



MATHEMATICAL ASPECTS OF THE CLASSIFICATION OF BASIC PITCH PATTERNS

G.Surmanowicz-Demenko*

ABSTRACT

A method of describing, analyzing and classifying fundamental frequency courses in speech is presented. For the analysis of variability of the F_0 parameter, the Karhunen - Loève transformation was used. In order to study the differences between the curves, a discriminant analysis was employed. The results of an automatic analysis demonstrated the possibility of describing time-variable F_0 as representing the following typical intonations: Low Rise, High Rise, Full Rise, Low Fall, Full Fall, Level, Low Rise-Fall and Full Rise-Fall in a system of 3 coordinates. A deterministic classification algorithm was developed. The training set included F_0 curves which had been judged to be correct imitations of prototype intonations in perceptual tests. The test set consisted of 360 imitations randomly selected from a collection of 1200 and 80% correct classification was obtained.

PREPARATION OF THE SIGNAL FOR THE ANALYSIS

The choice of a method for measuring fundamental frequency in speech, processing the results and correcting them depends heavily on the aim of the analysis. Accepting that the relation

$$F_0 = \frac{1}{T_0} \quad (1)$$

always holds however T_0 is defined, we can describe the F_0 parameter in different ways. In a short-term representation T_0 is defined as the average length of several successive periods. The manner of averaging depends on the particular method (ref 1). In the final stage of processing the results of the measurements, it is desirable to have procedures which permit a definition of the minimum of points representing a given curve. The criterion which defines the appropriate number of data may be formulated as follows (ref 2).

$$J_n = \frac{1}{2} \frac{\sum_{i=0}^n [R(t_{2i}, t_{2i}) - R(t_{2i}, t_{2i-1}) - R(t_{2i-1}, t_{2i}) + R(t_{2i-1}, t_{2i-1})]}{\sum_{k=0}^n [R(t_{2i}, t_{2i}) + R(t_{2i-1}, t_{2i-1})]} \quad (2)$$

where R_{t_1} is the autocorrelation function.

If $J_n \ll 1$ for the given number n of samples, it can be assumed that a suitable representation of the curve has been obtained. In the case of an analysis of F_0 , it can be shown that using this criterion leads to substantial initial data reduction. The F_0 courses may differ within intervals where they can be assumed to be a continuous time function by their duration and the incidence of interruptions, their location relative to time and frequency, as well as the rate and direction of change of the instantaneous values. When normalizing frequency a log scale should be used, but the debatable point is the reference value, which may be, for instance, the minimum, the mean, the middle of the range, or alternatively, a unit variance may be imposed.

*Acoustic Phonetics Research Unit, Institute of Fundamental Technological Research, Polish Academy of Sciences, Poznan.

Also time normalization is a complex problem. One method of matching two signals which differ in duration is time warping. After such frequency and time normalization, the distance between the curves should be minimum on the adopted criterion.

SELECTION OF FEATURES CHARACTERIZING THE F_0 CURVES

For any recognition procedure, the selection of features is critical. The Karhunen - Loève method (ref 3) is optimal with respect to the description of data because the mean square error of the approximation is less than in other transformations. The co-ordinate system is described by the eigenvectors and the eigenvalues of the covariance matrix:

$$R = \Phi D_\lambda \Phi' \quad (3)$$

where D_λ is the diagonal matrix of eigenvalues, and the transformation matrix Φ is a column matrix of orthogonal eigenvectors. The eigenvalues of the matrix R form a non-increasing sequence and the eigenvectors are arranged similarly. The product of the eigenvectors Φ of a real symmetric matrix R with the vector of the object X_i gives the vector C_i such that

$$C_i = \Phi' X_i \quad (4)$$

whose co-ordinates are uncorrelated and arranged according to decreasing variance. But the result obtained with the K-L method is not optimal with respect to its discriminant properties.

In order to find spaces that are optimal with respect to its discriminant properties, the Fisher criterion (ref 4) is used :

$$F = \frac{l' B l}{l' W l} \quad (5)$$

where l is the discriminant vector we seek. The optimization of the transformation reduces to the solution of the eigenvector problem

$$(B - \lambda W) l = 0 \quad (6)$$

The matrix B is the between-classes covariance matrix and W is the within-class covariance matrix. If the parameters of the random variable X observed in the population $\pi_i (i, j=1..k)$ are known, we can define the matrices B and W

$$B = N^{-1} \sum_{i=1}^k \sum_{j=1}^k N_i N_j (\bar{x}_i - \bar{x}_j) (\bar{x}_i - \bar{x}_j)' \quad (7)$$

$$W = \sum_{i=1}^k \sum_{j=1}^{N_i} (x_j^i - \bar{x}_i) (x_j^i - \bar{x}_i)' \quad (8)$$

The linear combinations

$$u_i = l_i' x \quad (9)$$

are termed discriminant variables.

By way of an illustration, Fig.1 presents the results of a K-L analysis and a discriminant analysis performed on the F_0 parameter within a fragment of an utterance (the syllable sequence / $\text{d}\text{z}\text{e}\text{m}\text{i}$ / repeated by 10 speakers (Fig.1a). The plots show interspeaker differences. When comparing Figs.1b and 1c, it can be observed that both analyses result in similar configurations. The discriminant analysis can be seen to emphasize the distances between speakers. In order to define (a) the difference between various F_0 curves and (b) the number of features necessary for their classification, a discriminant analysis of some typical Polish pitch patterns has been performed.

The phonetic material included eight utterances realizing the following types of intonation: Low Rise, Full Rise, High Rise, Low Fall Full Fall ,Level, Low Rise-Fall and Full Rise-Fall, each repeated 10 times by 15 speakers. For the purposes of the discriminant analysis, the materials were selected from recordings of speakers whose performance was found, in a formal perceptual test, to be similar to that of a prototypical voice. The results of the analysis are shown in Fig.2. The analysis leads to the following conclusions:

- The classes under investigation can be described in a 2-D space with 89% adequate distances between them and in a 3-D space with 99% adequate distances.

- All the distances between the classes are statistically significant. An analysis of the correlation coefficients between the discriminant variables and the original variables shows that it is possible to interpret the 1st variable as the ratio of the final to the initial value of the curve and the 2nd variable as the initial value alone. A third variable, which is the derivative in the extremum makes it possible to define the convexity or concavity of the curve.

It is necessary to check whether the three features actually permit a satisfactory classification of the curves.

CLASSIFICATION

One of the basic algorithms employed in the deterministic method of classification is the 'perceptron algorithm'. The decision functions are generated from patterns supplied to the computer using an iterative learning algorithm. It is assumed that for M classes $\omega_1 \dots \omega_M$ there are M linear discriminant functions. We assume that in the k-th iterative step during the learning phase the pattern Y_k belonging to class ω_1 is inputted. The decision rule is defined as follows (ref 2):

$$W_i^1 Y > W_j^1 Y \quad j=1, 2, \dots, M \quad j \neq i \quad \text{for } Y \in \omega_i \quad (10)$$

An algorithm for adjusting W then becomes

$$\begin{aligned} 1) \text{ If } & W_i^1 Y > W_j^1 Y & j=1, 2, \dots, M \quad j \neq i \quad \text{for } Y \in \omega_i & (11) \\ \text{then} & & & \\ & W_{k'} = W_k & k=1, 2, \dots, M & \end{aligned}$$

$$\begin{aligned} 2) \text{ If } & W_i^1 Y > W_i^1 Y \text{ and } W_i^1 Y > W_j^1 Y \text{ for } Y \in \omega_i & (12) \\ \text{then} & & \\ & W_i^1 = W_i^1 - cY & W_i^1 = W_i^1 + cY & W_j^1 = W_j^1 \end{aligned}$$

where c is the correction coefficient. Fig.3 shows the results, in the form of a printout, of the classification of F_0 curves for one of the speakers. Fig.4 depicts percent scores for 360 F_0 curves randomly selected from a collection of 1200 (150 for each of 8 patterns), broken down according to the speaker. The average score was 80%.

REFERENCES

1. W Hess, Pitch Determination of Speech Signals (Springer) (1975) p475
2. K Fukunaga, Introduction to Statistical Pattern Recognition (Academic Press, New York) (1972) p244
3. JT Tou, RC Gonzales, Pattern Recognition Principles (Addison-Wesley Publishing Company) (1974) p271
4. PA Lachenbruch, Discriminant Analysis (Hafner Press) (1975) p66.

