Speaker specificity in prosodic parameters[1]

J. Kraayeveld, A.C.M. Rietveld and V.J. van Heuven[2]
Dept. of Language and Speech, Phonetics Section, Nijmegen University
P.O. Box 9103, NL-6500 HD, Nijmegen, The Netherlands

ABSTRACT

Ten time-integrated prosodic parameters were used to assign read-out and spontaneous speech fragments to the speakers that had produced them. Fifty speakers of standard Dutch participated in the experiment. They were stratified by gender and age. It was found, that the parameters are independent of each other in terms of speaker specificity. In combination, they could correctly assign 73 % of 500 fragments to the 50 speakers. As expected, differences in mean F_0 were important for speaker identification, but an analysis without mean F_0 still resulted in 56 % correct classification. Identification improves when spontaneous and read-out speech are analysed separately. Thus, speech type is an important factor to control for in speaker identification.

INTRODUCTION

Research on speaker specific variation is often directed at applications of speaker recognition (e.g. in forensic research) and speaker verification (e.g. for electronic access systems). However, it is also a necessary step in the process of separating the contrastive and linguistically meaningful properties from speaker-dependent and meaningless variation. In this contribution results are reported of a research project that aims at mapping out individual variation along *prosodic* parameters in Dutch.

First, in carefully selected sentences prosodic parameters were measured that are closely related to the linguistic and prosodic structure of the utterance, e.g. F_0 at specific turning points in an utterance, such as the top of a pointed hat-pattern. These measures we call *point* measures.

Not all prosodic parameters require strictly controlled utterances. By averaging over larger stretches of time and over many different segments, some can be made more or less text-independent. Examples of these *time-integrated* measures are mean F_0, certain temporal measures, and perturbation measures (measures for the instability of a speaker's frequency and amplitude).

In our research project the usefulness of both the point measures and the time-integrated measures were studied. Preliminary results on speaker dependent characteristics of such point measurements were reported earlier in Kraayeveld *et al.* (1991). In the present contribution, however, we examine the possibility of separating individuals using non-linguistic, time-integrated prosodic measures only.

In ten 15 s.-fragments of both read-out and spontaneous speech ten different time-

[1] This research was supported by the Linguistic Research Foundation (projectnr. 300-173-006), which is funded by the Netherlands organization for scientific research, NWO

[2] Dept. of Linguistics, Phonetics Laboratory, Leiden University, The Netherlands

integrated prosodic measures were used. Some of these measures have been found earlier to be very powerful tools in speaker identification. Especially mean F_0 and the standard deviation of F_0 appear to be very useful (e.g. Sambur, 1975).
From segmental studies we know that there are large differences between read-out (or premeditated) and spontaneous speech (eg. Van Bergem *et al.*, 1989). We anticipate similar, large effects of speech style on the use of prosodic parameters. Still we would like to know to what extent a single individual displays the same prosodic behaviour in both speech styles.
One, almost trivial, speaker difference that is reflected in prosody, specifically in mean F_0, is gender: female voices have about double the F_0 of males. Therefore our research was set up to examine speaker dependence of prosodic parameters both across genders and for male and female groups separately. We want to find out if individuals can be successfully discriminated when the sex of the speaker is partialled out, and on the basis of which parameters.
In the experimental design the factors gender, age[3] and speech type will be controlled.

METHOD

Speakers: A sample of 50 speakers of standard Dutch was selected that was stratified for gender and age: 25 male and 25 female speakers, belonging to the age groups 18-25, 26-35, 36-45, 46-55 and 56-65. Thus, each age groups consisted of five males and five females each.
Elicitation: In the first task speakers were interviewed on everyday issues. After editing out irrelevant material such as interviewer intrusions, an otherwise contiguous stretch of speech of 75 s. was selected from the end of each recording, and divided into five stretches of 15 s. each.
The second task was to read out a newspaper-like story. This story consisted of five paragraphs. Of each paragraph, the first 15 seconds were included. Thus, the read-out material roughly contains the same lexical material.
Analysis parameters: the following time-integrated parameters were determined:

1 F_0MEAN	Mean F_0 in Hz.
2 F_0COV	Coefficient of variation of F_0, the standard deviation of F_0 divided by its mean)
3 PPQ	Pitch Perturbation Quotient, as defined by Davis (1976)
4 PZR	Pitch period Zero-crossing Rate, the percentage of triplets of adjacent periods where duration does not increase or decrease monotonically
5 AMPCOV	Coefficient of variation of the absolute peak-amplitude per period
6 APQ	Amplitude Perturbation Quotient, analogous to PPQ
7 AZR	Amplitude Zero-crossing Rate, analogous to PZR
8 VOICE	Percentage of time the signal is considered Voiced
9 PAUSE	Percentage of time the amplitude is below threshold
10 RATE	Speaking Rate, number of syllables per fragment.

[3] There are no clear indications that age is related to speaking behaviour, at least not in the 18-65 yr. age bracket. Although our sample was stratified by age (see method) we shall not study the influence of age on the question of speaker separation.

RESULTS AND DISCUSSION

To determine how well these ten prosodic measures can be applied to the task of speaker identification, five discriminant analyses were carried out. Discriminant analysis is primarily a data reduction method, in which parameters are collapsed onto orthogonal discriminant functions such that the functions maximally separate the groups. Discriminant functions are linear combinations of variables in which the weights reflect the importance of the associated variables.

In the first analysis all 50 speakers functioned as 'groups', with 10 data points (five 15s.-fragments of two speech types) per group. Next, the analysis was repeated with parts of the data set: the read-out and the spontaneous fragments (50 groups, 5 data points each), and the sets of female and male speakers (25 groups per analysis, 10 data points).

All discriminant analyses resulted in 10 significant discriminant functions (the maximum possible number given 10 variables). To show the influence of reducing the number of dimensions from 10 down to one, the analyses were repeated, limiting the number of discriminant functions. Figure 1 displays the percentage speakers that are correctly classified as a function of the number of discriminant functions. Although the percentage of classified fragments does not improve substantially when more than seven functions are included, the further discriminant analyses will be based on all ten functions. This enables us to compare these results with analyses from which F_0 was excluded. Below, in Table 1, for all five analyses the variables are specified that have the highest correlation with the first three (Varimax-rotated) functions. From this table it becomes clear, that mean F_0 is the most important prosodic variable for speaker characterisation. Apparently this is not only the result of the obvious fact that men and women differ considerably on this parameter, since in separate analyses of men and women mean F_0 was the most important discriminating variable as well.

It also becomes clear, that it is not *only* F_0 that contributes to the classification of the individuals. If only F_0 is allowed as a variable in the analysis, the amount of correct classification is small, ranging from 9 % for males, to a value of 32 % for read-out speech (male and female speakers). Apparently, for F_0 there is an interaction between the factors speaker and speech style. The behaviour of speakers in the two speech styles is different, and can therefore be better classified if only one speech style is taken into consideration. Another way to study the role of F_0 is to exclude it from the analyses. If the maximally possible number of functions (nine) are allowed, the percentages of correct classification of the fragments for read out and spontaneous

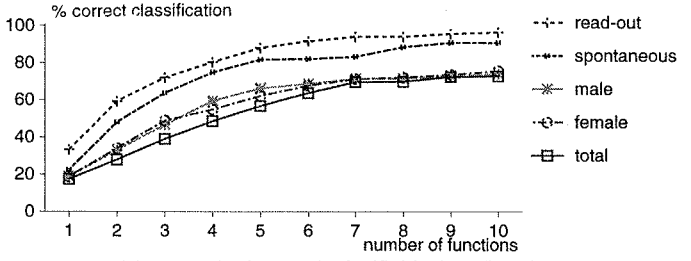

Figure 1: Percentage of fragments that is correctly classified in the *n*-dimensional space spanned by the discriminant functions

Table 1: Correlations between the three most important rotated discriminant functions and prosodic variables that exceed .50. Below the percentage correct classification for each analysis, the percentage correct classification is shown if F_0 is kept out of the analysis (resulting in an analysis with nine discriminant functions), and if only F_0 is allowed (resulting in one discriminant function):

	total	read-out	spontaneous	females	males
f. 1	F_0MEAN .87	F_0MEAN .97	F_0MEAN .95	F_0MEAN .87	F_0MEAN .84
f. 2	PZR .85	PZR .95	PAUSE .91	PZR .88	F_0COV .90
f. 3	F_0COV .92	F_0COV .95	PZR .92	F_0COV .93	PPQ .92
cor. class.	73 %	96 %	91 %	76 %	74 %
cor. class. without F0	56 %	93 %	82 %	58 %	62 %
cor. class., only F0	9 %	32 %	21 %	11 %	9 %

speech are somewhat lower than in the analysis with all variables. However, when we exclude mean F_0 from the separate analyses of male and female speakers, the percentage correct classification decreases considerably.

In summary, the differences between the speech types (i.e. read-out and spontaneous speech) appear to blur the speaker differences to some extent. If we compare the analysis of the total material with analyses of only parts of the data, we find that restricting the analysis to only one speech type improves the percentage correct classification more than analysing only one of the genders. Actually, the differences between the two speech types are so large, that in a discriminant analysis with the speech types as groups, 93.8 % of the fragments is correctly assigned to the speech types. The two parameters that correlate most with the only possible discriminant function are VOICE (.55) and AMPCOV (.52). Mean F_0 does not play any role in this function (-.10).

An analysis with the two genders as groups yields about the same percentage of correct classification as in the analysis of the speech styles: 98.4 % of the fragments was assigned to the correct gender.

REFERENCES

Bergem, D.R. van & Koopmans-van Beinum, F.J. (1989), "Vowel reduction in natural speech", *Proceedings of EUROSPEECH '89,* Paris, 285-288.

Davis, S.B. (1976), "Computer evaluation of laryngeal pathology based on inverse filtering of speech", *Speech Commun. Labs, Monogr.*, 13.

Kraayeveld, J., Rietveld, A.C.M. & Heuven, V.J. van (1991), "Speaker characterization in Dutch using prosodic parameters", *Proceedings of EUROSPEECH '91,* Genova, 427-430.

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