ARCHIVES OF ACOUSTICS
4, 1, 39-54 (1979)

COMPUTER-AIDED RECOGNITION OF POLISH VOWELS IN CONTINUOUS SPEECH

WIKTOR JASSEM, DANUTA GEMBIAK

Acoustic Phonetics Research Unit of the Institute of Fundamental Technological Research, Polish Academy of Sciences (Poznań)

ANDRZEJ DYCZKOWSKI

Department of Computer Techniques, Mickiewicz University (Poznań)

Vowels pronounced by male voices in typical Polish sentences were the object of the recognition. Vowel formants as variable time functions were measured from the spectrograms. In the first experiment, the patterns for each phoneme in the form of two-element mean vectors and the appropriate covariance matrices were averaged over various combinations of 10 voices. In the second experiment the patterns were averaged separately for each of 10 persons. Quadratic and linear discriminant functions were used for the recognition. In general, the recognition scores in experiment I reached 75% and in experiment II – 90%. It is assumed that in the automatic recognition of Polish vowels in a computer-aided man-machine system and using two features, high scores may be obtained. They are improved by an adaptation of the system to the operator's voice.

1. Introduction

In order to define the method and technique of automatic speech recognition several decisions concerning a number of fundamental questions must be taken, of which the most crucial are the following[1]:

1. The application range of the system.
2. One or more recognition levels of the system.
3. The number of operators of the particular system.
4. The range of the dictionary and grammar of the language being recognized.
5. The range of computerization.

[1] The fundamental problems of the automatic recognition of speech were presented by Newell at al. [21].

6. The number of distinctive features and their types.

7. The mathematical identification-classification model.

The present study continues the preliminary research-work where the following assumptions concerning the above questions were made:

1. The system may be applied wherever it is sufficient to have the dictionary limited to several hundred entries uttered and transmitted in favourable conditions – especially with respect to the signal-to-noise ratio, the transmission characteristics and the reverberation conditions.

2. The recognition of each item is executed in the following steps: (a) the extraction of acoustic-phonetic parameters from the speech signal, (b) the normalization (or adaptation) that takes into account the type of voice, (c) therecognition of phonoids, (d) the recognition of words.

3. The number of operators (or types of voices) necessitating the adaptation is approximately 10.

4. The dictionary entries may be chosen freely but their number must be limited (compare above).

5. The recognizing system is of the hybrid type, i.e. peripheral analysis of the signal is executed in an analog system.

6. The extraction of acoustic-phonetic parameters and the determination of the distinctive features is carried out in the frequency domain (spectral characteristics of the signal) and at the same time the number of the features should be minimum whilst ensuring high correct-recognition scores at the final stage (words).

7. Statistical models of classification are applied in multi-dimensional spaces with the use of discriminant functions.

Although there is no apriori definition of the application of the system, the maximum reduction of costs is presupposed. This will in turn induce mass production of sufficiently versatile systems, equipped with either simple and cheap minicomputers or relatively inexpensive microprocessors.

For the time being it is not possible to state more definitely the cost of the system because of the introductory character of the present research-work and changing prices.

The basic element to be recognized in the system is the phonoid. In view of the great variety of existing and designed recognizing systems, there is no accurate and generally-accepted definition of a phonoid. It is assumed in this study that phonoid is a segmental element (i.e. one that is time-limited and defined by acoustic-phonetic parameters) subject in a particular system to the process of recognition at one of the lower levels of this process and, therefore, is an object in the sense of statistical-mathematical theory of pattern-recognition. The phonoid is a unit of the sequences constituting elements of the higher rank subject to the process of recognition (in the present case – the words). From the linguistic point of view a phonoid is represented by an allophone (see. eg. [9]) at the lower level, and by a phoneme at the higher level

(ibid.). At the present stage of research it is admissible to ignore the distinction between these linguistically different levels. A phonoid may be a set of allophones of one or two phonemes (probably, never three or more). If a particular linguistic difference does not have a heavy distinctive load and necessitates many operations to ensure satisfactory discrimination (in the statistical sense), allophones of different phonemes may be treated as one phonoid[2].

2. The hypothesis

In a parametrically represented speech signal points in time which constitute boundaries between acoustic-phonetic segments can be defined — [7]. These segments are simply related to phonemes, idiophonemes and allophones (ibid.). Some of these segments also stand in particular perceptual-linguistic relations with the phones constituting stationary events. These are the segments representing the so called continuant (or liquid) sounds. Such phones may be pronounced in isolation, i.e. as separate sounds, with any duration between some 50 ms and several seconds. In continuous speech only some segments or parts of the segments may be treated as stationary events. However, there exist fairly simple relations between particular segments representing liquid sounds in continuous speech and their sustained correlates pronounced in isolation.

In the Polish language the syllabic vowels /i, ɨ, e, a, o, u/ belong to the continuant sounds. Formant frequencies are the acoustic-phonetic parameters describing them. It was indicated in [13], [14], [16] that stationary isolated Polish vowels pronounced by different male voices may be characterized by four formant frequencies allowing 100 %-correct classification and identification to be made with the adoption of appropriate probabilistic models. There is only a slight reduction of the scores after limiting the number of distinctive parameters to two, viz. F_1 and F_2. In running speech vowels do not constitute stationary events. Their formant frequencies in this case become variable time functions. These functions depend not only on the phonemic membership of the vowel but also on the proceeding and following phonetic context. Thus, the hypothesis adopted here can be formulated as follows: If the instantaneous values of the formant movements F_1 and F_2 for a particular vowel are measured at small time intervals and for each class (e.g.a phonoid, a phoneme, an allophone, an idiophoneme) a sufficient sampling of the sounds representing a given class is taken, then, on the basis of the set of data from this sample, the class may be represented in such a way as to obtain in a two-dimensional

[2] Should, for instance the difference between the allophones of the Polish phonemes /ɕ/ and /ʃ/ require the application of a considerable number of distinctive features, then this difference could be ignored, because in a lexicon of several hundred words it would never play a discriminating role. Only a few Polish words are differentiated purely by the opposition /ɕ/ : /ʃ/, e.g., proszę: prosię / proʃe(n̪)/ : /proɕe(n̪)/.

space (F_1, F_2), the identification area for a given class. A particular phone being the object of classification or identification at present will be represented as a trajectory in the (F_1, F_2) plane. It should be expected that at least an unambiguously definable part of this trajectory will be found in the identification area of the class which the phone to be recognized represents, so on the basis of this part the whole trajectory can unambiguously be assigned to one of the distinct classes.

3. Materials

The sequences of two, three or four phonemes most frequently occurring in Polish were given in [17]. For the purpose of the present study 6 short sentences satisfying the following conditions were composed: (a) in each sentence the most typical (i.e. the most frequent) triads (three-phoneme sequences) will be as numerous as possible; (b) the sentences will represent colloquial speech and constitute simple, casual utterances; (c) each sentence will be composed of various vowels in order to obtained an approximate balance of the total number of vowels in the materials. Trying heuristically to satisfy the above conditions the following sentences were formulated:

(1) Która godzina? /ˈktura goˈd͡ʒna/ 'what's the time?'

(2) Był pan tu już? /ˈbiwpan ˈtujuʃ/ 'Have you been here before?'

(3) Możesz nic nie mówić? /ˈmɔʒeʃ ˈɲit͡sɲe ˈmuvit͡ɕ/ 'Can you keep quiet?'

(4) Przyda ci się taki? /ˈpʃida ˈt͡ɕit͡ɕe ˈtaci/ 'Can you use one like this?'

(5) Wszyscy ludzie tak robili? /ˈfʃistʂi ˈlud͡ʑe ˈtak roˈbili/ 'Everyone did that?'

(6) Odpowiedział tylko: "Nie wiem" /otpoˈvjed͡ʑaw ˈtɨlko ˈɲevjem/ 'He only answered, 'I don't know'.

The total number of the individual phonemes represented in the above sentences is as follows: /i/7, /ɨ/5, /e/7, /a/7, /o/6 and /u/5.

It is a fact well known from the spectral analysis of speech that some speakers show more distinct vowel formants than others. There is a certain dependence (that has not been so far experimentally examined), among other things, upon the mean pitch of the voice. Therefore, in accordance with the assumptions discussed above, a greater number of voices than were actually used for the measurements were recorded and for all the voices introductory spectrographic analysis was made; the speakers whose spectrograms showed distinct formant movements, were chosen. This procedure may indicate the probable selection of the operators' voices for the future system realizing the method presented here. The number of rejected voices was however very small, the proportion being approximately 1 to 4. F_1 and F_2 values were read from the spectrograms, with an accuracy of 50 Hz, at intervals of $\Delta t = 20$ ms. The number of the bivariate measurements for each vocalic sound was always more than two, but never exceeded 10. Fig. 1 shows an illustrative wide-band spectrogram and the corresponding F_1 and F_2 formant movements, which in

the entire material were calculated either from the wide or narrow-band spectrograms, or if necessary, from both. In Fig. 2 the measurements of the successive (F_1, F_2) values within the vowels occuring in the sentence from Fig. 1 are marked as points in the (F_1, F_2) plane. These points are connected and the resultant trajectory represents the individual phones. The borders between

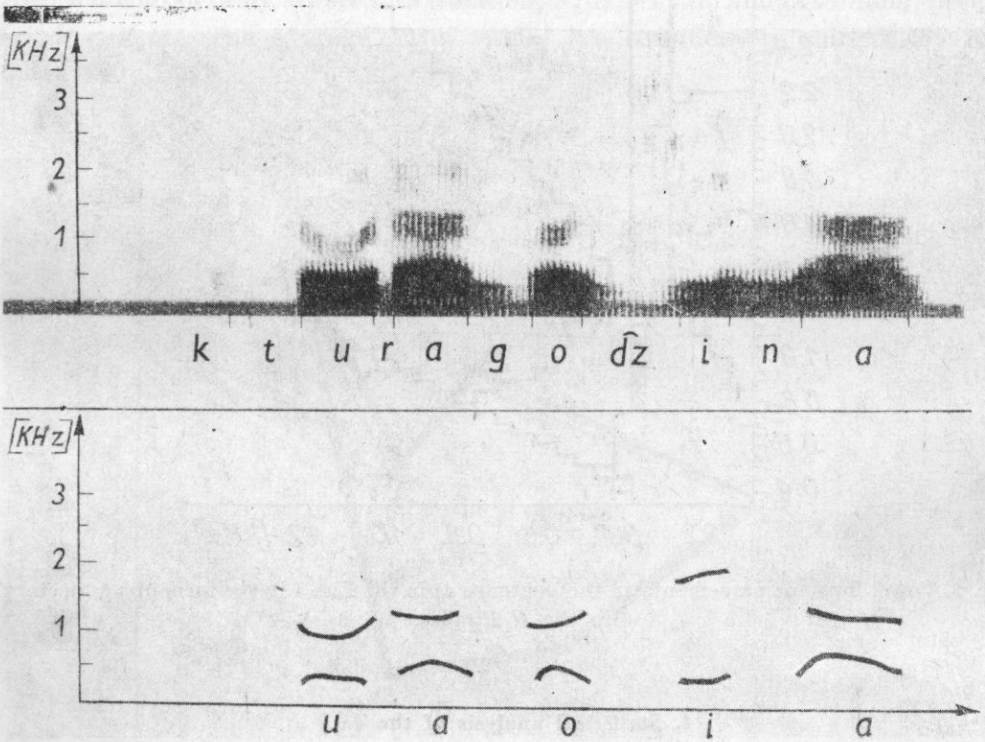

Fig. 1. Formants as time functions. A wide-band spectrogram and the formant movement in the phrase "ktura godzina" voice MN

the identification areas are marked in the (F_1, F_2) plane. These areas will be discussed in detail in the following sections.

In the present study the preliminary processing of the signal is of a semi-automatic nature because the data were obtained from visual measurements. A fully automatic system of speech recognition obviously presupposes the direct data input to the digital computer, the data being obtained either from the direct quantization of the signal or from an A/D converter after introductory processing of the signal in an analog system. At the present moment in the Acoustic Phonetics Research Unit definite progress has been made on the way to a complete automatization of the process of speech recognition (see, e.g., [19] and [20]).

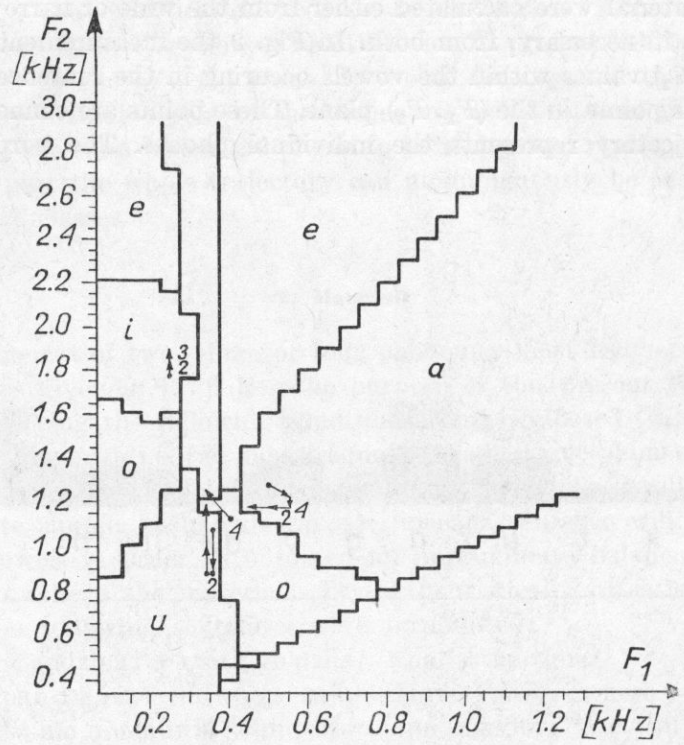

Fig. 2. Vowel-formant movements in the sentence as in the Fig. 1 in the form of a trajectory in the F_1F_2 plane

4. Statistical analysis of the data

11 speakers took part in experiment I. The data obtained from the measurements were processed in 11 variants. Each time the measurement obtained from 10 speakers constituted the design (learning) set for the individual phonemes, while the data obtained from the 11th voice constituted the test set. In each of the eleven variants the data from a different speaker were excluded from the design set. Each sample (design set) was described as a bivariate statistical distribution characterized by a mean vector and a covariance matrix (see, e.g. [5] and [24]). As each variant of experiment I differed from every other variant in the participation of one person (for 10 participants) the frequency values for a given formant of a given vowel are either identical or differ by some 1 to 2 %. The comparison of the formant values in continuous speech[3] with the data obtained for stationary vowels [13] shows the following differences:

[3] These values were presented in detail in [18].

(1) Average F_1 values are less differentiated for the vowels in continuous speech than for the isolated vowels and F_1 in continuous speech is higher for the closed vowels /i, i̇, u/, and lower for the open vowels /e, a, o/.

(2) The average F_2 values are also closer. In running speech F_2 is lower for the vowels /i, i̇/ and higher for the remaining vowels. The average value for /i̇/ is considerably lowered. In isolation, $F_2/\dot{i}/ > F_2$ /e/ and in running speech — $F_2/\dot{i}/ < F_2/e/$. The average $F_2/u/$ values are considerably shifted (by more than 400 Hz).

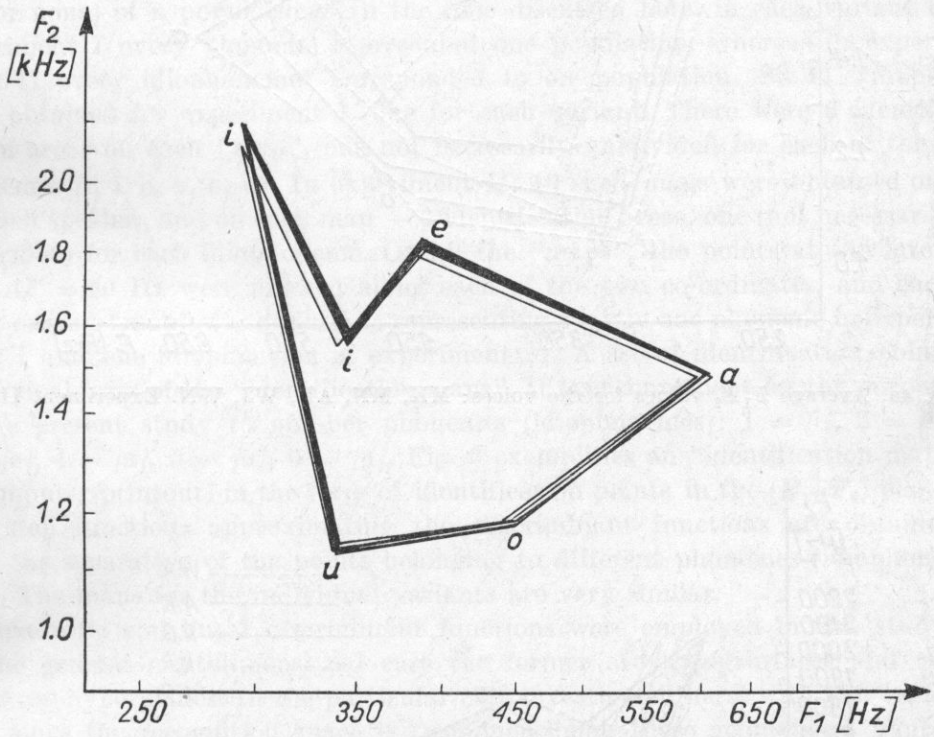

Fig. 3. Average F_1F_2 values. Experiment I

The covariance matrix is square and symmetric, so in the present case, where $s_{12} = s_{21}$, the three values s_{11}, s_{12} and s_{22} (cf. [18]) are necessary to construct it. For experiment II in which for each of the 10 voices, 5 repetitions of each sentence were analysed, the mean vectors and the covariance matrices were calculated from 4 replications constituting the design set; and the data from the fifth replication constituted the test set. Figs. 3 and 4a, b represent graphically the design data. The average F_1 and F_2 values are much more differentiated for the particular voices than for the variants of experiment I. On the other hand, the variations of individual idiophonemes in experiment II

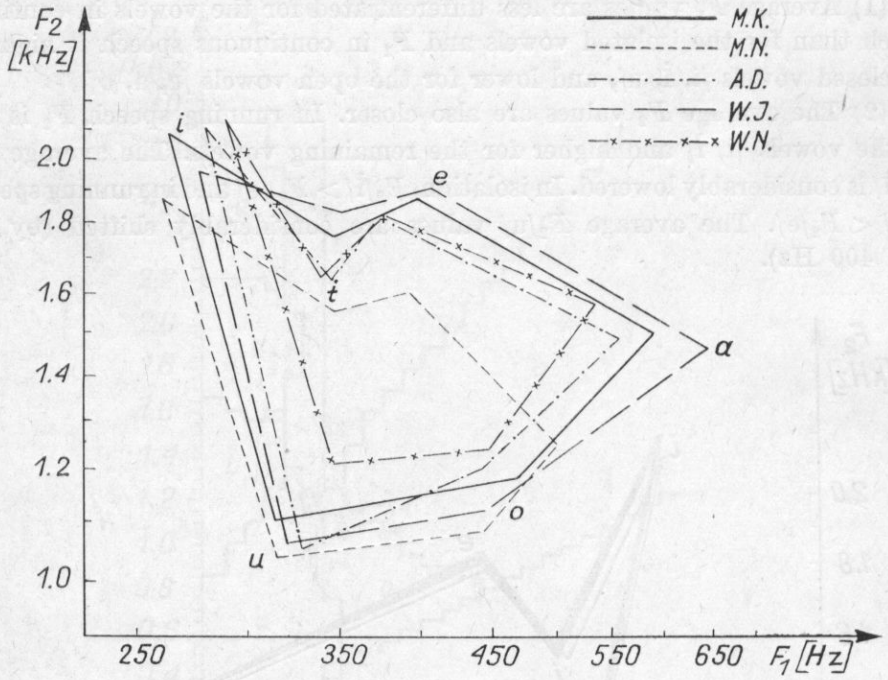

Fig. 4a. Average F_1F_2 values for the voices: MK, MN, AD, WJ, WN. Experiment II

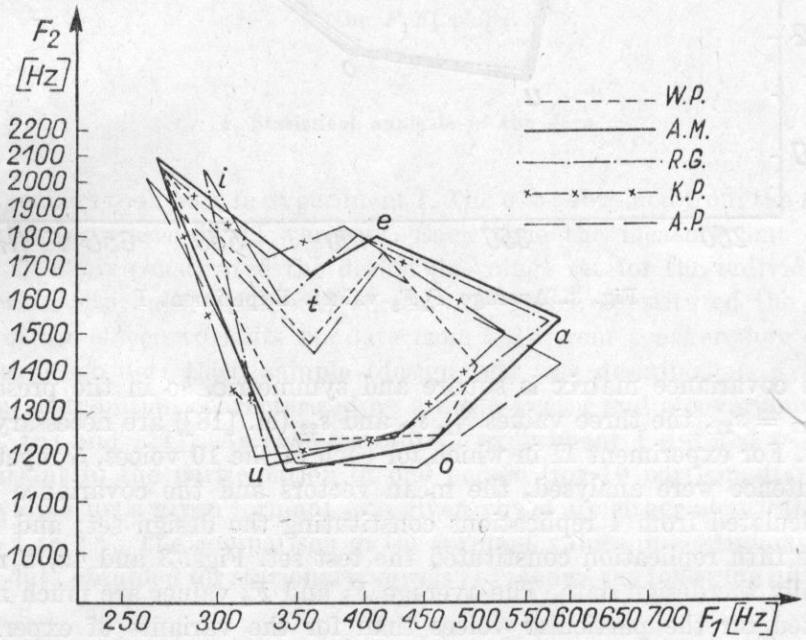

Fig. 4b. Average F_1F_2 values for the voices: WP, AM, RG, KP, AP. Experiment II

are distinctly smaller than those of the phonemes in experiment I. This suggests that the identification of vowels will give much better results in experiment II, than in experiment I.

5. Identification points

As indicated in [13], [14] and [15], in the plane with the coordinates F_1, F_2 in an orthogonal coordinate system, any point may be defined as being situated in exactly one identification area representing one (in the case of linear functions one or none) of n populations. In the case discussed here, in each variant of experiment I every phonome represented one population, whereas in experiment II every idiophoneme[4] coresponded to on population. So 11 "maps" were obtained for experiment I, one for each variant. There were 6 identification areas on each "map", one not necessarily unidivided for each of the 6 phonemes /i, i̇, e, a, o, u/. In experiment II, 10 such maps were obtained one for each speaker, and on each map — 6 identification areas, one (not necessarily undivided) for each idiophoneme. On all the "maps", the points at the intervals $\Delta F = 50$ Hz were marked along each of the two co-ordinates, and each point on any "map" was defined as representing exactly one phoneme in experiment I and one idiophoneme in experiment II. A set of identification points is the final form of the "identification maps". It was convenient for the purpose of the present study to number phonemes (idiophonemes): 1 = /i/, 2 = /i̇/, 3 = /e/, 4 = /a/, 5 = /o/, 6 = /u/. Fig. 5 exemplifies an "identification map" (a computer printout) in the form of identification points in the (F_1, F_2) plane. The step functions approximating the discriminant functions are obtained after the separation of the points belonging to different phonemes (idiophonemes). The maps for the individual variants are very similar.

Quadratic and linear discriminant functions were employed in this study. In the general multidimensional case the former are hypersurfaces and the latter are hyperplanes. In the particular case investigated here, with two variables, since the recognition space is twodimensional, these geometrical figures are reduced to second-order curves and straight lines respectively. The quadratic discriminant functions are expressed by the formula

$$v_{ij}(\boldsymbol{x}) = \boldsymbol{x}'(\Sigma_j^{-1} - \Sigma_i^{-1})\boldsymbol{x} + 2(\boldsymbol{\mu}_i'\Sigma_i^{-1} - \boldsymbol{\mu}_j'\Sigma_j^{-1})\boldsymbol{x} + \\ + \boldsymbol{\mu}_j'\Sigma_j^{-1}\boldsymbol{\mu}_j - \boldsymbol{\mu}_i'\Sigma_j^{-1}\boldsymbol{\mu}_i + \ln\frac{|\Sigma_j|}{|\Sigma_i|} + 2\ln\frac{q_1}{q_j}, \qquad (1)$$

whereas the linear functions are expressed by $\boldsymbol{b}'\boldsymbol{x} + c = 0$ in which

$$\boldsymbol{b}'\boldsymbol{x} + c \leqslant 0 \Rightarrow \boldsymbol{x}_0 \in \pi_1, \quad \boldsymbol{b}'\boldsymbol{x} + c > 0 \Rightarrow \boldsymbol{x}_0 \in \pi_2, \qquad (2)$$

[4] An idiophoneme is a phonologically distinctive class of sounds consisting of sound elements pronounced by a particular voice (particular speaker) (see [9]).

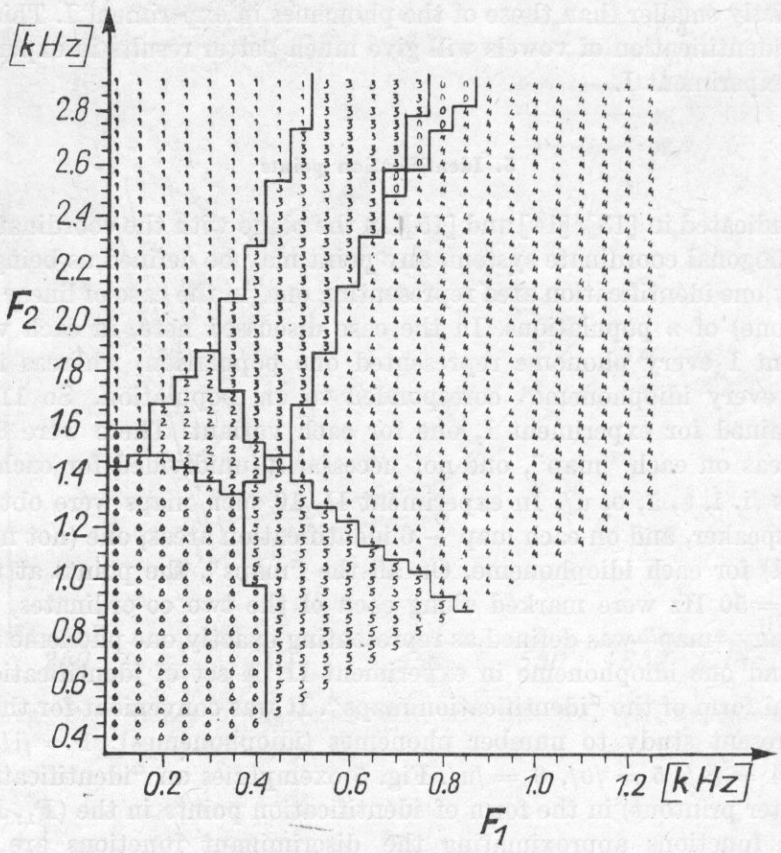

Fig. 5. The identification map for the first variant of experiment I (linear functions)

where $\boldsymbol{x}_0$ is the point being examined on the plane defined by the F_1 and F_2 variables, and π_1 and π_2 are two populations divided by a given discriminant function, such that

$$P(\pi_2|\pi_1) = P(\boldsymbol{b}'\boldsymbol{x}+c>0) = P\left[\frac{\boldsymbol{b}'\boldsymbol{x}-\boldsymbol{b}'\boldsymbol{\mu}_1}{(\boldsymbol{b}'\boldsymbol{\Sigma}_1\boldsymbol{b})^{1/2}} > \frac{-c-\boldsymbol{b}'\boldsymbol{\mu}_1}{(\boldsymbol{b}'\boldsymbol{\Sigma}_1\boldsymbol{b})^{1/2}}\right] \tag{3}$$

$$= 1-\Phi-\frac{c+\boldsymbol{b}\boldsymbol{\mu}_1}{(\boldsymbol{b}'\boldsymbol{\Sigma}_1\boldsymbol{b})^{1/2}} = 1-\Phi(t_1),$$

$$P(\pi_1|\pi_2) = P(\boldsymbol{b}'\boldsymbol{x}+c\leqslant 0) = P\left[\frac{\boldsymbol{b}'\boldsymbol{x}-b'\boldsymbol{\mu}_2}{(\boldsymbol{b}'\boldsymbol{\Sigma}_2\boldsymbol{b})^{1/2}} \geqslant \frac{-c-\boldsymbol{b}'\boldsymbol{\mu}_2}{(\boldsymbol{b}^0\boldsymbol{\Sigma}_2\boldsymbol{b})^{1/2}}\right]$$

$$= 1-\Phi\left[-\frac{c+\boldsymbol{b}'\boldsymbol{\mu}_2}{(\boldsymbol{b}'\boldsymbol{\Sigma}_2\boldsymbol{b})^{1/2}}\right] = 1-\Phi(t_2),$$

where μ_1 and Σ_1 are the parameters (the mean vector and the covariance matrix) characterizing the population π_1, and μ_2 and Σ_2 are the parameters of the population π_2.

More details about the mathematical methods employed are presented in [14] and [18].

6. Two-stage segment→(idio)phonoid recognition

The measurements of the F_1 and F_2 values at each point in time may be represented as a point on the appriopriate identification map, and at the same time each bivariate measurement (each bivariate observation) can be classifield (identified) as representing exactly one phoneme in experiment I or exactly one idiophoneme in experiment II. Then a sequence of bivariate measurements will correspond to a sequence of classificatory decisions. Each F_1, F_2 bivariate measurement is treated as defining one 20 ms segment, and each individual vowel sound subject to the classification (identification) and represented by a sequence of bivariate measurements may then be represented as a sequence of decisions determining the successive segments. Because the formant movements within the particular vowel sounds are slow-varying continuous time functions, it may be expected that at least some of the subsequent segmental decisions will be identical (i.e. will assign segments to the same (idio)phonemes).

The method of recognizing the vowel sounds as the sequences of segments is as follows:

Each sequence of identical decisions within the vowel sound is defined as a partial sequence eg. *EEE*, *AA*, *OOO*, etc. Some of the vowel sounds were recognized as single partial sequences (e.g. one vowel sound was represented by a sequence of segmental decisions *IIIII*). These cases will be defined as unitary recognitions. Other vowel sounds were represented by two or more partial sequences. For instance, one vowel sound consisted of the sequence of decisions *EYYY*, in which two partial sequences occur, the first consisting of one element, and the second of three. Another vowel sound was recognized at the level of segmental decisions as *OEEEY*, i.e. as three partial sequences.

If an entire sequence of segmental decisions corresponding to a given vowel sound was composed of two or more partial sequences, then either one partial sequence was more numerous than the others. e.g. *EEAOOO*, or two (or possibly more) partial sequences were equally numerous, eg. *YYEEY*, *EAO* (three patrial sequences). In the first case — recognition according to the majority — the final decision was taken in favour of the longest sequence, in the second — recognition according to the order — the final decision was taken in favour of the last of the equally numerous partial sequences. So for the examples quoted above the decisions defining the asignment of a vowel sound to a particular phoneme or idiophoneme (in this study synonymous

with a patricular phonoid or idiophonoid) are as follows: $IIIII \rightarrow I$, $EEAOOO \rightarrow O$, $YYEEY \rightarrow E$, $EAO \rightarrow O$.

The recognition of vowels in running speech according to the algorithm adopted here may be presented in the form of a flow diagram, as presented in Fig. 6. The computer Odra 1204 was used for the calculation of the statistical

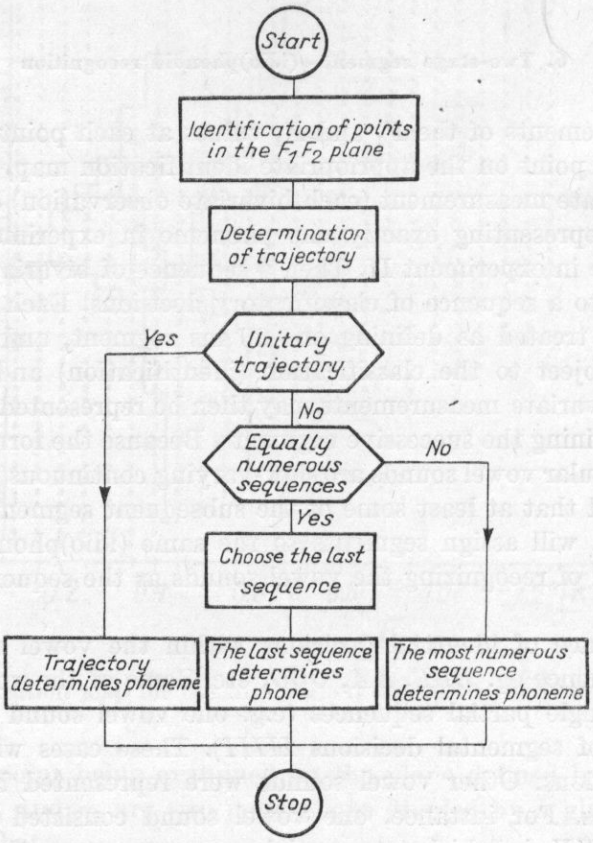

Fig. 6. The recognition of phonemes on the basis of 20 ms segments: a flow chart

parameters, the determination of the decision function, the construction of the identification maps and the execution of the decision algorithm. The programs for the calculations were presented in [13].

On the basis of the results the segment→(idio) phonoid recognition may be described as follows:

(1) Among the correct recognitions the unitary recognitions were considerably much more frequent than the recognitions according to the majority, whereas the recognitions according to the order were the least common.

(2) Among the wrong recognitions, the decisions taken according to the majority were the most numerous, the unitary decisions were slightly less

numerous, and the least frequent ones were the recognitions according to the order.

(3) The type of discriminant function seems to have no significant effect on the type of the classificatory decisions.

Among the recognitions according to the order there were more uncorrect decisions than correct ones. However, no conclusion as to the fallacy of this part of the method should be drawn from this fact for the following reasons:

(a) different types of sequences of various size were present among these recognitions and the decisions were based on various numbers of patrial sequences.

(b) according to the available material from both experiments, a slight increase in the number of the correct decisions in this type of sequences — comprising the equally numerous patrial sequences — may be acheived at the cost of considerably complicating the general form of the algorithm.

(c) the sequences of this type are very rare, and for the determination of the optimum algorithm which would take into consideration these cases, an examination of very extensive material would be necessary.

Adopting, for the present, the procedure described above, the general algorithm of the two-stage recognition on the bases of segment→(idio) phonoid, may be formulated as follows:

Choose the last of the equally numerous partial sequences all longer than the remaining sequences.

This algorithm comprises the three methods of recognizing the entire sequences representing vowel sounds as explained above. If an entire sequence is unitary, then the number of the equally numerous patrial sequences equals one. It is the only and at the same time the last patrial sequence and it is more numerous than the remaining ones, the size of which is 0. In the case of two or more patrial sequences one of which is the longest, the number of the remaining ones is 1 or more.

7. The results of the recognition

7.1. Experiment I. Table 1 presents the results of the recognition of the vowel sounds in each variant of experiment I using quadratic and linear functions. The figure representing the percentage of correctly recognized sounds of each variant are contained between 70 and 80% for the quadratic functions and 60-80% for the linear functions. The results of the classification in the design set (i.e. the recognition in training) differ only slightly form the results of one identification in the test set. This indicates that the choice of the discriminant functions was correct. The absence of a distinct difference in the overall results as between the methods based on the quadratic and the linear functions results from the fact that the segments of the conical curves separating the identification areas show very small curvatures.

Averaging over all the vowels and variants the percent recognition scores for each person are as shown in Table 2. Some voices (e.g. III and IX) give better results than the remaining ones. This reflects the fact that the formant movements of the vowels pronounced by these voices are more typical in the sense of the method adopted here.

The classification in the design set and the identification in the test set were much better in experiment I for the extreme vowels /i, a, u/ than for the remaining vowels[5].

7.2. Experiment II. As could be expected from the data Figs. 3 and 4a, b the results of the recognition of vowels in experiment II turned out to be much better than those of experiment I.

The data in Table 3 indicate that the differences between the voices are inconsiderable. The classification in the design set, as well as the identification in the test set resulted in scores of about 90 %. The best results were obtained for the idiophones /i, a, o/. The overall scores are considerably lowered by the results obtained for /e/, which are markedly lower than those for the remaining phonemes. It may be assumed that an allophonic division of the (idio) phoneme /e/ would yield better results. Two dependent phonoids would represent the phoneme /e/ in automatic recognition.

8. Conclusions

8.1. Experiment I simulates the situation in which a technical system recognizes the vowels within an utterance pronounced by a certain number (assumed to be sufficient) of other voices. Experiment II, on the other hand, simulates the situation where the system has either previously recognized the operator's voice (or at least the type of the operator's voice) or has been tuned to his voice.

If for certain purposes the recognition accuracy of vowels of about 75 % would be acceptable, then there is no need for tuning the system to the operator's voice or recognizing it by the system.

However, if accuracy of about 90 percent is required, then the tuning or recognition of the operator's voice is necessary.

8.2. If the spoken vowels are assigned to the proper idiophonemes, then some of them viz. those representing the phoneme /i/ will get almost one-hundred-percent correct recognition scores, whilst the others will get about 90 %.

Only /e/ and possibly /i/ result in less satisfactory recognition, /e/ and possibly /i/ could be recognized according to more idiophonemes than one

[5] See detailed data in [18].

Table 1. Correctness of recognition for individual variants of experiment II [%]

Variant	Quadratic functions		Linear functions	
	pattern set	test set	pattern set	test set
1	78	70	77	81
2	80	70	79	68
3	77	78	75	84
4	77	70	76	73
5	79	68	77	70
6	77	73	78	60
7	77	78	76	81
8	78	78	76	78
9	78	81	75	87
10	77	81	76	78
11	76	81	76	76
Average	78	75	77	76

Table 2. Correctness of classification for individual voices in experiment I [%]

Voice	Quadratic functions	Linear functions
I	79	81
II	69	67
III	80	84
IV	73	74
V	73	71
VI	74	62
VII	79	81
VIII	79	79
IX	81	86
X	82	79
XI	84	77

Table 3. Correctness of recognition for individual voices in experiment I [%]

Voice	Quadratic functions		Linera function	
	pattern set	test set	pattern set	test set
AD	95	95	93	92
WN	92	95	92	95
WJ	93	97	92	97
MŁ	92	97	92	95
MN	93	89	92	89
WP	80	81	83	81
AM	90	89	88	86
RG	92	92	93	92
KP	88	84	86	78
AP	81	92	78	86
Average	90	91	89	89

(probably not more than 2) should this turn out to be necessary. This would presumably result in overall recognition scores approaching one-hundred-percent correctness, at least for some selected voices.

8.3. It has been shown that the recognition of Polish vowels in continuous speech is possible with the use of only two classificatory features, namely the instantaneous values of F_1 and F_2 measured with an accuracy of 50 Hz, at 20 ms intervals. This method permits the identification to be executed in a hybrid analog-digital system processing a small number of data. Assuming the range for F_1 from 200 Hz to 1100 Hz, and 400-2700 Hz for F_2, and taking into consideration that by definition $F_2 > F_1$ and a part of the (F_1, F_2) plane is not used because of articulatory constraints (in real speech high F_2 values do not occur together with relatively high F_2 values), the number of the identification points may be limited to no more than about 500. The possibility of realizing the method adopted here in real time using hybrid systems is at present the object of further research.

References

[1] T. Caliński, A. Dyczkowski, Z. Kaczmarek, *Identification of observation using a dividing hyperplane* [in Polish], Rocznik Akademii Rolniczej w Poznaniu, *Algorytmy biometryczne i statystyczne*, No 4 (1975).

[2] G. I. Gjemel, *Recognition of speech signals* [in Russian], izd. Nauka, Moskwa 1971.

[3] M. Dierkacz, R. Gumieckij, Ł. Miszin, M. Owierczenko, M. Czaban, *The perception of speech in recognition models* [in Russian], izd. Iwowskowo Uniwersiteta, Lwów 1971.

[4] H. Dudley, S. Balashek, *Automatic recognition of phonetic patterns in speech*, JASA, **30**, 721-732 (1958).

[5] K. Hope, *Methods of multivariate analysis*, Univ. of London Press, London 1968.

[6] W. Jassem, *Vowel formant frequencies as cues to speaker discrimination*, *Speech Analysis and Synthesis*, Vol. I, PWN, Warszawa, 1968, 9-42.

[7] W. Jassem, *Phonetic-acoustic assumptions for the automatic recognition of phonemes* [in Polish], IFTR Reports 17/70, Warszawa (1970).

[8] W. Jassem, *Phonological segmental units in the speech signal, form and substance*, Akademik Verlag, Odense, 181-192 (1971).

[9] W. Jassem, *Speech and communication*, [in Polish], PWN, Warszawa 1974.

[10] W. Jassem, T. Caliński, Z. Kaczmarek, *Vowel formant frequencies as personal voice characteristics* [in Polish], IFTR Reports 5/70, Warszawa (1970).

[11] W. Jassem, T. Caliński, Z. Kaczmarek, *Investigation of vowel formant frequencies as personal voice characteristics by means of multivariate analysis of variance*, *Speech Analysis and Synthesis*, V. II, PWN, Warszawa 1970, 7-40.

[12] W. Jassem, L. Frąckowiak, *Vowel formant frequencies as a distinctive feature of speakers' voices* [in Polish], Biuletyn Polskiego Towarzystwa Językoznawczego, V. XXVI, Kraków, 67-99 (1968).

[13] W. Jassem, M. Krzyśko, A. Dyczkowski, *Classification and identification of Polish vowels on the basis of formant frequencies* [in Polish], IFTR Reports 14/72, Warszawa (1972).

[14] W. JASSEM, M. KRZYŚKO, A. DYCZKOWSKI, *Identification of isolated Polish vowels* [in Polish], Archiwum Akustyki, **9**, 3, Warszawa, 261-287 (1974).

[15] W. JASSEM, M. KRZYŚKO, A. DYCZKOWSKI, *Sequential identification of vowels* [in Polish], IFTR Reports 6/74, Warszawa (1974).

[16] W. JASSEM, M. KRZYŚKO, A. DYCZKOWSKI, *Verification of voices on the basis of vowel formant frequencies* [in Polish], Archiwum Akustyki, **9**, 1, Warszawa, 3-26 (1974).

[17] W. JASSEM, P. ŁOBACZ, *Phonotactic analysis of Polish text*, [in Polish], IFTR Reports 63/71, Warszawa (1971).

[18] W. JASSEM, D. SZYBISTA, A. DYCZKOWSKI, *Recognition of Polish vowels in typical sentences* [in Polish], IFTR Reports 43/75, Warszawa (1975).

[19] H. KUBZDELA, *Technical realization of the formant method for the recognition of Polish vowels* [in Polish], IFTR Reports 90/75, Warszawa (1975).

[20] K. MYTKOWSKI, *Analog function channel of the type KF-01 for the input and output of data in an "On — line" system to/from the storage of the minicomputer MOMIK-8B/100*, [in Polish], IFTR Reports 39/76, Warszawa (1976).

[21] A. NEWELL, J. BARNETT, J. W. FORGIE, C. GREEN, D. KLATT, J. C. R. LICKLIDER, J. MUNSEN, D. R. REDDY, W. A. WOODS, *Speech understanding systems, Final report of a study group*, North Holland, American Elsevier, 1973.

[22] G. E. PETERSON, M. L. BARNEY, *Control methods used in a study of the vowels*, JASA, **24**, 175-184 (1952).

[25] R. W. A. SCARR, *Zero-crossing as a means of obtaining spectral information in speech analysis*, IEEE Trans. on Audio Electroacoustics, AU-16, 247-255 (1968).

[24] M. M. TATSUOKA, *Multivariate analysis: techniques for educational and psychological research*, Wiley and Sons, New York 1971.

[25] N. G. ZAGORUJKO, *The methods of recognition and their application* [in Russian], izd. Sow. radio, Moskwa, 1972.

Received on 14th September 1977