COMPUTER-ASSISTED CLASSIFICATION OF BASIC POLISH INTONATIONS

WIKTOR JASSEM

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

ABSTRACT

An experiment was performed to explore, at a basic level, acoustic differences among FØ contours as related to linguistic and perceptual distinctions among intonation patterns. Each of 8 distinct pitch patterns was reproduced, in three sessions 10 times by 10 male and 5 female speakers of Polish. The FØ contours were treated as vectors in an 8-D space. Quadratic and linear discriminant functions were used for an automatic classification of the 1200 vectors with scores of over 80% correct. The misassignments were largely due to missing distinctions in the imitations. It is suggested that not all linguistic distinctions in intonation are categorical. The discriminant functions also permitted a study of similarities and dissimilarites among the different patterns.

THE PURPOSE OF THE EXPERIMENT AND THE DATA

The relations between the properties of an FØ curve as a representation of an acoustical, i.e., physical event and the corresponding linguistically distinct intonation pattern are largely unknown. The present study attempts to come to grips with the basic issue of describing some simple FØ curves so as to be able to assign them automatically to perceptually -- and -- presumably -- linguistically distinct classes.

A phonetician (WJ) recorded 8 versions of the Polish phrase "Dobrze." (/dobʒe/, approx. "OK"), each with a different intonation, viz. Low Rise (LR), Full Rise (FR), High Rise (HR), Low Fall (LF), Full Fall (FF), Level (L), Low Rise-Fall (LRF) and Full Rise-Fall (FRF), with pauses of 5 s. Both with respect to distribution (in discourse) and "meaning" (in a broad sense of the term), the intonations are all different. So it could be assumed that they might be treated as linguistically and perceptually distinct. 10 male and 5 female speakers of Polish listened to these utterances -- the prototypes -- and used the pauses to reproduce (repeat) them with the same "tone of voice". The reproductions (or imitations) were recorded in three different sessions, the first and the last being one month apart, and included 10 replications of each prototype by each speaker. All the 1200 new recordings were analyzed using a period-length meter and a minicomputer. Time normalization was obtained by dividing each utterance into 8 equal fragments and calculating average frequency in each fragment. Thus, the intonation contour of each utterance was represented by a sequence of 8 numbers. The raw data were also frequency normalized (after conversion to a log scale) by putting the mean for each individual voice from all his or her 80 utterances at zero and the variance at one (statistical standarization). This eliminated differences between speakers and allowed one average pattern to be obtained from 150 tokens for each of the 8 FØ contours as shown in Fig. 1.

STATISTICAL TREATMENT

After time and frequency normalization, each of the 1200 utterances was mathematically treated an an 8-element vector, i.e., as a point in an 8-dimensional space. The elements of each vector were the normalized successive frequency values. For each vector, 8 quadratic and 8 linear discriminant functions ere calculated to decide, in two ways, to which of the assumed eight classes: LR,FR,HR,LF,FF,L,LRF or FRF the vector belonged. This was indicated by the highest value of the discriminant function. Also, by observing the decreasing values of the remaining discriminant functions -- DFs --, the relative similarity of each utterance to each of the averaged patterns could be stated. The two kinds of DFs were: (1) The estimator of the quadratic discriminant function, EQDF, and (2) the estimator of the linear discriminant function, ELDF, of the following forms:

EQDF

$$\hat{u}_{ij}(x) = \frac{1}{2}\left[\frac{N_i - p - 2}{N_j - 1} D_j^2(x) - \frac{N_i - p - 2}{N_i - 1} D_i^2(x) + \ln\left(|S_j|/|S_i|\right) - h_4(p_i N_i, N_j) + h_3(p_i N_i, N_j)\right]$$

ELDF

$$\hat{e}_{ij}(x) = \tfrac{1}{2}[h_4(p,N,K)D^2_j(x) - D^2_i(x) + K^{-1}(p,N,K)h_3(p,N_i,N_j)]$$

In the above expressions, x is the observed vector, N is the sample size (here, 150 everywhere), p is the number of dimensions (here, 8), K is the number of classes (here, 8) and S_i , S_j are within-class covariance matrices:

$$D^2_i(x) = (x - \bar{x})' S_i^{-1} (x - \bar{x}) \quad , \quad i,j = 1,\ldots,k \quad j \neq i$$

The forms of the functions h_1, h_3 and h_4 are somewhat involved and are dealt with in [1] . They ensure that the estimators are unbiassed.

Tables 1 and 2 present results of classification obtained by observing the highest value of the DF for every utterance-vector.

		LR	FR	HR	LF	FF	L	LRF	FRF	
	LR	92	1	2	0	0	5	0	0	
a	FR	3	83	15	0	0	0	0	0	a
a	HR	5	17	78	0	0	0	0	0	a
E	LF	0	0	0	84	9	0	5	3	E
J	FF	0	0	0	7	89	0	0	5	J
L	L	5	0	0	0	0	95	7	0	L
L	LRF	0	0	0	12	2	0	82	4	L
d	FRF	0	0	0	1	12	0	4	83	d

classified

Table 1. Results of classification with EQDFs. The figures are percent scores.

		LR	FR	HR	LF	FF	L	LRF	FRF	
	LR	87	1	3	0	0	9	0	0	
a	FR	4	77	19	0	0	0	0	0	a
a	HR	3	25	72	0	0	1	0	0	a
E	LF	0	0	0	81	9	1	7	2	E
J	FF	0	0	0	7	82	0	0	11	J
L	L	6	6	0	0	0	94	0	0	L
L	LRF	7	0	0	12	3	2	80	3	L
d	FRF	0	0	0	1	11	0	4	83	d

classified — presented

Table 2. Results of classification with ELDFs. The figures are percent scores.

It can be seen from Tables 1 and 2 that (1) For all 8 patterns, EQDFs give better classifications than do ELDFs. (2) There is some confusion among the three Rises, between the two Falls and between the two Rise-Falls. (3) The Level and the Low Rise are sometimes mutually confused. (4) There is mutual confusion between the Falls and the Rise-Falls.

The overall results are 85.7% correct classification with the EQDFs and 81.8% correct with the ELDFs.

When the results of the classification of the individual vectors were compared between the two DFs, it was found that in 78.5% of the cases both gave correct and in in 10.9 % both gave the same incorrect assignment. The two methods gave different classification results 9.6% of the time.It is clear therefore that neither one nor the other kind of hypersurfaces separating the eight classes could be perfectly fitted to the entire data.However, a large proportion of the discrepancies between the results obtained by using the two DFs was due to the fact that the final decision was practically a random choice between two of the eight classes. As mentioned above, of the eight EQDFs and eight ELDFs it is those with the highest value that indicate the final assignment. We shall consider two cases here. For one imitation of a Full Rise (voice MC), the following DF values were obtained:

	EQDF	ELDF
LR	- 5.02	-10.50
FR	0.35	- 5.03
HR	1.56	- 4.39
LF	-41.55	-45.74
FF	-59.06	-57.53
L	-26.95	-16.54
LRF	-33.56	-43.35
FRF	-50.82	-54.09

Both functions have the highest values at HR, so both ways the particular expected FR was classified as HR. But in both columns, the difference between the values in the FR and HR rows are distinctly smaller than any other differences. So the ultimate decision between HR and FR is frail. In another case an HR imitation was classified as FR by the quadratic, but as HR (i.e., correctly) by the linear function:

	EQDF	ELDF
LR	- 7.43	-10.87
FR	1.41	- 4.70
HR	1.13	- 4.20
LF	-44.21	-46.85
FF	-59.06	-58.89
L	-26.95	-16.58
LRF	-33.56	-41.94
FRF	-50.82	-51.25

Again, the differences between the two highest values are much less than those between any of the remeining ones. Thus, even a correct decision is not convincing. Indeed, the two utterances were represented by the follwoing vectors (raw data, successive average frequencies in Hz):

(1)[220, 218, 210, 206, 258, 274, 300, 320]
(2)[209, 210, 205, 235, 273, 286, 308, 313]

There is nothing to indicate that the two

sequences of FØ values (or the two corresponding FØ contours) represent two different Rises. Many of the misassignments were of this kind, which is a strong indication that the misclassifications were largely due to an overlap between the 8 classes of utterance-vectors.

INTERSPEAKER DIFFERENCES

When the results of the classification were considered separately for each speaker, the following scores were obtained (percent error for EQDF, with ELDF results in parentheses):

Numbers indicate percent error

WJ Ø
IŁ Ø
LR 1 (1)
JI 4 (7.5)
BS 11 (16)
MC 11 (17)
AM 14 (17.5)
KK 14 (21)
HK 15 (20)
MK 20 (24)
PD 21 (25)
BI 25 (29)
TK 25 (22.5)
CW 25 (30)
BS 26 (38)

The speakers can be seen to have performed quite unequally. The top four speakers were phonetically trained. The remaining ones were all naive speakers. Should the 8 assumed classes of utterance-vectors be completely distinct at the linguistic level, one should have expe ted better individual scores. On the other hand, should they only be perceptually distinct after phonetic training, there would have been less variation in the scores of the 11 untrained subjects. A conclusion that suggests itself from these results is that though the 8 classes can be distinguished at the linguistic level, the differences between some of the classes are not entirely categorical.

SIMILARITIES BETWEEN THE CONTOURS

When the values of the DFs are arranged from the highest to the lowest, the relative similarity of each token to the eight patterns can be judged, the second highest DF indicating the most similar and the last, the most dissimilar pattern. The strength of the similarity and the dissimilarity in our entire materials may be evaluated by considering the number of times that the particular class (pattern) was indicated by the second-highest and the lowest DF. We shall here take into account the quadratic functions only. The results may be summarized as shown in Table 3.This Table contains, in the successive columns, the following:

1. The recognized pattern
2. The most similar pattern
3. The number of cases in which the pattern indicated as most similar actually occurred as the second-highest EQDF
4. The most dissimilar pattern.
5. The number of cases in which the pattern indicated as the most dissimilar actually occurred as the last EQDF. It is to be understood that other patterns occurred in the second and in the last places less frequently than indicated in the Table.

TABLE 3.

1	2	3	4	5
recog.	sim.	freq.	dissim.	freq.
LR	L	77	LRF	94
FR	HR	119	FRF	95
HR	FR	110	FF	77
LF	LRF	88	FR	94
FF	FRF	70	L	59
L	LR	110	FF	60
LRF	LF	61	HR	123
FRF	FF	66	LR	113

The following conclusions can be drawn from the results summarized in Table 3: (1) All the similarities are reciprocal. (2) The dissimilarities are mostly not reciprocal. (3) There is strong similarity between HR and FR, beween L and LR, and there is somewhat weaker similarity between the Falls and the corresponding Rise-Falls (Low with Low and Full With Full). (4) There is strong dissimilarity between the Rise-Falls and the Rises. (5) There is distinct dissimilarity between FF and L.

The similarities and dissimilarities among the 8 patterns may be studied in some more detail by considering also the third highest and the second-last DF. The results of such a study can best be shown by a three-dimensional bar graph like the one in Fig. 2., which by way of an example, refers to the 150 cases of (assumed) HR. The horizontal axis refers to the order of the DF. The highest-value DF, which indicates the assignment to a class, is No.1. The second highest, indicating the strongest similarity, is No.2.The third highest DF, referred to by No.3 shows second-order similarity.. Positions 4,5 and 6 are not very informative. No.8 is the strongest dissimilarity and No. 7 the second-order dissimilarity. The Figure shows that both FR and LR are similar to HR and that FRF and also FF are dissimilar to it.

REFERENCE

[1] G. DEMENKO, W. JASSEM & M. KRZYŚKO: Classification of basic FØ patterns using discriminant functions (forthcoming).

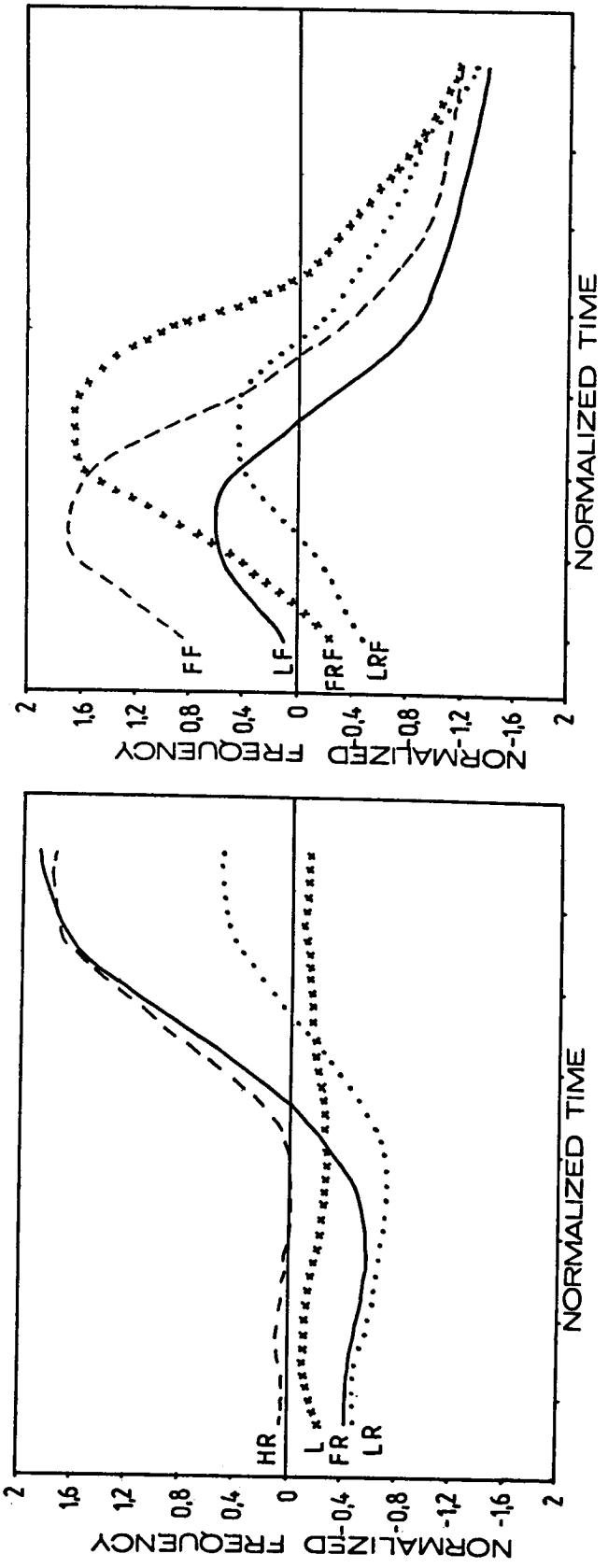

Fig. 1. The eight averaged F_o patterns.

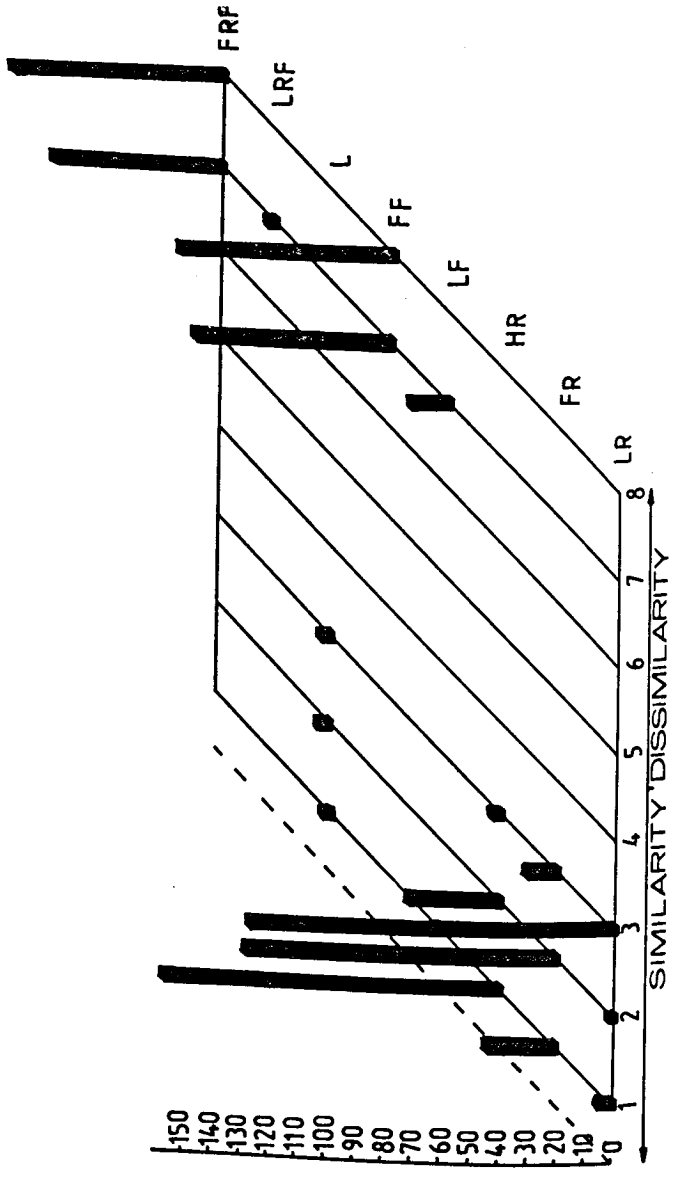

Fig. 2. The 150 tokens of HS with their similarities and dissimilarities.