitch periods for the 10 speakers ranged between 4.0 and 6.1 ms., which makes them comparable to the Atal study (4.3 - 5.6 ms.).
The speakers were instructed about the required accentuation of the sentences. The word "die" was to be accented by making rise 1, as described by Collier and 't Hart. Until the word "mee" at the end of the sentence, the pitch was to stay 'high' until the final fall on the word "mee", a so-called boundary tone (cf. Gussenhoven, 1988). All subjects agreed that this accentuation pattern was the most "natural" one. Nevertheless, different speakers used different pitch movements in the part of the sentence between "die" and "mee". Some, for instance, made 'half-falls' (fall 'E') on "doen", or immediately following the rise on "die".
This allowes us to test the hypothesis, that it is especially in these ambiguous parts that Atal's approach optimally separates the subjects.
In the last word of the utterance, "mee", most subjects realised a fall, either as a demarcation of the end of the utterance or as an accentuation movement.
Speech recordings were made in an audio studio. The speech material was digitized at a sampling rate of 10 Hz. Pitch analysis was performed using the algorithm developed by Hermes (1988).



the pitch contour that were used in Method 2.

2.2 *Data analysis.*
The recordings resulted in a set of 60 utterances: 6 replications by 10 speakers. Two methods were applied to these utterances in an attempt to use them for speaker characterisation.
1. Atal's 'wholistic approach'. In what follows, we shall refer to this method as 'Method 1'. As explained earlier, Atal used the entire pitch contours as the input for his analyses. Pitch analysis results in a number of pitch period samples that is too large to handle as individual variables. Therefore, like Atal, we reduced the data by dividing all utterances into 40 contiguous segments (after removing the unvoiced samples) and characterising these by the average value of the pitch samples in the segments. These segments were used as the predictor variables in a multiple discriminant analysis. We did not use the Karhunen-Loève coordinate system, since that would make it difficult to relate the outcome of the analysis to the properties of the pitch contour.
2. Our more analytical approach, to which we will refer as 'Method 2'. The most simple way to summarize a pitch contour appears to be to take measurements at a few 'turning points'. As is shown in Fig. 1, there are four of these points in the sentence used here: the start and end of the utterance,

the starting point of the rise in the accent and the peak of this accent. The measures taken are:

- The pitch at the four measurement points;
- The timing of the measurement points relative to the total duration of the utterance;
- The pitch difference between the starting point and the end point of the utterance and the slope, or declination, of it;
- The pitch difference between the starting point and the end point of the pitch rise and the slope of it;
- The pitch difference between the starting point and the end point of the pitch fall and the slope of it;
- The duration of the utterance.

Prior to making these measurements, the intonation contour was stylized by a program that was developed by Hermes. The stylization allowed us to take measurements in a standardized way.

3. RESULTS:

Two discriminant analyses were carried out on the data, in both of which the ten speakers functioned as 'groups', with six replications per group.

In the first analysis the 40 pitch values of the different sections in which, following Atal, we had divided the sentence served as variables. Nine discriminant functions accounted for 100 % of the variance in the data. The classification of the cases in the 9-dimensional space spanned by the discriminant functions was quite successful: all cases were correctly assigned to the ten groups.

In the second analysis the 13 measures of Method 2 functioned as the variables in the discriminant analysis. On the basis of these variables, six discriminant functions were extracted. The classification of the utterances in the 6-dimensional space spanned by the discriminant functions was correct in 86.67 % of the cases.

There are two possible reasons why Method turned out to be superior in the discriminant analyses: because it uses more variables, and because of the large number of discriminant functions. Therefore the analysis was repeated while allowing for only three discriminant functions. Again, Method 1 was clearly superior to Method 2. The percentage correctly attributed cases was 83.3 for the first method and 75 for the second.

For a better understanding of the correlations of the functions that were extracted in the first analysis with the various discriminating variables, Fig. 2 shows the means of the 40 segments and the mean squares between-mean squares within groups ratio, a measure for the speaker characterizing properties of the segments.

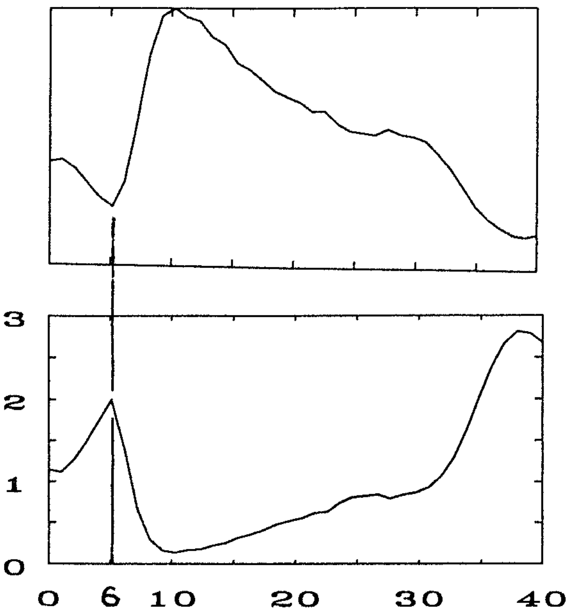


Fig. 2: (a) Means of the 40 segments in Method 1, and (b) Mean squares between groups devided by mean squares within groups for these segments.

The first function correlates highly with segment 20 to 34. The second function seems to be related to the segments 7 to 11, where the pitch rise in the accent takes place.

Many other functions correlate with some part of the contour; e.g. function 4 correlates with the first three segments.

The first function of the second analysis has much to do with the pitch of the starting point of the rise in the accent. The pitch at the start of the utterance is related to this function too. The second function correlates with the total utterance duration. The third function is concerned with the realization of the rise-fall accent; the pitch difference between the start and the end of it, the pitch at the top of the accent and the slope of the accent correlate highly with the function. The fourth and fifth function correlate with measures that are related to the declination line and the last function correlates with all measures that are concerned with the timing of the pitch accent.

It is interesting to note, that these observations do not replicate the finding reported by other researchers (e.g. Liberman and Pierrehumbert, 1984) who suggest that the low pitch values at the end of an utterance are particularly promising speaker-specific measures. In our data this is more true for the pitch at the starting point of the rise in the accent.

4. DISCUSSION:

The outcome of this study confirms Atal's finding that ".. pitch contours can be used effectively for speaker recognition".

Atal represented pitch contours by the mean pitches of 40 contiguous speech segments, into which the voiced parts of the contour had been divided. With respect to speaker recognition this method was superior to one in which the contours were represented by the pitch values, timing and slopes of some 'turning points' in the contour.

Although Atal's method leads to better classification of utterances to speakers and will therefore be the preferable method when it comes to speaker recognition, it has some important drawbacks. The main problem of this approach is the very global description of the contour that results from it. This makes it unsuitable for the investigation of the different ways in which speakers realise individual pitch movements in their utterances. The large amount of data that have to be handled can be a problem too.

In this study, we found the best speaker characterising properties for the pitch values of the segments in the second half of the utterance. As was explained in Section 2, different speakers realised this part of the utterance, the part following the accent on "die", in different ways. They chose for different underlying phonological forms which lead to differences in the pitch contours.

In utterance parts where subjects followed different phonological patterns, consistency in choice for one of the patterns will lead to relatively large inter-speaker variance, as compared to intra-speaker variance. Thus, speaker recognition is high, but results from different underlying forms, and not from speaker differences in the realisation of *specific pitch movements.*

By analysing utterances using predefined 'turning points', it is easier to investigate the speaker differences in pitch movements. The discriminant analysis that was presented for this second method shows, that classification of utterances to speakers is still rather good. Method 2 was particularly concerned with measures taken at the pitch accent. In the analysis of Method 1 this part of the utterance turned out to be closely related to the second-best discriminant function, indicating that at the level of individual pitch movements, speaker characteristization can take place too (cf. Fig. 2 where a peak in the MSbetween/MSwithin ratio at segment 6 shows the large contribution of this segment to correct utterance classification.

An important feature to note about Method 2 is that the resulting discriminant functions all showed a clear functional coherence. This strengthens our conviction that this method is useful in the study of speaker characteristics at the level of individual pich movements.

REFERENCES:

Atal, B.S. (1972) Automatic Speaker recognition based on pitch contours, *JASA*, 52, 1687-1697.

Collier, R. & 't Hart, J. (1981) *Cursus Nederlandse Intonatie.* Leuven: Acco.

Gussenhoven, C. (1988) Adequacy in Intonation Analysis: The Case of Dutch. In H. van der Hulst & N. Smith (eds), *Autosegmental studies on pitch accent.* Dordrecht: Foris Publication, 95-121.

Hermes, D.J. (1988) Measurement of pitch by subharmonic summation, *JASA*, 83.1, 257-264.

Jassem, W., Steffen-Batog, M. & Czajka (1973) Statistical characteristics of short-term average F0-distributions as personal voice features. In W. Jassem (ed), *Speech analysis and synthesis, Vol. 3*, 209-228.

Kraayeveld, J., Rietveld, A.C.M & Heuven, V.J. van (1990) Prosodic Speaker characteristics in Dutch. In J. Laver, M. Jack & A. Gardiner (eds), *Proceedings of the tutorial and research workshop on Speaker characterization in speech technology*, Edinburgh, European speech communication association, 135-139.

Liberman, M. & Pierrehumbert, J. (1984) Intonational Invariance under Changes in Pitch Range and Length. In M. Aronoff & R. Oehrle (eds), *Language Sound Structure.* Cambridge (Ma.): MIT Press, 157-233.

O'Shaughnessy, D. (1986) Speaker Recognition. *IEEE ASSP Magazine*, 4-17.

Sambur, M. (1975) Selection of acoustic features for speaker identification. *IEEE Trans. ASSP*, ASSP-23, 176-182.