MULTIDIMENSIONAL ANALYSIS OF THE SIMILARITY OF PITCH CONTOURS

GRAZyna DEMENKO

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

Abstract

In order to find the relations between the physical and perceptual analysis of fundamental frequency, a number of listeners' 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 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 in 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.

In [3], the various sources of variability of F0 were briefly discussed. If it is desired that most of the manifold variability 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 attempts 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 mimic. 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 characteristi
c features. A method which is optimal with respect to data description uses eigenvectors of the covariance matrices (the Karhunen - Loève 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. (2), (3). The aim of a discriminant analysis is to find a linear combination of the total dispersion of the data collection will be maximized, for the classification of high-class disease.

It is assumed that using the discriminant analysis it would be possible to examine the slopes of prototypes, 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 value was subtracted from the logarithm of the frequency value in successive frequency values within a curve.

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 result. The logarithm of the ratio between these and the Prototype, as well as reducing data, each utterance was divided into 8 parts within which average frequency was calculated as the reciprocal of measured length. It was accepted that the Prototype utterances and their 10 replications by each of 3 of the imitators (3 male voices) will form the classes to be examined. The pitches of these voices differed significantly. The female voices were 65 Hz and that for the male voices was 100 Hz. Each individual class was represented by 80 readings, Fig. 2 depicts a dendrogram tree over the mean vectors of the classes under examination. The values of the F-ratios of c = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 were computed, where for all the classes are statistically significant. The discriminant analysis leads to the following conclusions:

1. The classes under examination may be defined by the 80 per cent accuracy, correct distances between them in a 0-D space. The differences between the individual classes are statistically significant.

2. As the features corresponding to the discriminant variables represent an optimal set to separating the classes. A description of the FO curves in terms of these features appears desirable.

The first discriminant variable is integrated into a straight line passing through the initial and terminal points of the time and frequency normalized curve.

The second characteristic of the set under analysis is the initial frequency value of the curve. Although a slight increase of the frequency value obtained with only two variables, a third variable, the negative value, will slightly improve the classification. It is related to the magnitude of convexity or concavity of the curve. An increase in the magnitude of the value will be obtained when a fourth feature is introduced, under the condition obtaining a classification.

One of the basic methods used in describing each of defined as finding a simple straight, in recognition literature, as the "perception 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 7 decision functions having the property that if $x^n > 0$, then $y^n = 1$, and if $y^n = 0$, then $x^n = 1$. The $x^n$ being the vector to be classified and the $y^n$ to the class of the $x^n$.

Let us consider the classes $W_1$, $W_2$, $W_3$, $W_4$, $W_5$, $W_6$ and $W_7$ as the terminal points of the iterative step during the learning stage the pattern x belonging to class 2 presented to the computer. The decision function $d(x^n) = w(x^n)$ is calculated.

4. PERCEPTUAL ANALYSIS.

The advances in methods of computer optimization and the development of a method of evaluating the results of perceptual experiments, known as Multidimensional Scaling, make it possible to find a configuration of the stimuli in the multidimensional space such that the distances should correspond to subjective dissimilarities between observed objects. A nonmetric relation between the distances and the dissimilarities is required. The concept of stress is introduced to reflect the magnitude of nonmonotonicity, i.e. of the error in the approximation to the experimental data. Except for degenerated configurations, however, there is usually an optimum configuration. The quality of the configuration is generally defined as the goodness if the stress is 5 percent or less, good if 50 percent and acceptable up to 20 percent. An extensive study of the psychological process involved in the perception of tone in speech was presented by Gandour (20). On the basis of the results obtained in Multidimensional Scaling, the author accepted two features as characteristic: the mean frequency and the directrix frequency. The linearity of the stress functions at the second dimension is 30 percent difficult to interpret.

The purpose of the listening experiment to be reported here was to seek the answer to the following questions: Is one of the different voice qualities perceptually similar? Do the listeners consist of a single population? Is there a systematic relation between the perceptual and physical features of the pitch curves? The references were prepared to test the hypotheses in a number of experiments in the conditions of the very good imitator (WJ), one good imitator (GJ) and one (TR) before we show. A panel of 20 listeners (all untrained) were asked to judge similarities between pairs of stimuli. 2 replications were used in which the order of the stimuli was randomized. This experiment was carried out with 10 Voss voice samples. For each voice a number of 1200 curves were used, each of which was randomly selected for each of the listeners. This number is very close to the mean number of 0.1, "with an increase in the rating reflecting the measure of similarity. 28 pairs of stimuli were then assessed. The listeners judged the similarity between the stimuli to be the similarity of 0.20 to 0.50. in a 0-D space it can be that the stimuli are represented by a cluster of the other replications group together, and that for TX there are two clusters, (1, 2, 3, 4, 5), (6, 7, 8, 9, 10), and TX whilst Fig. 4 shows the replicates of patterns 2 and 3 as produced by WJ and GJ whereas Fig. 4 shows the replicates of patterns No. 4, 5, 7 and 8 as produced by voice TX. It is clear from Fig. 4 that the intonations that are confused in perception are also indistinguishable as FO curves. The two perceptual dimensions obtained in the present study may be interpreted as distinguishing the form of the curve (the first dimension, i.e. the strongest discriminant, termed as the second dimension, i.e. the weaker distinctive feature).
Fig. 1. Mean vectors of the 8 classes in a coordinate system of discriminant variables.

Fig. 2. Error scores for F contours (a) in the deterministic algorithm (blank area) 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", "9", "0", Voice TK

Se 52.1.4