ARCHIVES OF ACOUSTICS
19, 2, 147–159(1994)

AUTOMATIC DISCRIMINATION OF POLISH STOP CONSONANTS BASED ON BURSTS ANALYSIS

P. DOMAGAŁA AND L. RICHTER

Department of Acoustic Phonetic
Institute of Fundamental Technological Research
Polish Academy of Sciences
(61-704 Poznań, ul. Noskowskiego 10)

The aim of the work reported is to test the possibility of speaker- and context-independent automatic discrimination of Polish stop consonants. A new approach to stop consonant discrimination has been proposed based on a linear combination of autocorrelation function values. By computing BETWEEN and WITHIN matrices for parameters representing different populations and by solving the general eigenproblem, the direction of the eigenvector is determined, corresponding to the maximum eigenvalue. When projected in this direction, objects belonging to one population are most clustered and pair-wise projection of microsegments ("each with each") representing stop consonants was performed. Multidimensional parameter space was reduced to one dimension (axis). The material consisted of nonsense words, with most common Polish stop consonant contexts, produced by 20 speakers (10 male and 10 female). Experiments were conducted for male and populations themselves maximally separated. It is in this direction that female voices separately as well as for all the voices pooled. The burst segment has been found to provide better cues for phone identification than the frication segment. The average identification rate (for voices pooled) was. 764.

1. Introduction

In automatic speech recognition, identification of stop consonants encounters particular difficulties from the very nature of the corresponding acoustic events. It seems that selection of proper parameters, optimum for discrimination purposes, may be crucial here. In the most recent studies of stop consonants, various parametrization methods have been applied [1], [2], [3], [4]. In the present paper, an entirely new approach has been put forward, based on a certain linear combination of autocorrelation function values.

2. Speech signal parametrization

Nonsense words containing stop consonants were tape-recorded and input into an IBM PC AT memory. A 13-bit A/D converter and a sampling rate of 10 000 kHz were

applied. The signal was limited to 5 kHz. The preemphasis filter followed the dependence $y(n)=x(n)-x(n-1)$. The first 13 autocorrelation function values were used for signal parametrization. The function R is defined by the formula:

$$R(\tau) = \lim_{T\to\infty} \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t) * x(t+\tau)\, dt. \tag{1}$$

For $\tau=0$ the autocorrelation function is the mean square of the function, which corresponds to the mean energy of the signal. For time-limited segments of the discrete signal (frames), formula (1) has the form[1]:

$$R(n) = \sum_{i=0}^{N-1} x(i) * x(i+n). \tag{2}$$

Since for a given N (corresponding to frame duration) $R(0)$ is the maximum value, it was used for normalization of the autocorrelation function. Autocorrelation function expresses, in a sufficient degree, the most important features of the speech signal. It makes a departure point for an LPC analysis enabling one to obtain estimates of the speech signal spectrum envelope, cross-section functions of the vocal tract or reflection coefficients, thus making resynthesis possible. For the purposes of the present work, a 128-sample frame (12.8 ms) has been adopted. It moved along the time axis in 64-sample steps. In other words, parametrization converts the sequence of numbers corresponding to the speech signal stored into a sequence of 12-element vectors.

3. Theoretical bases of automatic discrimination

Theoretical bases were presented in detail in [5] and [6], hence only the main idea of the present method is sketched below.

It is convenient to represent a p-element vector as a point in a p-dimensional space. A large set of such points belonging to a given class (pattern) in this space makes a cluster of a certain shape and concentration. Assuming the normal distribution of such points, the confidence region at a given percent level forms a p-dimensional ellipsoid. Even though the assumption of the normality of distribution is usually not fulfilled, it often leads to optimum results. For a multi-dimensional ellipsoid, the direction of its longest axis can be determined. It coincides with the direction of covariance matrix Σ eigenvector corresponding to the maximum eigenvalue. In Fig. 1, an example for two dimensions is shown. The axis determines the direction along which the scatter of elements is the greatest. Eigenvectors and eigenvalues are a result of solution of the eigenproblem (3):

[1] It is assumed that $x_i=0$ for $i<0$ and $i>N-1$

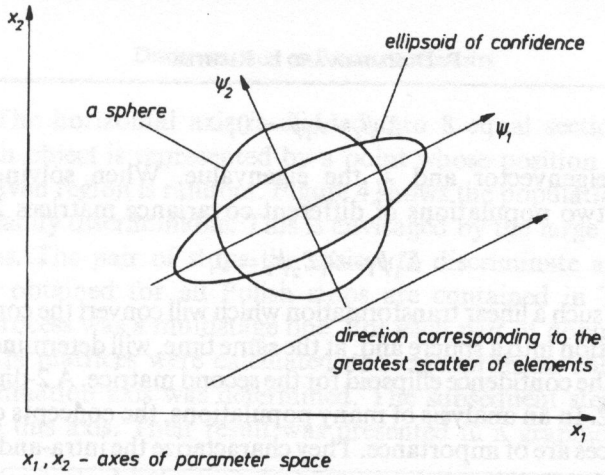

Fig. 1. Ellipse of confidence in parameter space

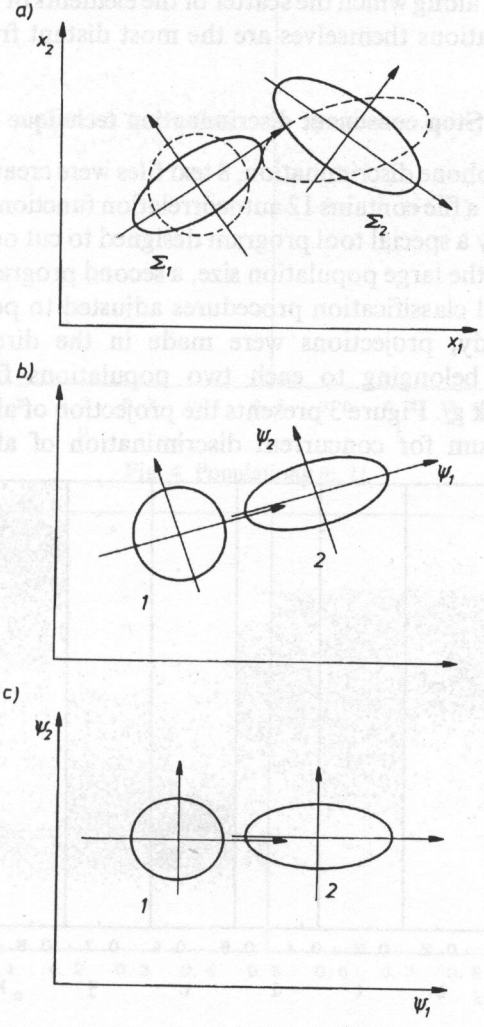

Fig. 2. Diagonalization for 2 populations.

$$\Sigma\psi_i - \lambda_i\psi = 0, \tag{3}$$

where ψ_i is the eigenvector and λ_i the eigenvalue. When solving a generalized eigenproblem for two populations of different covariance matrices Σ_1 Σ_2

$$\Sigma_1\psi_i - \lambda_i\Sigma_2\psi_i = 0, \tag{4}$$

it is possible to find such a linear transformation which will convert the confidence ellipsoid for the first population into a sphere and, at the same time, will determine the direction of the longest axis of the confidence ellipsoid for the second matrice. A 2-dimensional case is illustrated in Fig. 2. In an analysis of many populations, the concepts of WITHIN and BETWEEN matrices are of importance. They characterize the intra-and interpopulation scatter, respectively. Solving the appropriate generalized eigenproblem for them determines the direction along which the scatter of the elements of one population is the smallest and the populations themselves are the most distant from each other.

4. Stop consonant discrimination technique

For the purposes of phone discrimination, 8 text files were created (one for each stop consonant). Each line of a file contains 12 autocorrelation function values ($R(1)..R(12)$). The files were created by a special tool program designed to cut out selected fragments of the signal. In view of the large population size, a second program was written which automatically generated classification procedures adjusted to populations analyzed.

In the present study, projections were made in the direction optimum for separation of objects belonging to each two populations from among 8 stop consonants /p b t d c ɟ k g/. Figure 3 presents the projection of all the bursts collected in the direction optimum for concurrent discrimination of all the 8 populations

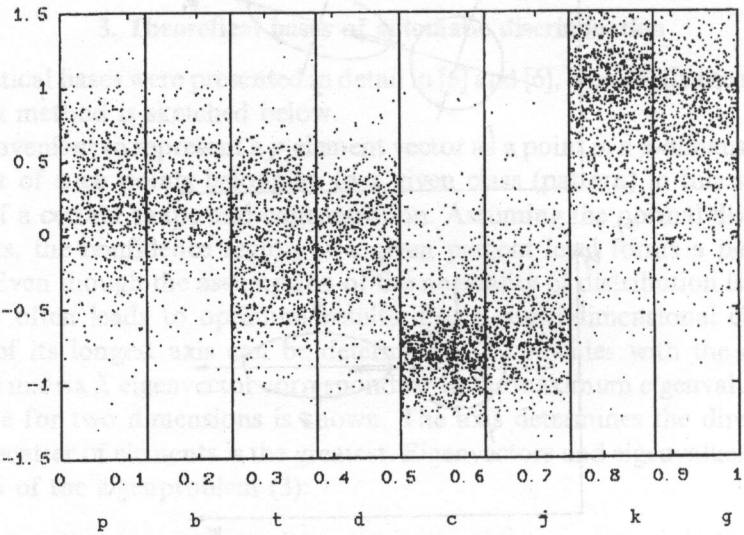

Fig. 3. Projection of all the populations onto the axis determined by the eigenvector.

(vertical axis). The horizontal axis is divided into 8 equal sections, one for each population. Each object is represented by a point whose position on the horizontal axis within the given region is random. Figure 4 shows the populations of /p, t/ bursts which are not readily discriminable. This is envisaged by the large area belonging to both populations. The pair of stops the easiest to discriminate are /c, k/ (Fig. 5). Detailed results obtained for all Polish stops are contained in Tables 1–4. The discrimination process was a multistage one. For each pair of populations, BETWEEN and WITHIN matrices were calculated and eigenproblem solved, whereby the optimum discrimination axis was determined. The subsequent step was to perform projections onto this axis. Their result was presented in a graphic form. Boundary

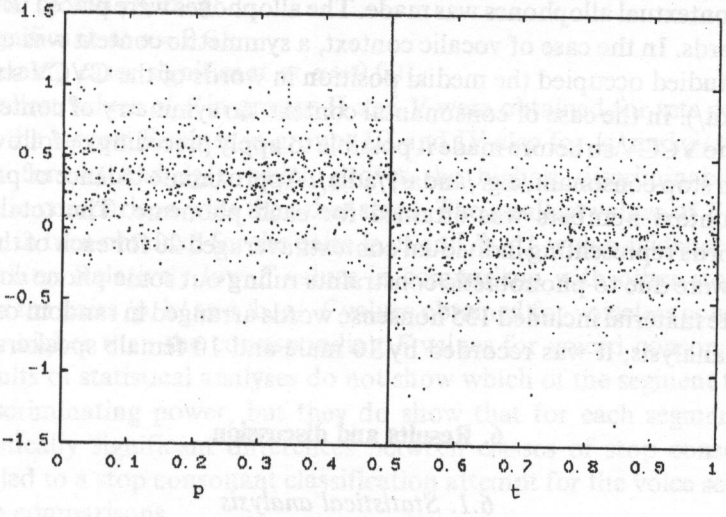

Fig. 4. Populations /p, t/.

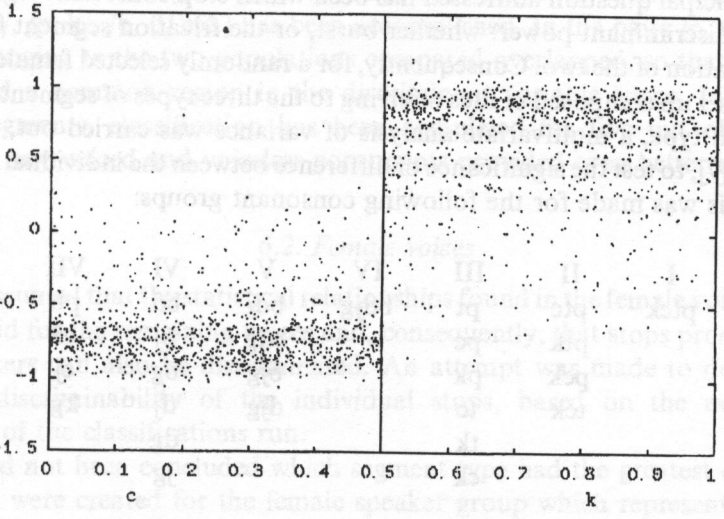

Fig. 5. Populations /c, k/.

values were determined, which made it possible to unequivocally assign each region to a population. Alien elements present in a given region (classification errors) were scrupulously counted. Subsequently, objects already classified were removed from both populations. The procedure was then reiterated (starting from BETWEEN and WITHIN matrix calculation) until no further assignment of a region to any population was possible. At the final stage, unclassified objects were counted.

5. Speech material

For each Polish stop phoneme /p, b, t, d, c, ɟ, k, g/, a list of all phonotactically admissible contextual allophones was made. The allophones were placed in two-syllable nonsense words. In the case of vocalic context, a symmetric context was used, and the consonant studied occupied the medial position in words of the CVCV structure (e.g. /dudu/, /kaka/). In the case of consonantal context, no symmetry of contexts could be obtained. The VCCV structure made it possible to apply preceding or following context for any given stop consonant, e.g. /anda/, /adla/. Approximate balance of preceding and following context numbers was attained for each phoneme. The total number of nonsense words representing individual contexts averaged 20 for each of the stops. The differences were due to phonotactic constraints ruling out some phone combinations. The complete material included 155 nonsense words arranged in random order. For the purpose of analysis, it was recorded by 10 male and 10 female speakers.

6. Results and discussion

6.1. Statistical analysis

The principal question addressed has been which stop consonant segment has the maximum discriminant power: whether burst, or the frication segment following, or the combination of the two. Consequently, for a randomly selected female voice a data base was a set up composed of files referring to the three types of segments mentioned. Within each type, a multivariate analysis of variance was carried out, using Wilks' L criterion [7], to test the significance of difference between the individual populations. The analysis was made for the following consonant groups:

I	II	III	IV	V	VI	VII
ptck	ptc	pt	bdjg	bdj	bd	pb
	ptk	pc		bdg	bj	td
	pck	pk		bjg	bg	cj
	tck	tc		djg	dj	kg
		tk			dg	
		ck			jg	

The following results were obtained:

1) segment type: burst + frication
   all F values significant at $\alpha=0.001$,
2) segment type: burst
   all F values significant at $\alpha=0.001$,
3) segment type: frication
   voiceless consonants:
   all F values significant at $\alpha=0.001$
   voiced consonants:
   for the pairs /b, d/ and /b, g/
   F values not significant
   for /d, g/ and /b, d, g/
   F significant at $\alpha=0.01$,

in the remaining cases significant at $\alpha=0.001$.

The smallest values of F in groups II and V were obtained for sets comprising /c/ and /ɟ/ and the largest F values in groups III and IV also for /c/ and /ɟ/ (considerably larger than for /k/ and /g/), which proves the highest discriminability of these consonants in comparison with the other stops. The smallest values of F in groups III and IV were in turn obtained for the pairs /p, t/ and /b, d/, which evidences their lowest discriminability. Relatively low F values in comparison with other stops were also achieved for the pairs /p, k/ and /b, g/. F values obtained for voiceless consonants were consistently higher than the corresponding F values for voiced consonants.

The results of statistical analyses do not show which of the segment types has the greatest discriminating power, but they do show that for each segment type there occur statistically significant differences between classes of stop consonants. This conclusion led to a stop consonant classification attempt for the voice selected, based on pairwise comparisons.

In the case of burst segments, classification of ten out of twelve consonant pairs (/p, t/, /p, c/, /p, k/, /b, d/ etc.) has been accomplished. In the pairs /p, t/ and /b, d/, objects belonging to the two populations compared overlapped, so that the populations shared a common region in the discriminant variable space. In the case of frication segments, classification has been completed for five out of nine pairs. Comparing any voiced and voiceless consonants provided even better results.

6.2. Female voices

It was assumed that the statistical relationships found in the female voice examined are also valid for the remaining voices and, consequently, that stops produced by the other speakers can also be discriminated. An attempt was made to determine the degree of discriminability of the individual stops, based on the evaluation of correctness of the classifications run.

As it had not been concluded which segment type had the greatest discriminant power, files were created for the female speaker group which represented both the burst and frication segments.

Consonant classification were run in 28 burst file pairs and 28 frication file pairs. Results referring to the burst segments are shown in Tables 1 a)–c).

Table 1 a) provides the total numbers of misclassifications of objects belonging to the individual populations i and j (e_{ij}) for discrimination values adopted in consecutive steps.

Table 1 b) presents, for the individual populations, the values of recognition coefficient r_{ij} determined from the dependence:

Table 1. Evaluation of stop consonant classification — bursts, female voices.

a) Number of errors for the individual populations e_{ij}.

pop.	size	p	b	t	d	c	j	k	g
p	390		322	220	95	17	10	55	108
b	326	148		36	217	7	59	26	139
t	517	191	42		337	24	29	19	22
d	418	156	60	130		52	31	24	50
c	539	56	5	35	52		367	5	48
j	433	5	8	22	20	172		0	3
k	567	219	61	24	17	6	0		410
g	463	204	169	18	50	15	5	245	

b) Recognizability coefficient for the individual populations r_{ij}.

pop.	p	b	t	d	c	j	k	g	mean	st. dev.
p		.174	.436	.756	.956	.974	.859	.723	.697	.293
b	.546		.890	.334	.979	.819	.920	.574	.723	.240
t	.631	.919		.348	.954	.944	.963	.957	.817	.238
d	.627	.856	.689		.876	.926	.943	.880	.828	.121
c	.896	.991	.935	.904		.319	.991	.911	.849	.237
j	.988	.982	.949	.954	.603		1.00	.993	.924	.143
k	.614	.892	.958	.970	.989	1.00		.277	.814	.272
g	.559	.635	.961	.892	.968	.989	.471		.782	.220
grand mean grand standard deviation									.804 .072	

c) Masking coefficients for the individual populations p_{ij}.

p	b	t	d	c	j	k	g
.359	.292	.134	.269	.078	.165	.094	.241
mean standard deviation discrimination coefficient Q						.204 .101 3.943	

$$r_{ij} = 1 - e_{ij}/N_i \tag{5}$$

where N_i denotes population size.

The lowest r_{ij} values (the lowest discriminability) characterize the pairs /p, b/, /k, g/, /t, d/, /c, ɟ/, whereas the highest r_{ij} values (the highest discriminability) — the pairs /ɟ, g/ and /c, k/. The last two columns of Table 1 b) contain population means and standard deviations. /ɟ/ turned out to be the most readily recognizable consonant (on average, 92 per cent of object classified correctly), with /c/ and /k/ doing slightly worse. The lowest recognizability characterized /p/ (70 per cent of objects classified correctly). Grand mean for all the population was 80 per cent correct.

Table 1 c) shows, for the individual populations, the values of masking coefficients p which specifies the number of errors per one population object, i.e. expresses in terms of numbers to what extent objects of a given population "hinder" correct classification of objects from the remaining populations. It was calculated using the data from Table 1 a), as the ratio of the sum of errors in a given column to the size of the population associated with this column multiplied by 7. For example, the overall number of errors caused by the objects of population /g/ in the other populations was 780 (108 errors in population /p/, 139 errors in population /b/ etc.). This makes 0.240 errors (780/7*463) per one population /g/ object (population /g/ size is multiplied by 7, since that was the number of times this population was involved in classification).

It follows from Table 1 c) that the greatest number of errors were caused by objects belonging to population /p/ ($p=0.358$), and the smallest by objects from population /c/ ($p=0.077$).

The overall evaluation of stop consonant recognition rate is provided by the discrimination coefficient value Q, expressed as the ratio of mean r to mean p. For burst segments in female voices $Q=3.0055$.

Tables 2 a)–c) refer to frication segments. Here, the distribution of the highest and lowest values is somewhat different than in Table 1, and the mean r values are strikingly consistently lower. Consequently, grand mean in Table 2 (0.631) is considerably lower than that in Table 1 (0.804).

Table 2. Evaluation of stop consonant classification — frications, female voices.

a) Number of errors for the individual populations e_{ij}.

pop.	size	p	b	t	d	c	ɟ	k	g
p	638		208	453	165	49	61	545	159
b	95	88		59	66	15	6	80	81
t	971	341	223		237	143	164	325	157
d	235	154	115	129		62	62	187	42
c	1862	172	62	1089	59		946	76	69
ɟ	553	124	2	425	84	484		225	3
k	1252	771	295	527	308	92	61		553
g	252	182	115	165	57	118	4	179	

b) Recognizability coefficient for the individual populations r_{ij}.

pop.	p	b	t	d	c	j	k	g	mean	st. dev.
p		.674	.290	.741	.923	.904	.146	.751	.633	.300
b	.074		.379	.305	.842	.937	.158	.147	.406	.347
t	.649	.770		.756	.853	.831	.665	.838	.766	.083
d	.345	.511	.451		.736	.736	.204	.821	.543	.229
c	.908	.967	.415	.968		.492	.959	.963	.810	.246
j	.776	.996	.231	.848	.125		.593	.995	.652	.353
k	.384	.764	.579	.754	.927	.951		.558	.703	.207
g	.278	.544	.345	.774	.532	.984	.290		.535	.265
grand mean grand standard deviation									.631 .133	

c) Masking coefficients for the individual populations p_{ij}.

p	b	t	d	c	j	k	g
.410	1.534	.419	.593	.074	.337	.185	.603
mean standard deviation discrimination coefficient Q						.519 .448 1.215	

The values of coefficient p exhibited in Table 2 are, in turn, higher than those in Table 1, the exception being /c/, for which these values are equal. Analysis of results shows that frication segments are less readily recognizable than burst segments, which is evidenced by the lower Q value in Table 2 than in Table 1.

6.3. Male voices

Of the three segment types considered, the burst segment has been selected for discrimination purposes, as it proved more readily recognizable than the frication segment. Taking into account the third segment (burst + frication) was considered pointless, since its second element would lead to a deterioration of classification results.

Tables 3 a)–c) present evaluation of burst segment classification in male voices. The mean r value is slightly lower and the mean p value slightly higher than in female voices, which results in the lower Q value. Results indicate a somewhat poorer recognizability rate in this speaker group. As in female voices, /ɟ/ is the easiest consonant to recognize and /c/ shows the smallest masking effect.

Table 3. Evaluation of stop consonant classification — bursts, male voices.

a) Number of errors for the individual populations e_{ij}.

pop.	size	p	b	t	d	c	j	k	g
p	346		54	149	61	30	22	106	109
b	193	58		14	127	22	20	13	181
t	393	164	35		166	82	10	80	53
d	332	108	169	85		50	21	58	90
c	531	178	21	122	70		249	73	27
j	345	75	38	17	46	36		16	15
k	556	97	69	66	47	43	26		377
g	443	106	118	52	102	14	12	95	

b) Recognizability coefficient for the individual populations r_{ij}.

pop.	p	b	t	d	c	j	k	g	mean	st. dev.
p		.844	.569	.824	.913	.936	.694	.685	.781	.135
b	.699		.927	.342	.886	.896	.933	.062	.678	.344
t	.583	.911		.578	.791	.975	.796	.865	.786	.154
d	.675	.491	.744		.849	.937	.825	.729	.750	.144
c	.665	.960	.770	.868		.531	.863	.949	.801	.157
j	.783	.890	.951	.867	.896		.954	.957	.899	.063
k	.826	.876	.881	.915	.923	.953		.322	.814	.221
g	.761	.734	.883	.770	.968	.973	.786		.839	.101
grand mean grand standard deviation									.793 .065	

c) Masking coefficients for the individual populations p_{ij}.

p	b	t	d	c	j	k	g
.325	.373	.184	.266	.075	.149	.113	.275
mean standard deviation discrimination coefficient Q						.220 .106 3.608	

6.4. Voices pooled

Files containing bursts from male and female voices were pooled, and classification was run using the combined data. Joint results for all the voices are presented in Tables 4 a)–c). Grand mean of recognizability coefficient is equal to

Table 4. Evaluation of stop consonant classification — bursts, all voices.

a) Number of errors for the individual populations e_{ij}.

pop.	size	p	b	t	d	c	j	k	g
p	736		540	274	334	85	49	222	477
b	519	167		74	323	7	17	40	237
t	910	535	101		419	104	108	157	287
d	750	319	275	261		45	108	80	111
c	1070	173	44	343	170		351	96	97
j	778	13	93	240	91	219		31	140
k	1123	337	276	80	131	64	49		604
g	906	380	447	173	108	69	64	327	

b) Recognizability coefficient for the individual populations r_{ij}.

pop.	p	b	t	d	c	j	k	g	mean	st. dev.
p		.266	.628	.546	.885	.933	.698	.352	.615	.251
b	.678		.857	.378	.987	.967	.923	.543	.762	.235
t	.412	.889		.540	.886	.881	.827	.685	.731	.192
d	.575	.633	.652		.940	.856	.893	.852	.772	.147
c	.838	.959	.679	.841		.672	.910	.909	.830	.113
j	.983	.880	.692	.883	.719		.960	.820	.848	.112
k	.700	.754	.929	.883	.943	.956		.462	.804	.180
g	.581	.507	.809	.881	.924	.929	.639		.753	.175
grand mean grand standard deviation									.764 .072	

c) Masking coefficients for the individual populations p_{ij}.

p	b	t	d	c	j	k	g
.373	.489	.227	.300	.079	.137	.121	.308
mean standard deviation discrimination coefficient Q						.254 .140 3.006	

0.764, which indicates that 76 per cent of objects were classified correctly. As in the two speaker groups treated separately, /j/ was the most readily recognized consonant (85% correct), and /c/ showed the smallest masking effect. /p/ turned out the most difficult stop to recognize (62 per cent of correct classifications).

Recognizability of the individual stops, expressed by means of the coefficient r_{ij}, is illustrated in Fig. 3 which shows projections onto the discrimination axis of all the objects belonging to populations representing bursts of the particular stops.

7. Final remarks

The presented method of stop consonant discrimination, based on autocorrelation function values for burst segments, makes it possible to observe relations between particular stops. Statistical analysis and classification results show that the greatest similarity characterizes /p, b, t, d/ (see Fig. 3), and the greatest dissimilarity /c, ɟ/ and /k, g/, /c, ɟ/ being the most distinct of all the stops. Since this tendency occurred at every stage of the experiment (single voice, female voices, pooled voices), it can be considered a general rule. In the applied method of stop discrimination, palatalness of the consonant plays the most essential role, with proximity of place of articulation coming next.

The results of classification can be regarded as quite good, considering the phonetic material used: stop consonants are known to pose special identification problems in automatic speech recognition.

The ultimate verification of the method will be in segmental recognition of the speech signal based on adopted discrimination threshold values.

References

[1] L.C. LIU, L.M. LEE, H.C. WANG, Y.C. CHANG, *Layered neural nets applied in the recognition of voiceless unaspirated stops*, IEE PROCEEDINGS–1, **138**, 2, 69–75 (1991).

[2] K. FORREST, G. WEISMER, P. MILENKOVIC, R.N. DOUGALL, *Statistical analysis of word-initial voiceless obstruents: Preliminary data*, JASA, **84**, 1, 115–123 (1988).

[3] D. KEWLEY-PORT, *Time-varying features as correlates of place of articulation in stop consonants*, JASA, **73**, 1, 322–334 (1983).

[4] Z. NOSSAIR, S. ZAHORIAN, *Dynamic spectral shape features as acoustic correlates for initial stop consonants*, JASA, **89**, 6, 2978–2991 (1991).

[5] P. DOMAGAŁA, *Automatic segmental recognition of Polish words* (in Polish), unpublished doctoral thesis, Reports of IFTR PAS, Warsaw 1991.

[6] K. FUKUNAGA, *Introduction to Statistical Pattern Recognition*, Academic Press, New York 1972.

[7] M. TATSUOKA, *Multivariate analysis*, John Wiley & Sons, New York 1971.

Received April 14, 1993, English version October 15, 1993