MULTIDIMENSIONAL ANALYSIS OF THE SIMILARITY OF PITCH CONTOURS

GRAZYNA DEMENKO

Acoustic Phonetics Research Unit
Institute of Fundamental Technological Research
Polish Academy of Sciences
Noskowskiego 10, 61-704 Poznan, Poland

ABSTRACT

In order to find the relations between the physical and perceptual analysis of fundamental frequency, a number of listening tests were performed and evaluated by means of Multidimensional Scaling. The experimental materials consisted of utterances with eight different intonation patterns. On the basis of results obtained from automatic pitch pattern recognition, such cases were selected as would represent (1) a 100% recognition (2) fair recognition (about 50% correct) (3) poor recognition (about 20% correct). The listening panel judged the proximity between the elements in each case including two replications of each of two patterns. The purpose of the experiment was (a) to establish the perceptual dissimilarities between the patterns (b) to create a basis for the classification and (c) to compare the results of an objective and a subjective analysis.

1. INTRODUCTION

The analysis of prosodic features takes a significant position in an acoustical and a perceptual description of the speech signal. The F0 parameter (the fundamental frequency) is the subject of much theoretical and experimental work. Experimental investigations may be performed at the perceptual or the physical level. A selection of just one of them does not ensure proper analysis procedures. Perceptual experiments may be objected to on the grounds of subjectivity. On the other hand, purely instrumental analysis may lack a clear relation to linguistic entities. As it is generally accepted that variations of fundamental frequency produce, at the perceptual level the sensation of tone height, a psychoacoustic analysis of this parameter appears to be very much to the point. Temporal variations of fundamental frequency are due to a number of effects that vary themselves during an utterance. It is essential. for the analysis of this parameter, to define which of the many possible sources of variability are effective in a given case.

In [3], the various sources of variability of F0 were briefly discussed. If it is desired that most of the manifold variabity sources be kept out, the experimental material should include only simple utterances. An analysis of more complex melodic structures requires a prior discrimination of the functional units of intonation. The present work attemps to find possibilities of evaluating the physical and the perceptual similarities between various simple pitch curves and to classify the curves on the basis of a limited set of prototypical natural Polish utterances.

2. PREPARATION OF THE EXPERIMENTAL MATERIAL.

The Polish phrase "Dobrze" = approx."all right" was uttered by a phonetician with 8 different intonation patterns. The utterances were recorded at intervals of approx. 5 s. The patterns (treated as prototypes) were reproduced over loudspeakers to be immediately and without reflection imitated by the test person who was always asked just to repeat what he or she had heard, as naturally as possible, with their own natural voice, without any attempt to mimick. 15 native speakers of Polish (10 males and 5 females) were used as test subjects. Three of them had previously been exposed to professional phonetic training. The reproduction of each of the 8 Prototypes: (1) Low Rise, (2) Full Rise, (3) High Rise, (4) Low Fall, (5) Full Fall, (6) Level. (7) Low Rise Fall, (8) Full Rise Fall was performed in several sessions, altogether 10 times by each subject.

3. A MULTIDIMENSIONAL STATISTICAL ANALYSIS.

A fundamental problem in any recognition procedure is the selection of characteristic features. A method which is optimal with respect to data description uses eigenvectors of the covariance matrices (the Karhunen - Loeve method). It was used for data reduction in F0 curves. e.g., by ATAL ([1]). But the aim of recognition is a discrimination of classes, so better possibilities are offered by subspaces constructed on the basis of discriminant vectors.

The problems of discriminant analysis are presented in a number of publications, e.g. [5]. The aim of a discriminant analysis is to find a subspace in which the total dispersion of the data collection will be maximum relative to the within-class dispersions.

It was assumed that using the discriminant analysis it would be possible to examine differences between F0 curves, to define the features necessary for their correct discrimination and to establish the possibility of their classification.

In order to eliminate differences caused by varying pitches of the individual voices, frequency normalization was performed. The logarithm of the lowest value was substracted from the logarithm of successive frequency values within a curve. Then, the difference between the means for the frequency variation ranges of the given voice producing the prototypes was added or subtracted leading to the desired relation among the reproductions as well as between these and the Prototypes.

In order to normalize for time, as well as reducing data, each utterance was divided into 8 parts within which average frequency was calculated as the reciprocal of mean period length. It was accepted that the Prototype utterances and their 10 replications by each of 3 of the imitators (the phoneticians) will form the classes to be examined. The pitches of these voices differed: the lowest frequency for the two male voices was 65 Hz and that for the female voice, 160 Hz. Each of the individual classes was thus represented by 30 replications. Fig.1 depicts a classification tree over the mean vectors of the classes under examination. The values of the Hotelling T^2 statistic are placed over each of the connecting lines. By comparing these with the critical value at the 5 percent significance level, (which was 88.35) it was found that all distances between the classes are statistically significant. The performed analysis leads to the following conclusions:

(1) the classes under examination may be defined in a 2-D space with 90 percent correct distances between them or in a 3-D space with 99 percent accurancy in the distances

(2) the differences between the individual classes are all statistically significant.

3. CLASSIFICATION.

As the features corresponding to the discriminant variables represent an optimal set with respect to recognition, a description of the F0 curves in terms of these features appears desirable.

The first discriminant variable is interpretable as the slope of a straight line passing through the initial and the terminal point of the time and frequency normalized F0 curve.

The second characteristic of the set under examination was defined as the initial frequency value of the curve. Although a satisfactory description of the curves was obtained with only two variables, a third variable (see above) will slightly improve the classification. It is related to the degree of convexity or concavity of the curve. A still more precise description may be obtained when a fourth feature is introduced, viz. the location of the extremum.

One of the basic methods used in deterministic classification is referred to, in recognition literature, as the "perceptron algorithm", with the decision functions generated from patterns provided for the computer by an iterative learning algorithm. The coefficients of the decision function have here been defined as follows:

It was assumed that there exist M decision functions having the property that if $x \in \omega_i$, then $d_i(x) > d_j(x)$ for all $j \neq i$, x being the vector to be recognized and ω_i being the class ω_i. Let us consider M classes $\omega_1 \ldots \omega_M$ and assume that in the kth iterative step during the learning stage the pattern x belonging to class ω_i is presented to the computer. The decision functions

$d_j[x(k)] = W_j'(k)\, X(k)$ are calculated, and if

$d_i[x(k)] > d_j[x(k)]$ for $j = 1,2,\ldots M$ and $j \neq i$, then the weighting vector W_j remains unaltered in the next iterative step:

$W_j(k+1) = W_j(k)$ for $j = 1,2..M$.

Otherwise, the weighting vector is altered in accordance with the relation

$W_i(k+1) = W_i(k) + cX(k)$

else

$W_l(k+1) = W_l(k) - cX(k)$

where c is a constant. If the classes are linearly disjoint, then the algorithm is convergent in a finite number of iterations for an arbitrary initial weighting vector.

Our learning set included all the replications produced as imitations of the Prototypes by the three phonetically trained subjects. For each speaker, 3 replications of each of the eight classes were selected at random and subjected to the recognition procedure. Fig. 2 presents the results which were in agreement with the assumed classes 80 percent of the time. This suggests that the algorithm should be modified by using a greater number of features. As the classes turned out not to be in fact linearly disjoint, an alternative type of decision functions may be preferable. Fig.2 also shows the results of recognition of the entire collection of 1200 curves using a different method, viz. quadratic statistical discriminant functions. The method and the results will not be here discussed (see the companion paper by W.JASSEM presented at this Congress) except for mentioning that 8 features were used there. But it is noteworthy that though the deterministic algorithm yielded distinctly poorer results, both methods

divided the test subjects into identical three groups of very good imitators (LR, JI, WJ, WI) good ones (AM, HK, KK, BS, MC) and bad ones (TK, RC, MK, CW, BI, PD).

4. PERCEPTUAL ANALYSIS.

The advances in methods of computation and optimization of the recent 15-20 years permitted the development of a method of evaluating the results of perceptual experiments known as Muldidimensional Scaling [4]. Its aim is to find a configuration of n elements such that the distances between them should correspond to subjective dissimilarities between observed objects. A monotonic relation between the distances and the dissimilarities is required. The concept of stress is introduced to reflect the measure of non-monotonicity, i.e., of the error in the approximation of the experimental data. Except for degenerated systems, the stress is minimum for the optimum configuration. The quality of the configuration is generally described as very good if the stress is 5 percent or less, good if between 5 and 10 percent and acceptable up to 20 percent. An extensive study of the psychological process involved in the perception of tone in speech was preesented by Gandour [2] On the basis of results obtained in Multidimensional Scaling, the author accepted two features as being characteristic: the mean frequency and the direction of pitch movement. He confirms the stability of these features and concludes that other dimensions are difficult to interpret.

The purpose of the listening experiment to be reported here was to seek the answer to the following questions: (1) Are some of the different intonations perceptually similar? (2) Do the listeners consistently use the similarity measures ? (3) Is there a systematic relation between perceptual similarity and some physical features of the pitch curves?

By reference to the results of automatic recognition of the intonations (the deterministic model), listening tests were prepared which consisted of the utterances of one very good imitator (WJ), one good imitator (MC) and one bad imitator (TK, see above). A panel of 20 listeners (all university students) judged the similarities between pairs of stimuli. 2 replications were randomly selected for each of the three voices and each of the 8 intonation patterns producing for each voice a collection of 136 stimulus pairs. The listeners judged the similarity between the members of each pair on a scale of "0" to "4",with an increase of the rating reflecting the measure of similarity. 28 pairs of stimuli were administered in one session. The measure of similarity between any two stimuli was defined as the sum of the ratings obtained from all listeners.

The results of the test are presented in Fig.3, with the second selected replication indicated by a prime. In a 2-D space it can be seen that for voice WJ the replications form distinct clusters, that for voice MC the utterances 2, 2', 3, and 3' form a single cluster whilst the other replications group together, and that for TK there are five clusters: (1) 2, 2', 3, 3', (2) 4, 4', 7, 7', (3) 8, 8', 5, 5', (4) 1, 1', (5) 6, 6'. In order to show how these perceptual results are related to the physical properties of the stimuli Fig. 4 a presents the 10 replications of patterns 2 and 3 as produced by MC whilst Fig.4 b shows the replications of patterns No. 4, 5, 7 and 8 as produced by voice TK. It is clear from Fig. 4 that the intonations that are confused in perception are also indistinguishable as F0 curves. The two perceptual **dimensions** obtained in the present study may be described as relating to the steepness of the curve (the first dimension, i.e. the strongest distinctive feature) and to the terminal pitch (the second dimension, i.e. the weaker distinctive feature).

CONCLUSIONS

1. Both the automatic and perceptual analysis permitted a classification of the F0 patterns.
2. The set under examination can be described using a few features. The first two are statistically and perceptually most significant.
3. A final automatic classification of the intonation curves requires more stringent methods.
4. Perceptual classification would be improved by considering differences between individual listeners (INDSCAL).

REFERENCES

1. ATAL, B.: Automatic Speaker Recognition Based on Pitch Contours, JASA, vol.52, No. 6, 1687 - 1697, 1972.
2. GANDOUR, J.T.: Perceived Dimensions of 13 Tones: A Multidimensional Scaling Investigation, Phonetica, vol. 35, No.3, 169 - 180, 1978.
3. JASSEM, W., DEMENKO,G.: On extracting linguistic information from F0 traces, Studies of Intonation in Discourse, (C.Johns-Levis ed.) London, 1984.
4. KRUSKAL, J.B.: Nonmetric multidimensional scaling: a numerical method, Psychometrika, vol.29, NO.2, June, 115 - 129, 1964.
5. LACHENBRUCH, P.A.: Discriminant analysis, Hafner Press, New York, 1975.

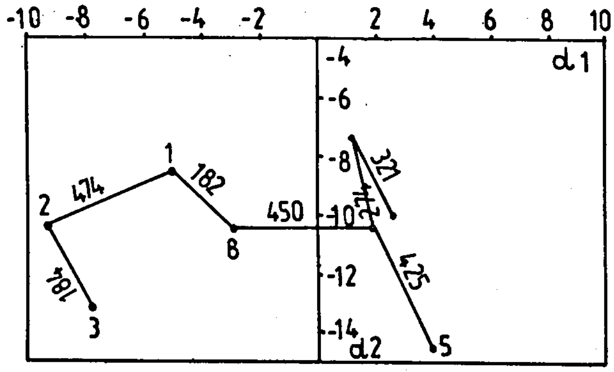

Fig.1. Mean vectors of the 8 classes in a coordinate system of discriminant variables.

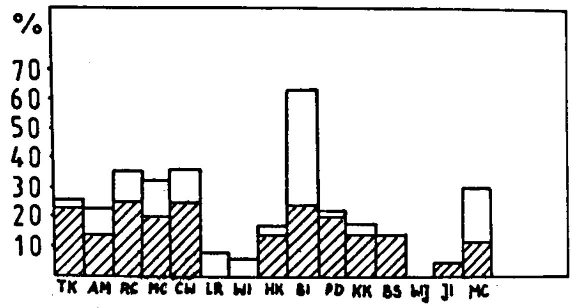

Fig.2. Error scores for F_0 contours (a) in the deterministic algorithm (blank areas) and (b) using quadratic discriminant functions (shaded areas).

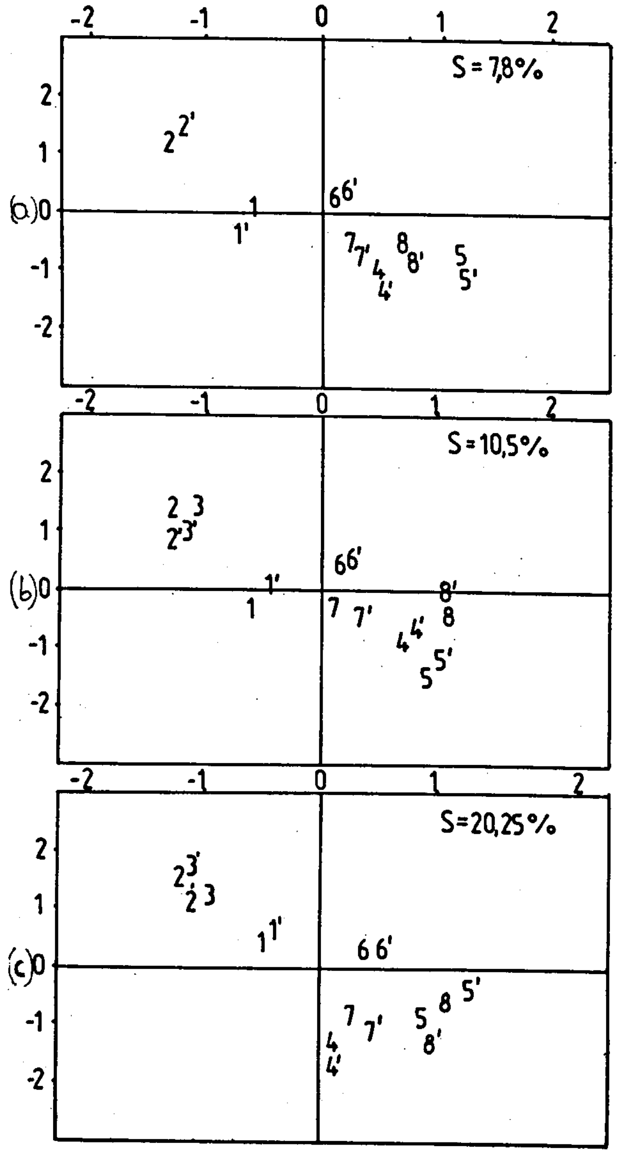

Fig.3. Results of Multidimensional Scaling
(a) voice WI
(b) voice MC (c) voice TK

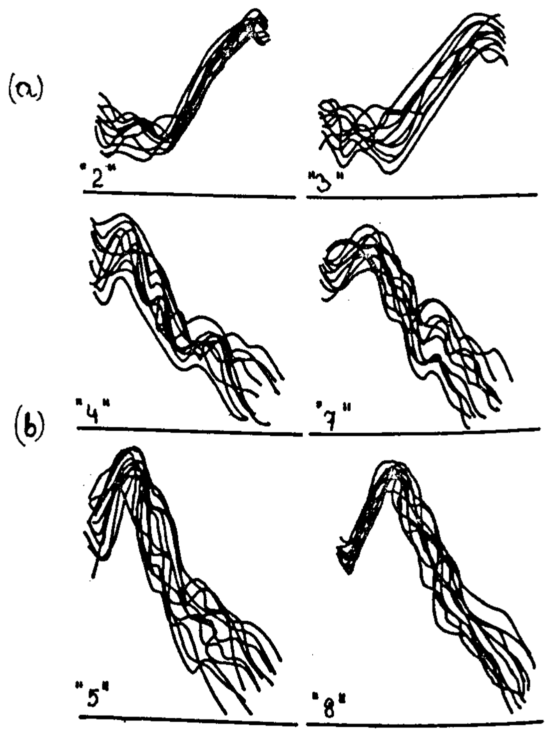

Fig.4. Replications of patterns
(a) "2", "3". Voice MC
(b) "4", "7", "5", "8". Voice TK