MAN – MACHINE COMMUNICATION: SPEAKER – INDEPENDENT SPEECH RECOGNITION

INFORMATICA 1/88

UDK 681.3:534.44

Zdravko Kačič
Bogomir Horvat
Štefan Greif
Faculty of Technical Sciences, Maribor

Abstract. With a proper selection of feature description methods sufficient accuracy of the speaker - independent speech recognition should be achieved. The speech signal features are described with the three sets of feature (the set of descriptive features , the set of selected features, and the set of characteristic features). The feature description methods are described with the three sets of map (the set of descriptive features map, the set of selected features map , and the set of characteristic features map). As an example two feature description methods are dismembered - zero - crossing method (variant a and b) and method of formant frequencies energy classes (variant a and b). It has been shown that the Fourier transformation as a map of descriptive features was more convinient as a measurement of interval lenght between two succesive zero-crossings of the signal. The mapping rule in variant b of the method of formant frequencies energy classes was more convinient map of selected features than the mapping rule in variant a. With these more convinient maps the smallest feature overlapping and consequently a better average recognition accuracy (greater than 92.5%) has been achieved.

Keywords. Speech recognition, independent speaker, recognition base element, set of features, set of maps, recognition accuracy, feature description, feature overlapping.

1. Introduction

In spite of fast develqpment of computer tecnology, digital signal processing theory, phonetics , linguistics and artiffical intelligence, solution of the problem regarding man - machine communication on the basis of speaker-independent speech recognition, remains entirely the job of the feature.

To solve this problem a very good knowledge of all above mentioned fields shall be required .

Nowadays commercial speech recognition systems recognize successfully a large vocabulary of words only in the case of isolated word recognition and are mostly dependent on speaker [10]. In systems which recognize connected speech or even continuous speech the vocabulary of words is much smaller.

A special problem represent systems which recognize the speaker-independent speech signal.

In systems which recognize isolated words the extent of vocabulary decreases already (on about 40 words). Of course, the same recognition accuracy as in the speaker-dependent systems shall be required.

Today the speaker-independent continuous speech recognition systems exist as prototypes only and their vocabulary is not greater than 10 words [10].

The 'complexity', and first of all the great 'heterogeneity' of speaker-independent speech signal represent one of the major obstacles for solving this problem more successfuly.

The speech signal can be recognized on the basis of the so called recognition base elements (words , syllables, phonemes etc.).

This paper describes some problems which appear in the process of speaker-independent speech recognition on the basis of phoneme recognition, and indicates the ways of their solution .

'Features overlapping' of different recognition base elements (i. e. when features were described by feature extraction methods and presented in n-dimensional space) and great dispersion of features of same base element (i.e. when spoken by an independent speaker) represent a great problem in the speaker independent speech recognition process.

Features overlapping mostly means recognition error when the classification is made.

Speech signal characteristics can be described by various features extraction methods [1 ,3,6, 7].

Differences in speech signal features of the same recognition base element (for an independent speaker) are due to speaker's age, sex , psychophysical condition , etc [1,2,6].

Different feature extraction methods describe speech features in different ways.Consequently, the rate of features overlapping is different and depends upon the method which has been used.

To achieve high recognition accuracy of recognition base elements , the features overlapping of different base elements should be as small as possible.

So the proper selection (definition) of a feature extraction method for particular groups of recognition base elements is an important condition for a good recognition accuracy.

In the next sections the feature extraction process of recognition base elements with definition of some mapping rules and features sets shall be described and the basic notion with dismembers of two feature extraction methods - zero - crossing method (variant a and b) and method of formant frequencies energy classes (variant a and b) shall be presented.

2. Description of recognition base element features

We shall try to describe feature extraction process by means of three sets of speech signal features and three sets of map.

The three sets of features are : the set of all descriptive features, the set of all selected features and the set of all characteristic features. Each of the sets should be mapped with the following mapping sets : the set of descriptive features maps, the set of selected features maps and the set of characteristic features maps.

Such distribution of speech signal features has been assumed to estimate the convinience of a single feature extraction method which shall be used in the feature extraction process.

Analytic evaluation of the importance of a single feature description and with it a definition of 'optimum' description might also be possible.

Let us describe now briefly single sets of features and the sets of maps.

A) Sets of features

a) Set of the recognition base elements articulation - $\mathcal{A}$.

$$\mathcal{A}=\{A_1,A_2, \dots A_n, \dots A_N\}, \qquad (1)$$

N - the number of different recognition base elements

$$A_n=\{A_{n1},A_{n2}, \dots A_{nm}, \dots \}, \qquad (2)$$

A_{nm} -the m - th articulation of the n-th recognition base element

$$A_{nm}=\{a_{nm1},a_{nm2}, \dots a_{nml}, \dots a_{nmL}\}, \qquad (3)$$

L - the number of windows of the m - th articulation

a_{nml} - the l - th window of the m - th articulation of the n-th recognition base element

$$a_{nml}=\{a^1_{nml},a^2_{nml}, \dots a^u_{nml}, \dots a^U_{nml}\}, \qquad (4)$$

U - the number of elements in the l-th window

b) Set of all descriptive features - D

$$D=\{D^1,D^2, \dots ,D^i, \dots \}, \qquad (5)$$

D^i - the set of descriptive features described by the i-th description

$$D^i=\{D^i{}_1,D^i{}_2, \dots ,D^i{}_n, \dots D^i{}_N\}, \qquad (6)$$

$D^i{}_n$ - the set of descriptive features of the n-th recognition base element described by the i-th description

$$D^i{}_n=\{D^{i1}{}_n,D^{i2}{}_n, \dots ,D^{im}{}_n, \dots \}, \qquad (7)$$

$D^{im}{}_n$ - the set of descriptive features of the m-th articulation of the j-th recognition base element

$$D^{im}{}_n=\{D^{im}{}_{n1},D^{im}{}_{n2}, \dots D^{im}{}_{nl}, \dots D^{im}{}_{nL}\}, \qquad (8)$$

L - the number of windows of the m - th articulation of the j - th recognition base element

$D^{im}{}_{nl}$ - the set of descriptive features of the l-th window of the m-th articulation of the n - th recognition base element

$$D^{im}{}_{nl}=\{d^{im}{}_{nl1},d^{im}{}_{nl2}, \dots ,d^{im}{}_{nlk}, \dots d^{im}{}_{nlK}\}, \qquad (9)$$

K - the number of descriptive features

c) Set of all selected features - S

$$S=\{S^1,S^2, \dots ,S^j, \dots \}, \qquad (10)$$

S^j- the set of selected features defined by the j-th description

$$S^j=\{S^j{}_1,S^j{}_2, \dots ,S^j{}_n, \dots S^j{}_N\}, \qquad (11)$$

N - the number of different recognition base elements

$S^j{}_n$ - the set of selected features of the n-th recognition base element defined by the j-th description

$$S^j{}_n=\{S^j{}_{n1},S^j{}_{n2}, \dots ,S^j{}_{np}, \dots\}, \qquad (12)$$

$S^j{}_{np}$ - the p-th selected feature of the n-th recognition base element defined by the j-th description

$$S^j{}_{np}=\{s^j{}_{np1},s^j{}_{np2}, \dots ,s^j{}_{npR}\}, \qquad (13)$$

R - the number of elements of the p-th selected feature

d) Set of all characteristic features - C

$$C=\{C^1,C_2, \dots ,C^u, \dots \}, \qquad (14)$$

C^u - the set of characteristic features defined the u-th description

$$C^u=\{C^u{}_1,C^u{}_2, \dots ,C^u{}_n, \dots ,C^u{}_N\}, \qquad (15)$$

N - the number of different recognition base elements

C^u_n – the characteristic feature of the n-th recognition base element defined by u-th description

$$C^u_n=\{c^u_{n1},c^u_{n2}, \dots ,c^u_{nv}, \dots c^u_{nV}\}, \quad (16)$$

V – the number of elements of the characteristic feature

B) Sets of maps

1) Set of descriptive feature maps – F_D

- elements of the set are mapping the set of recognition base elements articulation into the set of descriptive features

$$F_D=\{f_{D1},f_{D2}, \dots ,f_{Di}, \dots \}, \quad (17)$$

$$f_{Di}:\mathcal{A} \longrightarrow D^i \quad (18)$$

The map $f_{Di}:\mathcal{A} \longrightarrow D^i$ is surjective.

2) Set of selected feature maps – G_s

- elements of the set are mapping the set of descriptive features into the set of selected features

$$G_s=\{g_{s1},g_{s2}, \dots ,g_{sj}, \dots \}, \quad (19)$$

$$g_{sj}:D^i \longrightarrow S^j \quad (20)$$

The map $g_{sj}:D^i \longrightarrow S^j$ is surjective.

3) Set of characteristic feature maps – G_c

- elements of the set are mapping the set of descriptive features into the set of characteristic features

$$G_c=\{g_{c1},g_{c2}, \dots ,g_{cn}, \dots \}, \quad (21)$$

$$g_{cu}:D^i \longrightarrow C^u \quad (22)$$

The map $g_{cu}:D^i \longrightarrow C^u$ is surjective.

The map g_{cu} is mapping the set of descriptive features D^i into the set of characteristic features C^u so that the set of all intersection of the elements of set C^u is an empty set.

The set of all intersections of the elements of selected features set – S^j is not an empty set.

That means, that the elements of characteristic features set C^u are disjunctive sets . This is not valid for the elements of selected features set S^j.

If $f_{Di}:\mathcal{A} \longrightarrow D^i$ and $g_{cu}:D^i \longrightarrow C^u$ are maps, then we may compose f_{Di} and g_{cu} to obtain a map $f_{Di} \circ g_{cu}:\mathcal{A} \longrightarrow C^u$.

We shall define such maps , which are mapping the set of recognition base elements articulation into the set of characteristic features.

3. An example of feature extraction method dismembers

Considering the maps and sets mentioned below, as an example the two feature extraction methods shall be dismembered. The first one is the so called zero-crossing method (method from the time domain) and the second one is the method of formant frequencies energy classes (frequency domain).

There are various variants of the zero-crossing method [5].

Almost all have in common the mapping rule of descriptive features , i.e. measuring the time between the two successive zero-crossings of a signal.

Single variant 'evaluates' these intervals in different ways.

We shall briefly describe two of them.

Elements of the descriptive features set are defined as:

$$d_k=\sum_j T_s \quad , k = 1,2, \dots , K \quad (23)$$

where:

T_s is the time between two successive samples
j is the number of samples with equal sign
d_k is the lenght of k-th interval
K is the number of intervals
In this way , the set of descriptive features D^{im}_n is composed of subsets which contain lenght of intervals between two successive zero-crossings.

$$D^{im}_{nj}=\{d^{im}_{nj1},d^{im}_{nj2}, \dots ,d^{im}_{njk}, \dots d^{im}_{njK}\}, \quad (24)$$

Variant a (ZCa)

Elements of the selected features set S^1_{np} are defined as:

$$s^1_{np}(\tau_j,\tau_{j+1}) = \frac{d(\tau_j,\tau_{j+1})}{P}, \quad (25)$$

where:

-$d(\tau_j,\tau_{j+1})$ is the number of intervals in the time class (τ_j,τ_{j+1})

Value of P is defined by

$$P= \sum_{j=0}^{k-1} d(\tau_j,\tau_{j+1}), \quad (26)$$

K is the number of all intervals .

The subset S^1_{np} of selected features set S^1_n is composed of elements which represent portion of intervals lenght in particular time classes.

$$S^1_{np}=\{s^1_{np1},s^1_{np2}, \dots s^1_{npR}\}, \quad (27)$$

variant b (ZCb)

Secondly , elements of the selected features set S^k_{np} are defined as follows:

$$s^k_{np}(\tau_j,\tau_{j+1})=\frac{n(\tau_j,\tau_{j+1}).(\tau_j+\tau_{j+1})/2}{W.(\tau_{j+1}-\tau_j)}, \quad (28)$$

where

$n(\tau_j,\tau_{j+1})$ is the number of intervals in the time class (τ_j,τ_{j+1})
W is the window width
τ_j, τ_{j+1} are the boundary values of the j-th time class.

By means of factors $(\tau_j,\tau_{j+1})/2$ and $(\tau_{j+1}-\tau_j)$ a better evaluation of high and low frequency components should be achieved.

$$S^k_{np}=\{s^k_{np1},s^k_{np2}, \dots ,s^k_{npT}\} \qquad (29)$$

b. Method of formant frequencies energy classes (FFEC)

Like the zero-crossing method this method knows various variants as well.

All variants use the discrete Fourier transformation as the mapping rule of the descriptive features [6,9] :

$$G(s)= \sum_{u=0}^{U-1} g(u) \exp(-j2\pi su/U) \qquad (30)$$

The subset D^{jm}_{nL} of the descriptive features set D^{jm}_{n} is composed of frequency samples.

$$D^{jm}_{n1}=\{G^{jm}_{n11},G^{jm}_{n12}, \dots ,G^{jm}_{n1K}\}, \qquad (31)$$

Variant a (FFECa)

To define elements of the selected features set S^t_{np} the following prescription has been used:

$$s^t_{np}(r) =(\sum_{u=f_m/R_f}^{f_{m+1}/R_f} \log G_v^2(u))/(\sum_{u=1}^{K} \log G_v^2(u)); \quad r=1,2, \dots ,R \qquad (32)$$

where

$G_v(u)$ is the u-th element in descriptive features set D^{jm}_{n1}
$s^t_{np}(m)$ is the m - th element in selected features set S^t_{np}
R_f is the resolution factor of DFT
K is the number of all elements in the descriptive features set D^{jm}_{n1}
R is the number of elements in the selected features set S^t_{np}
f_m,f_{m+1} are the boundary values of the m-th formant frequencies class

The selected feature set S^t_{np} is composed of elements , which represent parts of common energy in particular formant frequencies classes.

$$S^t_{np}=\{s^t_{np1},s^t_{np2}, \dots s^t_{npR}\}, \qquad (33)$$

Variant b (FFECb)

This variant defines elements of the selected features set S^v_{np} as follows:

$$s^v_{np}(j)=\log G_{vmax}^2(j)/(\sum_{m=1}^{M} \log G_{vmax}^2(m)), \qquad (34)$$

where

$G_{vmax}^2(j)$ is the maximum frequency component in j-th formant frequencies class
M is the number of maximum components of all classes and the number of elements in the selected features set S^v_{np}

The subset of the selected features set S^v_{np} is composed of elements , which represent a portion of maximum frequency components in a single formant frequencies class.

$$S^v_{np}=\{s^v_{np1},s^v_{np2}, \dots s^v_{npM}\} \qquad (35)$$

Out of it arises a question how efficiency , or better 'convenience' of a single map should be estimated in order to be used in the base element recognition process .

For this purpose , the recognition results obtained by the dismembered feature extraction methods mentioned above , will be presented in the next sections.

4. Experimental results of isolated Slovene vowels recognition

The recognition of the five isolated Slovene vowels (/a/,/e/,/i/,/o/ and /u/) was carried out by the recognition experiment.

One hundred and ten articulations of each vowel, pronounced by 110 different speakers, has been performed. All articulations were recorded in an studio environment.

Speakers were of different age categories. Female - male rate was 3/7.

The speech signal was passed through a band-pass filter (600 Hz - 3.4kHz) and sampled at 10k Hz with a 12 bit A/D converter.

The time window width (W) was limited to 20 ms.

Because of such a great amount of different speakers we might presume that the recognition results (see Table 1) are the recognition results of an independent speaker.

Elements of the selected features set (for a single method) were combined into the feature vector and the number of vector elements was limited to ten :

$$Z_{np}=[z(1),z(2), \dots ,z(10)], \qquad (36)$$

Classification was made on the bassis of multivariate normal distributions with equal covariances [7].

Recognition results for single methods are given in the Table 1a-b.

	Zero - Crossing Method									
	variant a					variant b				
	recognized as [%]									
	A	E	I	O	U	A	E	I	O	U
A	96.4	1.8	0.0	1.8	0.0	97.3	0.9	0.0	1.8	0.0
E	0.0	71.8	13.7	5.5	9.0	0.9	78.2	10.0	7.3	3.6
I	0.0	5.4	90.9	0.0	3.7	0.0	3.7	94.5	0.0	1.8
O	6.4	24.5	0.9	62.7	5.5	7.3	22.7	0.0	63.6	6.4
U	1.9	11.8	11.8	10.9	63.6	2.7	11.8	17.3	10.9	57.3

a

	Method of Formant Frequencies Energy Classes									
	variant a					variant b				
	recognized as [%]									
	A	E	I	O	U	A	E	I	O	U
A	83.6	1.8	0.0	8.2	6.4	97.3	0.0	0.0	2.7	0.0
E	6.3	70.0	16.4	2.7	4.6	0.0	92.8	5.4	0.9	0.9
I	3.6	16.6	69.0	3.6	7.2	0.0	4.5	92.8	0.0	2.7
O	13.8	4.5	1.8	58.1	21.8	1.8	0.0	0.0	92.8	5.4
U	5.5	7.2	1.8	10.1	75.4	0.0	0.9	0.0	10.0	89.1

b

Table 1a-b: Experimental results of five isolated Slovene vowels recognition

5. Efficiency of feature extraction methods

We shall now try to estimate efficiency of single maps, or better, their 'convenience' for the use in the base elements recognition process on the basis of recognition results.

By using map rules in the zero-crossing method (variant a) a somehow better recognition accuracy was achieved only for the vowel /a/ (96.4%) - less for the vowel /i/. For the vowels /e/ , /o/ and /u/ a rather worse recognition accuracy was achieved.

The variant b of the zero-crossing method showed a little bit better recognition results, but the rate of vowels recognition error was rather the same as at the variant a.

The reason for a worse recognition accuracy when zero-crossing method was applied , should be searched in the usage of the map of descriptive features.

In this method (for both variants) the measurement of intervals lenght as mapping rule for mapping the descriptive features was used.

Anyhow, this 'function' is 'incapable' to 'ignore' phase changes between particular frequency components in a signal.

In other words - it is a phase dependent function.

Human ear is insensitive to phase changes in a speech signal [4], whereas this is not true for the 'simple' measurements of intervals lenght.

Two signals with equal frequency components and with different phases sound the same. However, they can be formed in very different subsets of descriptive features , if the rule of the measuring interval lenght between the two successive zero-crossings of the signal was used as the mapping rule.

This is of great importance for phase changes at low frequencies (first two formants), which have ussualy the greatest amplitude and as such a greater influence on the zero-crossing rate.

Fig.1a shows the first three elements of the feature vector formed by the zero-crossing method (variant a) and the method of formant frequencies energy classes (variant b) for all articulations of the vowel /e/. They 'describe' low frequencies in the frequency spectrum. Fig. 1b represents the last three elements of the feature vector for all articulations of the vowel /e/ , for the both methods. They describe high frequencies in the frequency spectrum.

It could be noticed, that the dispersion of the first three elements of the feature vector formed by the ZCa method (marked by ^), is much greather than the dispersion of the feature vector elements formed by the FFECb method (they are labeled as .).

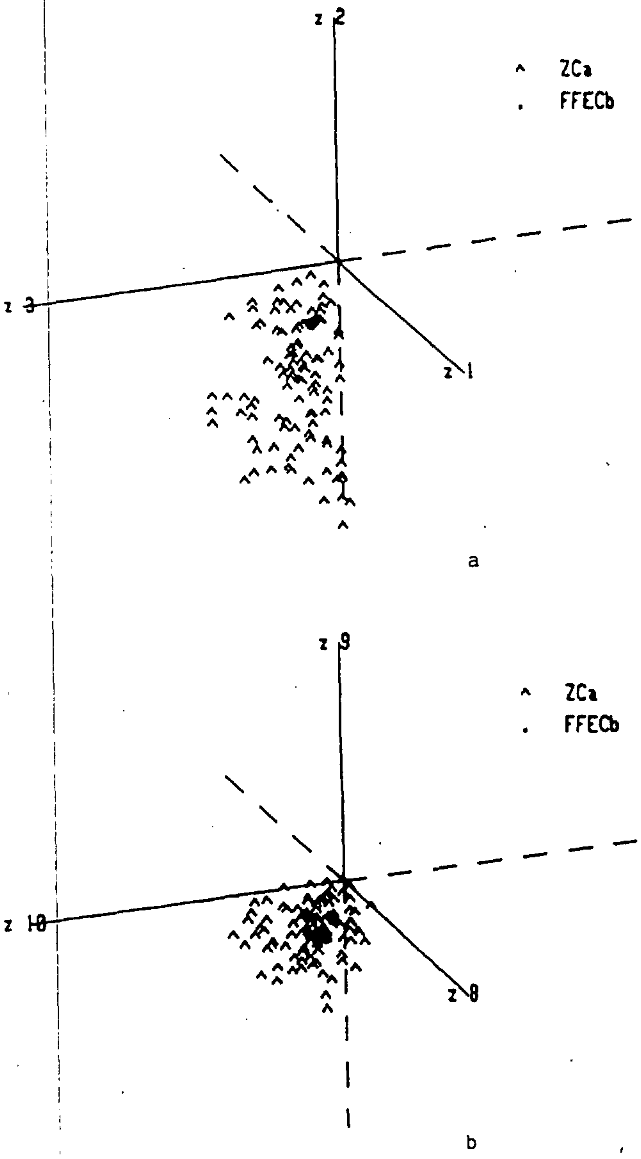


Fig. 1a-b : Distribution of the first three a) and the last three b) elements of the feature vector, for vowel /e/, formed by ZCa (^) and FFECb (.) methods.

A rather smaller dispersion could be seen at the last three elements of the feature vector formed by the ZCa method.

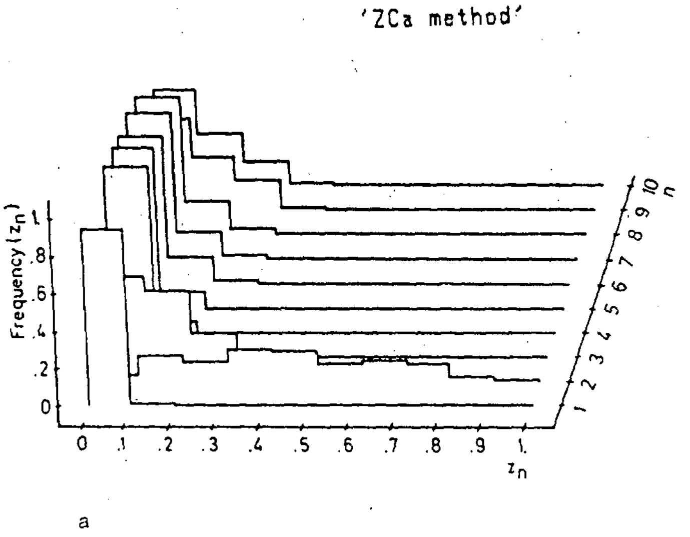
In the both cases the dispersion of feature vectors elements formed by the FFECb is very similar.

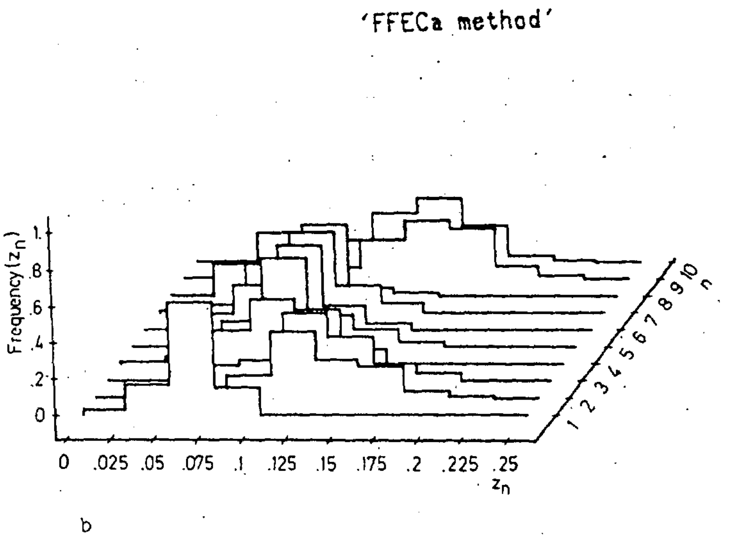
From the above mentioned the importance of the fact of phase changes between single frequency components in the frequency spectrum might be noticed - first of all , for low frequencies being present in a speech signal of an independent speaker.

This fact also indicates the recognition results of the vowels /o/ and /u/, for which first of all the first formant is dominant.

From the Fig. 2a it can be also seen, that features description of recognition base elements with measurement of interval lenght as mapping rule of descriptive features was less successful as with Fourier transformation. This should be evident from the dispersion rate of single feature vector elements , which is greater than the one for the other two methods. This is particulary true for the second and the third element of the feature vector (they first of all describe the first formant).

Comparision of recognition results for variants FFECa and FFECb (see Table 1b) and considerations of dispersion rates of vector elements for both variants (Fig. 2 b - c) give indication of the fact that common normalized energy of single formant frequencies classes calculated by this variant was a 'worser criteria' than the ratio of normalized energy of maximum components was. This might point out that the common energy contents per single formant frequencies classes for some recognition element change with an independent speaker. It was reflected as an increase of dispersion for almost all elements of the feature vector (Fig. 2b). This means a worse recognition accuracy (Table 1b).





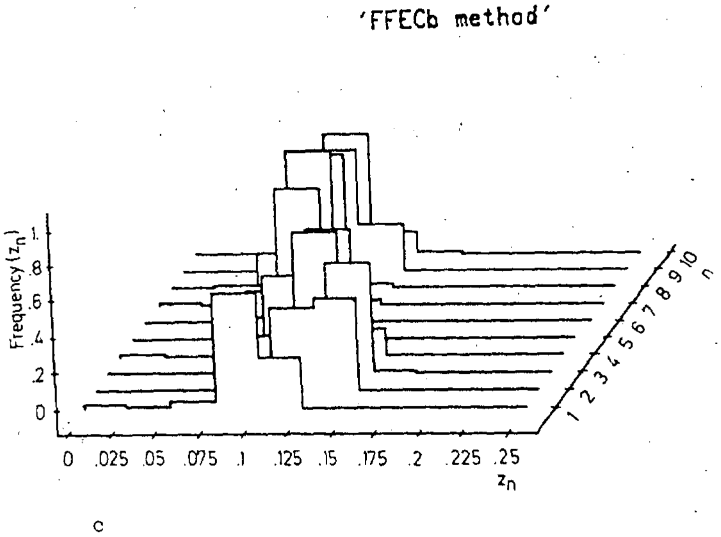


Fig. 2a-c : Histograms of the feature vector elements , for vowel /e/, formed by ZCa a), FFECa b) and FFECb c) methods.

A better recognition accuracy and the smallest features vector elements dispersion was achieved when the mapping rule of method FFECb was used.

The mapping rule of the selected features for this variant 'enables selection' of frequency components. In each class only the maximum component was choosen. In this way only energy of the maximum component for a particular class was described. But because of the fact that ten formant frequencies classes were defined, they are not all maximum frequency components of formants.

With this variant the best average recognition accuracy was achieved - greater than 92.5 %.

6. Conclusion

By the speaker-independent speech recognition such features maps should be defined that 'differences' in speech features, appearing in the case of an independent speaker shall be expressed as small as possible. That means that such functions should be defined where features overlapping was as small as possible.
This should be valid for maps of descriptive features (e.g. measurement of intervals lenght - discrete Fourier transformation) and for maps of selected features (e.g. variant a - variant b of FFEC method) as well.

The mapping rules discussed in our paper showed that the discrete Fourier transformation as the mapping rule for the descriptive features maps and the variant b of the FFEC method as the mapping rule for the selected features maps gave the best recognition results.

With above mentioned methods the smallest features overlapping and consequently the best average recognition accuracy has been achieved - i. e. more than 92.5 % .

References

[1] L.R. Rabiner and R. W. Schafer , Digital Processing of Speech Signals, Prentice - -Hall , Englewood Cliffs , NJ , 1978.

[2] A. H. Seidman and I. Flores , Handbook of Computers and Computing , Van Nostrand Reinhold Company , New York , 1984.

[3] R. De Mori and C.Y. Suen , New Systems and Arhitectures for Automatic Seech Recognition and Synthesis , Springer - Verlang, Berln, 1985, Chap. 1, pp. 1 - 72 .

[4] James C. Anderson , "Improved zero-crossing method enhances digital speech " , EDN Magazine , vol. 27, No. 20 , october 13 1982 , pp. 171 - 174 .

[5] R.J. Niederjohn and P.F. Castelaz, "Zero - crossing analysis methods and their use for automatic speech recognition " ,Proc. IEEE Computer Society Workshop on Pattern Recognition and Artifical Intelligence, 1978 , pp. 274 - 281 .

[6] F. Fallside and W.A. Woods, Computer speech processing , Prentice - Hall , Englewood Cliffs , NJ , 1985

[7] J. C. Simon, Spoken Language Generation and Understanding,D. Reidel Publishing Company, 1980 , pp. 129 - 145

[8] R.J. Senter,Analysis of Data, Scot,Foresman and Company,Illinois , 1969 .

[9] I. H. Witten, "Digital storage and analysis of speech", Wireless world, november 1981, pp. 44 - 48 .

[10] P. Willich, "Putting speech recognizers to work" , IEEE Spectrum , april 1987 , pp. 55 - 57 .

[11] Z.Kačič, Š.Greif and B.Horvat, "Uspešnost metod opisovanja skupnih značilnosti osnovnih elementov govornega signala", Elektrotehniški vestnik , Vol. 53 (1986), No. 3, pp. 121 - 129 .