ARCHIVES OF ACOUSTICS
16, 3–4, 413–432 (1991)

EVALUATION OF THE EFFECTIVENESS OF PARAMETERS OF THE GLOBAL DESCRIPTION OF WORDS IN SIMPLE AUTOMATIC SPEECH RECOGNITION SYSTEMS

CZ. BASZTURA, W. MAJEWSKI, W. BARYCKI

Institute of Telecommunication and Acoustics Technical University of Wrocław
(50-370 Wrocław, Wybrzeże Wyspiańskiego 27

In this paper, oriented at the so-called simple systems of speech recognition, the effectiveness of 6 sets of parameters in the global description of words in a given vocabulary was analysed. The usability of simple parametrization methods and such parameters as: density of zero-crossings, distribution of intervals between zero-crossings, parameters of two so-called phase planes, spectral parameters in octave and tertiary bands, was investigated and analysed on the basis of sound material for one operator, mainly and a vocabulary from 5 to 40 words. It was proved that such parameters as the density of zero-crossings and the distribution of time intervals between zero-crossings can be applied in simple systems with a vocabulary preferably not exceeding 10 words, unless it would be possible to select certain words from the vocabulary. Parameters of the first phase plane and spectral parameters exhibited positively weak discrimination ability, especially in octave bands. Also the usability of the NN algorithm and the Camberra distance which standarizes parameters in such ASR systems was confirmed.

1. Introduction

As yet the problem of automatic recognition of continuous speech has not found a satisfactory solution [1, 12]. By far better results from the point of view of potential applications have been achieved in investigations concerning the recognition of a limited set of isolated words [5, 6, 7, 8, 9, 10, 11, 14, 15, 16]. DAXTER and ZWICKER, for example describe in paper [7] a simple on-line recognition system, which consists of an input analyser (12 filters in the range 170–10 000 Hz). A vocabulary of 38 words was used. Their duration was standarized to 500 ms, while the actual time ranged from 280 to 1000 ms. Every word was described by five groups of 12 spectral parameters each. It was proved that a 3–4 bit amplitude quantization is sufficient to

Research performed within program CPBP 02.03.9.3.

obtain satisfactory results. The mean recognition error was equal to 5.3% for a full frequency band.

Brown and Rabiner proved in another paper [5] that in definite experimental conditions the joint weighted functions of the energy distance and prediction coefficients reduce the recognition error by 6 to 25% at the average in comparison to the use of prediction coefficients only.

In SAMBURS and RABINER'S paper [16] 10 digits are recognized with the application of developed systems on acoustic level. In this case there was no need to tune the system to speakers voice. Every 10 ms the system measured the energy, zero-crossings, frequencies of poles determined with the use of linear prediction and prediction error. The recognition error, determined for 10 speakers, did not exceed 2.7%. These and other papers indicate that the recognition of isolated words remains in the range of interest of scientists. This is a result of a demand for simple automatic recognition (ASR) systems, which could recognize vocabularies consisting of tens, or even several words. Robotics should be mentioned as the main consumer of simple ASR systems. There are also trials of applying ASR instead of an alphanumeric keyboard, i.e. dat a input into a computer with the use of voice. Telephone inquiry systems are another example of ASR application.

Several approaches to the recognition of isolated words can be distinguished [1, 12, 18]. One of these are presented by KUBZDELA [11]. It consists in the global description of the whole word by a set of parameters. While another approach assumes that all segments, into which the word is divided, must have their references. The second method is much more complex, what also makes measuring systems and decision algorithms more complicated. Yet, the effectiveness of the segmentation method is better, what can lead to a larger recognized vocabulary, independence of the speaker's voice and smaller recognition error.

However, the mentioned above demand for simple ASR systems prefers global descriptions of words. This manifests itself in the form of e.g. one parameter vector for the whole word. Many factors influence the widely understood effectiveness of simple ASR systems [12]. Parametrization, aimed at the extraction of parameters used in the global description of words, is one of the most important factors. Some parametrization methods extract parameters which are time functions, such as the envelope curve; while other methods use parameters averaged over time. Parameters averaged over time have this advantage that they do not have to undergo time normalization e.g. paper by DAXTER and ZWICKER [7].

The aim of this paper was to study chosen parametrization methods from the point of view of their effectiveness of recognition of isolated words from a vocabulary consisting of 40 words. In consequence studies aimed at the determination of the best algorithm and similarity function were carried out also. The relationship between the effectiveness of parameters and size of the vocabulary was investigated. Such a complex approach made it possible to specify more fully the evaluation of the usability of studied parameters in simple ASR systems.

2. Methods

The fundamental acoustic material of these studies was based on a 40-word vocabulary recorded by two operators (first one – 5 times, second one – 2 times). The 40-word vocabulary can be considered among average size vocabularies, while the recognition of e.g. 10 digits refers to small vocabularies. The used vocabulary is presented below together with phonetic transcriptions. The same vocabulary was used at the same time in another paper concerned with automatic segmentation of a speech signal [4]:

JEDEN	(jeden) ("one")	PLUS	(plus) ("plus")
DWA	(dva) ("two")	MINUS	(minus) ("minus")
TRZY	(tʃɨ) ("three")	MNOŻ	(mnuʃ) ("multiply")
CZTERY	(t͡ʃteri) ("four")	DZIEL	(d͡zel) ("divide")
PIĘĆ	(pieɲt͡ɕ) ("five")	POTĘGA	(poteɲga) ("power")
SZEŚĆ	(ʃeɕt͡ɕ) ("six")	KROPKA	(kropka) ("dot")
SIEDEM	(ɕedem) ("seven")	WYNIK	(vɨnik) ("result")
OSIEM	(oɕem) ("eight")	TAK	(tak) ("yes")
DZIEWIĘĆ	(d͡zevjeɲt͡ɕ) ("nine")	NIE	(ɲe) ("no")
ZERO	(zero) ("zero")	NAWIAS	(navjas) ("bracket")
START	(start) ("start")	PROGRAM	(program) ("programme")
STOP	(stop) ("stop")	TEKST	(tekst) ("text")
GOTOWY	(gotovɨ) ("ready")	FUNKCJA	(fuŋkt͡sja ("function")
ZAJĘTY	(zajentɨ) ("occupied")	ARGUMENT	(argument) ("argument")
PISZ	(piʃ) ("write")	A	(a)
CZYTAJ	(t͡ʃɨtaj) ("read")	DO	(do) ("to")
ŁADUJ	(waduj) ("load")	LUB	(lup) ("or")
SKOCZ	(skot͡ʃ) ("jump")	NIECH	(ɲex) ("let")
POWTÓRZ	(poftuʃ) ("repeat")	I	(i) ("and")
ŁĄCZ	(wont͡ʃ) ("join")	APOSTROF	(apostrof) ("apostrophe")

The vocabulary was divided into groups of 10 word: the first group included digits, the second group – mathematical operations, the third group – certain commands, while the fourth group – certain names or elements in programming.

Possibly smallest dimensions of the observation space for individual parametrization methods were accepted, because simple ASR systems must have a compact vector description. For example, for parameters of the zero-crossings density, the space dimension is $P = 40$, for spectral parameters $P = 6$ (octaves) and $P = 16$ (tertiary).

Simple ASR systems should have possibly simplest classifiers which make it possible to realize a quick classifier "learning" process without a need of many repetitions of the same word. Two heuristic algorithms were applied, i.e. NN (nearest neighbour) and NM (nearest mean).

The experimental systems was provided for one operator, but the results of recognition were verified for a second operator.

3. Parameters

On the basis of the principle of system simplification, parameter which do not require time normalization were chosen mainly to globally represent words included in the vocabulary. The set of parameters of zero-crossings of the speech signal is an exception. Observation parameter spaces, x^P, with dimension P are as follows:

- discrete function of zero-crossings density $\varrho_0(p)$ with dimension $P = 40$ and linear normalization of function,
- parameters of the distribution of time intervals between zero-crossings of the speech signal $x(p)$, $P = 8$,
- parameters of the so-called first phase plane $Pf_1(r)$, $R = 8$,
- parameters of the so-called second phase plane $Pf_2(r)$, $R = 8$,
- spectral parameters in octave bands $F_{oct}(p)$, $P = 6$,
- spectral parameters in tertiary bands $F_{ter}(p)$, $P = 16$.

3.1. Parameters of the density function of zero-crossings

The density of zero-crossings is a parameter frequently used in the analysis and recognition of speech signals [1, 2, 8, 17]. The mean density of zero-crossings of a signal in a time interval T_p is expressed by

$$\varrho_0[u(t), T_P] = \frac{1}{T_P} \frac{\int_{-\infty}^{+\infty} f^2 P(f)\, df}{\int_{-\infty}^{+\infty} P(f)\, df} \tag{1}$$

where $P(f)$ – spectrum of signals power density in time interval T_p, f – frequency of signal.

If T_p values are sufficiently small and the signal with length T is divided into P segments, then the value from expression (1) will be a component of a vector

$$\varrho_0(P) = \mathrm{col}\{\varrho_0(1), \varrho_0(2), \varrho_0(3), \ldots, \varrho_0(P)\}. \tag{2}$$

The digital measurement of components of the density of zero-crossings is very simple in practice. It consists in the summation of moments in which samples change

sign in windows with length T_p. Therefore, the length of a vector depends on the number of windows. It is subject to time normalization in the recognition process. Non-linear time normalization would be most advantageous here [1, 18]. However, linear time normalization was applied, because of the simplicity of measurements [1]. Figure 1 presents discrete $\varrho_0(p)$ functions, with linear normalization on the length $P = 40$ for the first four words in the vocabulary, i.e. "jeden", "dwa", "trzy", "cztery".

Detailed parametrization data for $\varrho_0(p)$ is as follows: $f_{pr} = 10\,000$ samples/s, dynamics described with 10 bits, window length $T_p = 10$ ms. Dimension of parameter vector after normalization $P = 40$ for the whole vocabulary and all repetitions.

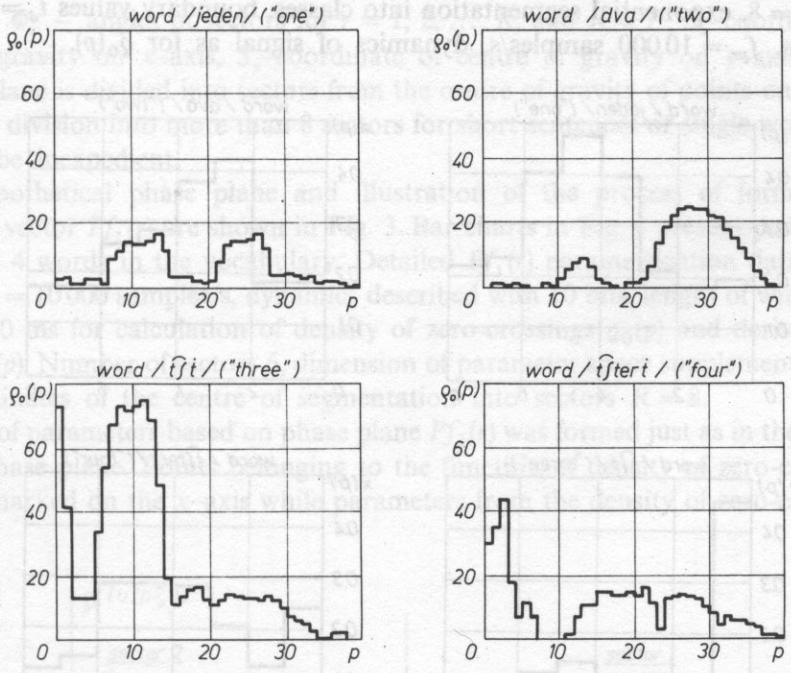

FIG. 1. Examples of parameter ϱ_0(p) values, P = 40, for first four words in the vocabulary

3.2. Parameters of distributions of time intervals between zero-crossings of a speech signal

In spite of its long name parameters of distributions of time intervals between zero-crossings of a speech signal $x(p)$ are a significant "reduction" of the parameter space dimension in relation to the zero-crossings density. The detailed analysis and description of this problem can be found in several papers – [1, 2], among others. It

should be mentioned that

$$x(p) = \mathrm{col}\{x(1), x(2), \ldots, x(P)\} \tag{3}$$

has components calculated according to

$$x(p) = x(t_{p-1}, t_p) = \begin{cases} x(t_{p-1}, t_p)+1 & \text{for } t_j \in (t_{p-1}, t_p\rangle \\ x(t_{p-1}, t_p) & \text{for } t_j \notin (t_{p-1}, t_p\rangle \end{cases} \tag{4}$$

where: $(t_{p-1}, t_p\rangle$ – range of hidden time, called the p-time channel 2, t_j – interval between the j–1 and j zero-crossing of a speech signal.

Figure 2 presents examples of distributions of time interval of the first four words in the vocabulary as in Fig. 1. Detailed $x(p)$ parametrization data are as follows: $P = 8$, exponential segmentation into classes, boundary values $t_d = 0.2$ ms, $t_g = 6.2$ ms, $f_{pr} = 10\,000$ samples/s, dynamics of signal as for $\varrho_0(p)$.

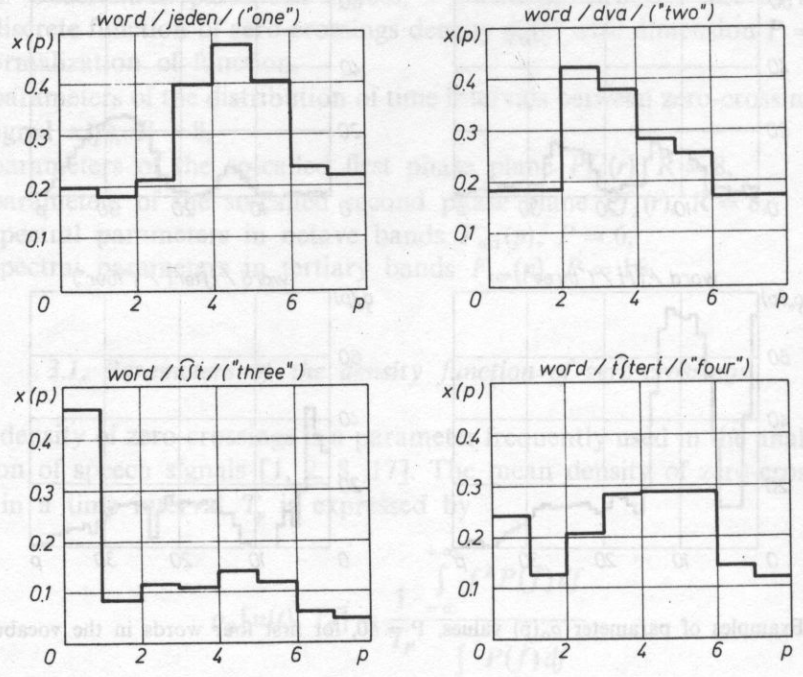

FIG. 2. Examples of parameter x(p) values, $P = 8$, for first four words in the vocabulary

3.3. *Parameters of phase planes Pf_1 and Pf_2*

Two short comings of the density function of zero-crossings of a speech signal, i.e.:

a) large dimension of the parameter vector, and
b) necessity of time normalization

have induced a search for such a modification of this parameter that it would be possible to avoid the shortcomings with possibly smallest effectiveness loss. The idea of new sets of parameters on so-called phase planes originated [1, 2].

It is possible to create a new set of parameters on the first phase plane Pf_1. The number of points in assigned R sectors will be their components. The x– axis of the plane forms the density of zero-crossings $\varrho_0(p)$, the y–axis-derivative of the density of zero-crossings $\varrho_0'(p)$. The principle of forming components of the parameter vectors is as follows

$$Pf_l(r) = Pf_l(r)+l \text{ if}$$

$$\Phi_{r-1} \leqslant \operatorname{Arg}\{\varrho_0[u_n(pT_p)]-S_x, \quad \varrho_0'[u_n(pT_p)]-S_y\} < \Phi_r \tag{5}$$

where Φ_r – angles of sectors for $r = 1, 2, \ldots, R$, and $R \ll P$, S_x–coordinate of centre of gravity on x–axis, S_y–coordinate of centre of gravity on y–axis.

The plane is divided into sectors from the centre of gravity of points on the Pf_1 plane. The division into more than 8 sectors for short sentences or single words was found to be inexpedient.

A hypothetical phase plane and illustration of the process of forming the parameter vector $Pf_1(r)$ are shown in Fig. 3. Bar charts in Fig. 4 present parameters of the first 4 words in the vocabulary. Detailed $Pf_1(r)$ parametrization data are as follows $f_{pr} = 10\,000$ samples/s, dynamics described with 10 bits, length of window T_p equal to 10 ms for calculation of density of zero-crossings $\varrho_0(p)$ and derivative of density $\varrho_0'(p)$. Number of sectors 6, dimension of parameter space supplemented with two coordinates of the centre of segmentation into sectors $R = 8$.

A set of parameters based on phase plane $Pf_2(r)$ was formed just as in the case of the first phase plane. Points belonging to the function of density of zero-crossings $\varrho_0(p)$ are marked on the x–axis while parameters from the density of zero-crossings

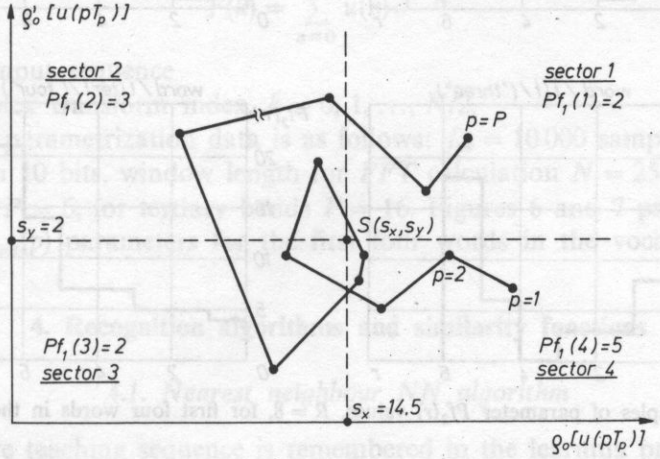

FIG. 3. Method of forming parameters of phase plane $Pf_1(r)$, $R = 4+2$

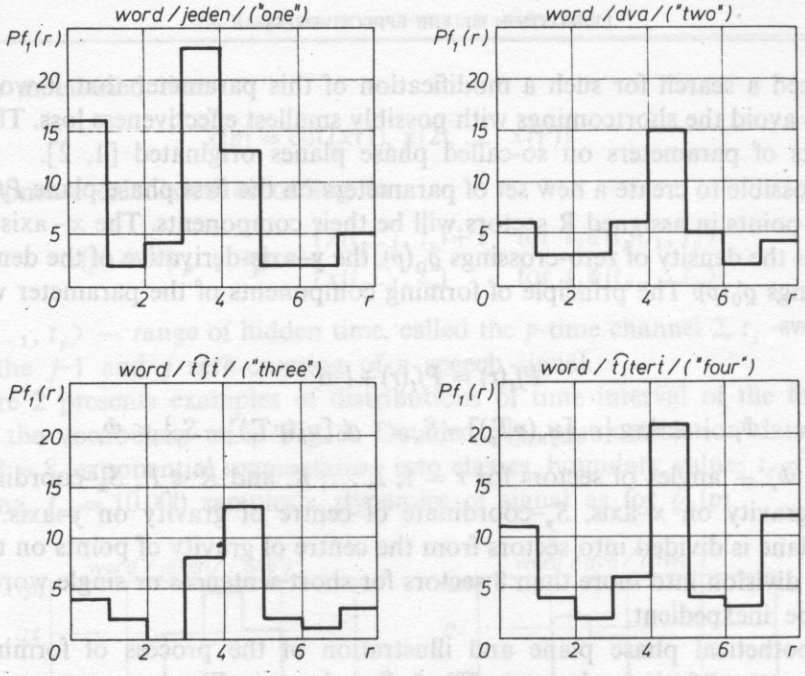

FIG. 4. Examples of parameter $Pf_1(r)$ values, $R = 8$, for first four words in the vocabulary

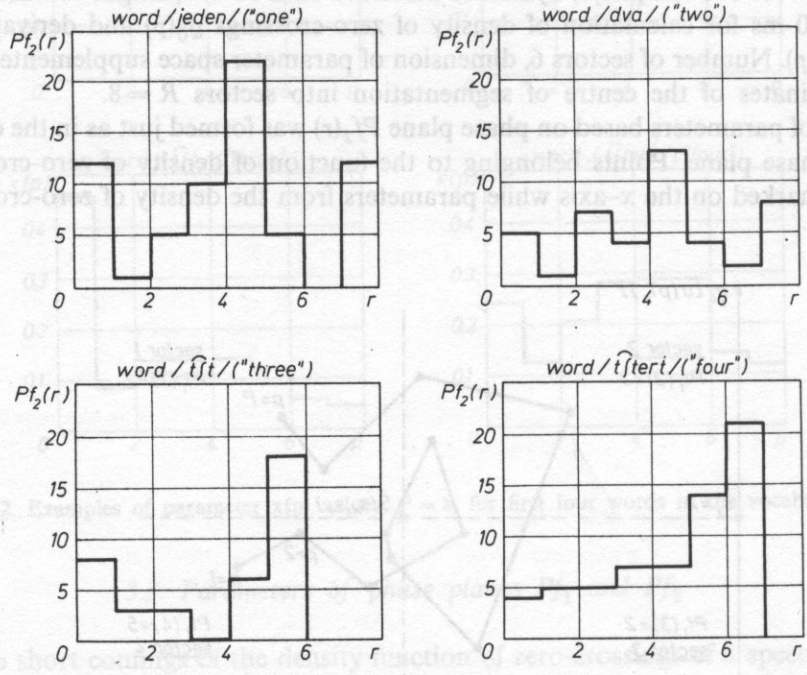

FIG. 5. Examples of parameter $Pf_2(r)$ values, $R = 8$, for first four words in the vocabulary

of the speech signals derivative are marked on the y–axis. The principle of forming components of the parameter vector $Pf_2(r)$ is as follows:

$$Pf_2(r) = Pf_2(r)+1 \quad \text{if}$$
$$\Phi_{r-1} \leqslant \operatorname{Arg}\{\varrho_0[u_n(pT_p)]-S_x, \quad \varrho_0[u'_n(u_n(pT_p)]-S_y\} < \phi_r \tag{6}$$

denotation as in (5).

Figure 5 presents the set of Pf_2 parameters for the 4 first words in the vocabulary. Detailed Pf_2 parametrization data is the same as for $Pf_1(r)$.

3.4. Spectral parameters F

Vectors with components that are signal energies in $P = 6$ and $P = 16$ frequency channels were used as spectral parameters. Components, corresponding with individual bands, were calculated according to expression

$$F(p) = \frac{1}{K_p - P_p + 1} \sum_{j=P_p}^{K_p} [ReF(j)]^2 + [\operatorname{Im} F(j)]^2 \tag{7}$$

where

$$P_p = f_p \frac{N}{f_{pr}}$$

$$K_p = f_{p+1} \frac{N}{f_{pr}}$$

f_p – boundary frequency between the $(p-1)$ and p frequency band, N – number of FFT samples, f_{pr} – signal sampling frequency,

$$F(k) = \sum_{n=0}^{N-1} u(n)\, e^{-j\frac{2\pi}{N}nk} \tag{8}$$

$u(n)$ – real input sequence
$F(k)$ – complex transform index, $k = 0, 1, \ldots, N/2$.
Detailed $F(p)$ parametrization data is as follows: $f_{pr} = 10\,000$ samples/s, dynamics described with 10 bits, window length for FFT calculation $N = 256$ samples. For octave bands $P = 6$, for tertiary bands $P = 16$. Figures 6 and 7 present values of $F_{ter}(p)$ and $P_{oct}(p)$ parameters for the first four words in the vocabulary.

4. Recognition algorithms and similarity functions

4.1. Nearest neighbour NN algorithm

The entire teaching sequence is remembered in the learning process. A set of pairs: vector (describing the word) – vocabulary indication word number is called

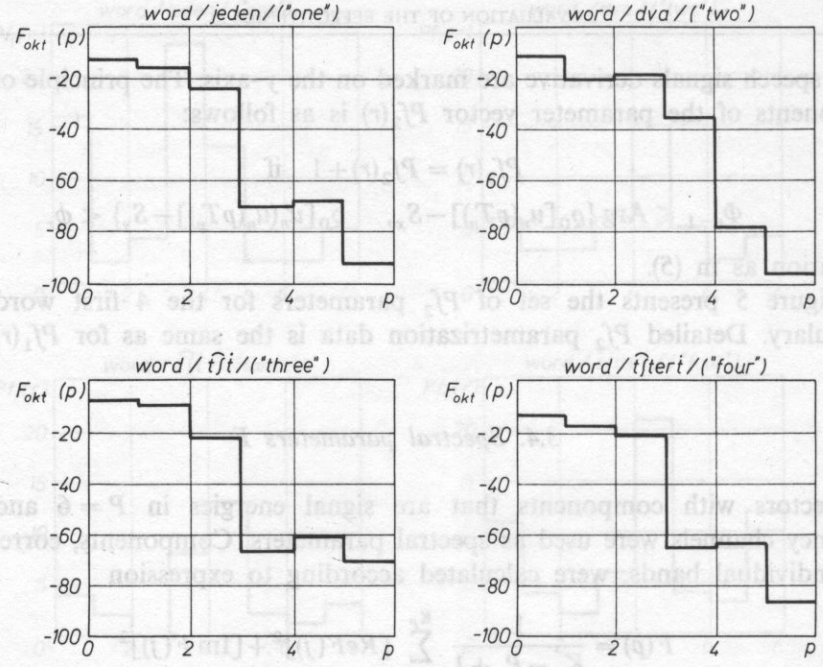

FIG. 6. Examples of parameter $F_{oct}(p)$ values, $P = 6$, for first four words in the vocabulary

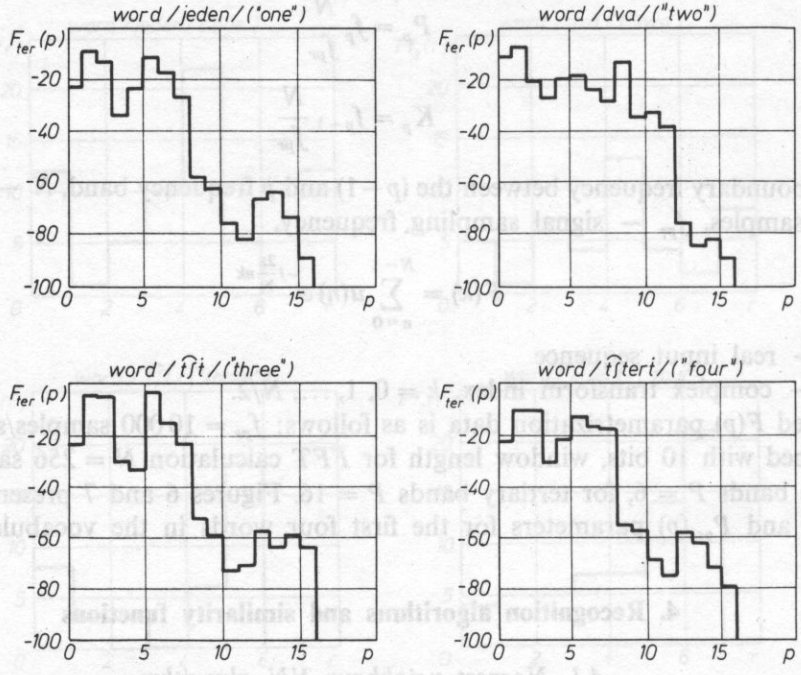

FIG. 7. Examples of parameter $F_{ter}(p)$ values, $P = 16$, for first four words in the vocabulary

the teaching sequence. The length of the sequence is a result of multiplication of the numbers M of words and number I_m of repetitions of the m–th word. In the process of actual recognition a sequence of unknown images (y) appears, described in the same parameter space as the teaching sequence (x). The NN procedure calculates a definite similarity function FP (e.g. distance) between all images in the teaching sequence and successive unknown image y:

$$FP(y, x_{m,i}) \tag{9}$$

while $i = 1, 2, \ldots, I_m$, $m = 1, 2, \ldots, M$ } parameters of teaching sequence

$x_{m,i}$ – i–repetition of m-class word of teaching sequence.

When all FP are calculated, the smallest one is found (in the case of distance). The class (word) number, which includes the image of the teaching sequence and which is found to be closest to the recognized image (in the sense of similarity function) is given as the classifier decision.

$$\Phi(y) = k \text{ i.e. } y \rightarrow k \tag{10}$$

if

$$FP(y, x_{k,i}) < FP(y, x_{l,i})$$

$$m = 1, 2, \ldots, M$$

$$l = 1, 2, \ldots, k-1, k+1, \ldots, M$$

In this case FP is a distance function. The algorithm is very simple. It assures good recognition results. The necessity of storing a great number of reference images is its only shortcoming.

4.2. Nearest mean NM algorithm

The necessity of storing all images of the teaching sequence is eliminated in the case of the NM algorithm. And this is an advantage of this algorithm with respect to the previous one. These images are replaced with the storage of mean images, as most typical for a given class. The decision rule has the following form:

$$\Phi(y) = m \text{ i.e. } y \rightarrow m$$

if

$$FP(y, W_m) < FP(y, W_l) \tag{11}$$

where

W_m – mean image of class word

$l = 1, 2, \ldots, m-1, m+1, \ldots, M$.

The *NM* algorithm is not recommended in these situations in which distributions of probability density of images are multimodal distributions.

4.3. Similarity functions

Descriptions and interpretations of similarity functions can be found in papers [1, 12, 18], among others. Thus, we will limit ourselves to the formulation of the mathematical notation of two similarity functions, i.e. Euclidean and Camberra. The first one was chosen, because of its wide application in ASR systems. The second one, because of its form which has a standarizing effect on parameters [1, 3, 18].

a) Euclidean distance

$$D^{\mathrm{EU}}(x, y) = \Big[\sum_{p=1}^{P} (x_p - y_p)^2\Big]^{1/2} \tag{12}$$

b) Camberra distance

$$D^{\mathrm{CAM}}(x, y) = \sum_{p=1}^{P} \frac{|x_p - y_p|}{|x_p + y_p|} \tag{13}$$

5. Experiments and results

The experiments were carried out in several stages, because of the complexity of accepted input data, i.e., various parametrization methods, classification algorithms and similarity functions. The structure of experiments is shown in Fig. 8. It contains four different combinations of the "source", i.e. vocabulary, and two types of test sequences (CT_w – individual test sequence, i.e. the same speaker recorded the teaching and test sequence and CT_0 what means that another speaker recorded the test sequence). 6 sets of parameters, two decision algorithms and two similarity functions were investigated. The number of possible combinations is tremendous. The needless variation was eliminated by accepting the following order of research:

stage I – the classification algorithm and similarity function were determined for the entire experimental material,

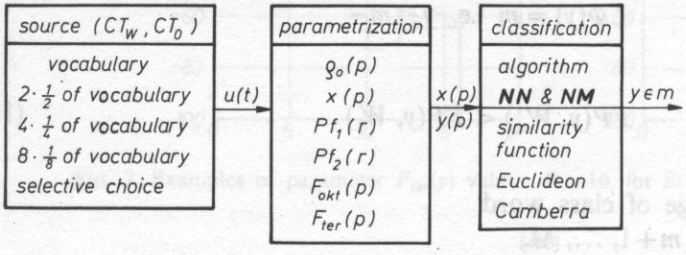

FIG. 8. Arrangement of experiments

stage II – the influence of the size of the vocabulary on the effectiveness of recognition was determined for all parameters,

stage III – evaluation of the system after eliminating from the vocabulary words, which in stage I were recognized worst of all. All these analysis and evaluations were performed for all 6 sets of parameters. At the same time their effectiveness and directions of changes due to investigated external factors were observed. It was accepted that the teaching sequence will consist of the first 3 repetitions of the first speaker, while the test sequence – of 2 succeeding repetitions from among the recordings of the first speaker and two repetitions of the second speaker.

5.1. Evaluation of algorithms and similarity functions

Within the first stage of experiments the whole vocabulary was recognized for both algorithms, both probability functions and all 6 sets of parameters. Cumulative results are presented in Table 1.

Table 1. Results of the first stage of recognition experiments [in %] for $M = 40$ words including all parameters of classifier and different sets of parameters

Parameters	Algorithm							
	NN				*NM*			
	Euclidean distance		Camberra distance		Euclidean distance		Camberra distance	
	CT_w	CT_0	CT_w	CT_0	CT_w	CT_0	CT_w	CT_0
$\varrho_0(p)$ $P = 40$	68.75	13.75	73.75	20.00	60.00	17.50	70.00	22.50
$x(p)$ $P = 8$	47.50	15.00	52.50	20.00	41.25	12.50	53.75	13.75
$Pf_1(r)$ $R = 8$	25.00	10.00	21.25	7.50	28.75	10.00	31.25	11.25
$Pf_2(r)$ $R = 8$	37.50	8.75	27.50	12.50	42.50	8.75	32.50	10.00
$F_{oct}(p)$ $P = 6$	16.25	1.25	23.75	6.25	21.25	6.25	35.00	10.00
$F_{ter}(p)$ $P = 16$	20.00	3.75	33.75	10.00	12.50	7.50	40.00	12.50

Results from Table 1 were considered from the point of view of further studies on the algorithm and similarity function. We can see that best results were achieved for most parameters for the Camberra distance. A small number of repetitions of the teaching and test sequence could not be decisive in the choice of algorithm. Yet, because of the fact that as a rule 3 repetitions of teaching sequence do not determine the reference as a mean exactly the *NN* algorithm was used in further research. The *NN* algorithm and the Camberra distance will be classified in further experiments. The results of experiments provided a rather clear explanation of the problem of using the system in the recognition of statements of another operator (CT_0). From results gathered in Table 1 we can see that for all cases of CT_0 effectiveness values are several times smaller than for CT_w. This means that global recognition parameters are strongly dependent on individual features of the voice. Therefore further analysis will refer to CT_w only.

5.2. *The effect of word set size on recognition results*

The second stage of recognition experiments was aimed at the determination of the relationship between recognition effectiveness and number of words in the vocabulary. Tables 2, 3 and 4 contain results for individual parameters. These tables indicate a monotone effectiveness increase accompanying the decrease of the number

Table 2. Recognition results [in %] for two parts of the vocabulary (2 × 20), separately, Algorithm *NN*, Camberra distance, test sequence CT_w

Parameters	Vocabulary words		Mean value
	1÷20	21÷40	
$\varrho_0(p)$ $P = 40$	82.50	80.00	81.25
$x(p)$ $P = 8$	65.00	65.00	65.00
$Pf_1(r)$ $R = 8$	22.50	42.50	32.50
$Pf_2(r)$ $R = 8$	52.50	22.50	37.50
$F_{oct}(p)$ $P = 6$	42.50	32.50	37.50
$F_{ter}(p)$ $P = 16$	47.50	35.00	41.25

Table 3. Recognition results [in %] for 4 parts of the vocabulary (4×10), separately. Algorithm *NN*, Camberra distance, test sequence CT_w

Parameters	Vocabulary words				Mean value
	1÷10	11÷20	21÷30	31÷40	
$\varrho_0(p)$ $P = 40$	85.00	90.00	100.00	75.00	87.50
$x(p)$ $P = 8$	55.00	90.00	90.00	70.00	76.25
$Pf_1(r)$ $R = 8$	45.00	40.00	50.00	55.00	47.50
$Pf_2(r)$ $R = 8$	55.00	75.00	40.00	45.00	53.75
$F_{okt}(p)$ $P = 6$	60.00	80.00	40.00	45.00	56.25
$F_{ter}(p)$ $P = 16$	65.00	80.00	40.00	60.00	61.25

Table 4. Recognition results [in %] for 8 parts of the vocabulary (8×5), separately. Algorithm NN, Camberra distance, test sequence CT_w

Parameters	Vocabulary words								Mean value
	1÷5	6÷10	11÷15	16÷20	21÷25	26÷30	31÷35	36÷40	
$\varrho_0(p)$ $P = 40$	90	100	100	90	100	100	80	80	92.50
$x(p)$ $P = 8$	100	70	100	100	90	90	80	80	88.75
$Pf_1(r)$ $R = 8$	70	60	50	60	90	80	60	60	66.25
$Pf_2(r)$ $R = 8$	80	50	60	90	80	60	70	70	70.00
$F_{oct}(P)$ $P = 6$	90	70	80	100	40	80	50	80	73.75
$F_{ter}(p)$ $P = 16$	90	70	90	90	40	70	60	80	73.75

of classes (words). The next observation results from Fig. 9.: when the number of classes is decreased then differences in effectiveness between individual parameters also decrease. A rather big differentiation of recognition effectiveness between individual groups of recognized words can be clearly noticed on the basis of all tables. These differences range from 25% for $\varrho_0(p)$ up to 20% for $Pf_1(r)$, for $M = 20$, and are equal to about 60% for a set of $M = 5$ words and parameter $F_{ter}(p)$.

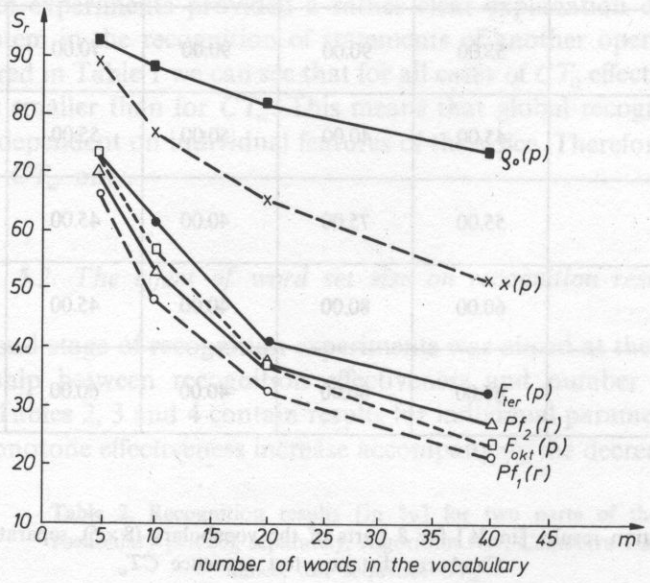

FIG. 9. Recognition effectiveness s_r [in %] in terms of the number of words in the vocabulary [m]

5.3. *Selective choice of vocabulary*

Differences in recognition results stated for various parts of the vocabulary clearly indicate that the choice of vocabulary influences the results of recognition. Hence, at this stage of research ten and twenty words which obtained worst recognition results for most parameters were eliminated from the vocabulary. The goal of the experiment was to determine quantitative effectiveness changes relative to control groups of words, i.e. average from successive sets of 30 and 20 words. Research results are presented in Table 5. A significant increase of correct recognition is visible, especially for the 20-word vocabulary. For example: for $\varrho_0(p)$ the effectiveness increased by 18.75%, for $x(p)$ by 20.00% and for $F_{ter}(p)$ by 26.25%. Next trials consisted in the arbitrary choice of the "best" with respect to achieved results, words from the whole vocabulary. Previously the effectiveness criterion with respect to most parameters was used, while now in two following trials the effectiveness of $x(p)$ selection a and $Pf_2(p)$ selection were applied as the criterions.

Table 5. Recognition results [in %] when words with worst recognition results achieved in the stage with reference to average results for sets of words taken in a fixed order are eliminated

Parameters	30-word vocabulary		20-word vocabulary	
	succesive words	Selected words	succesive words	selected words
$\varrho_0(p)$ $P = 40$	79.17	86.67	81.25	100.00
$x(p)$ $P = 8$	59.16	66.67	65.00	85.00
$Pf_1(r)$ $R = 8$	30.00	31.67	32.50	47.50
$Pf_2(r)$ $R = 8$	32.50	38.33	37.50	55.00
$F_{oct}(p)$ $P = 6$	31.67	46.67	37.50	57.50
$F_{ter}(p)$ $P = 16$	40.00	46.67	41.25	67.50

Table 6. Recognition results [in %], when 10 words with best recognition results in the I stage with reference to average results to sets of words taken in a fixed order are chosen. Selection a) — according to criterion $x(p)$, selection b) — according to criterion $Pf_2(r)$

Parameters	10-word vocabulary		
	mean for 10 words	selection a)	selection b)
$\varrho_0(p)$ $P = 40$	87.50	90.00	90.00
$x(p)$ $P = 8$	76.25	100.00	90.00
$Pf_1(r)$ $R = 8$	47.50	65.00	60.00
$Pf_2(r)$ $R = 8$	53.75	85.00	80.00
$F_{oct}(p)$ $P = 6$	56.25	50.00	70.00
$F_{ter}(p)$ $P = 16$	61.25	70.00	80.00

Results are presented in Table 6. The advantages of selecting words according to a definite set of parameters are clearly visible. For $x(p)$, for example, the recognition correctness improved from 76.25% to 100%. Similarly, the effectiveness for $Pf_2(r)$ increased from 53.75% to 80%.

5.4. *Usability analysis of individual sets of parameters*

Six sets of parameters were selected initially: $p_0(p)$, $x(p)$, $Pf_1(p)$, $Pf_2(p)$, $F_{oct}(p)$, $F_{ter}(p)$. The first parameter required time normalization – the others did not. Spectral parameters were derived from *FFT* calculations. This was the most complex analysis from among all applied parametrization methods. From the point of view of potential implementation the simplicity and speed of the parametrization procedure is very important. Spectral analysis in time close to real time requires an adequate signal processor or a bank of analogue or digital filters. Thus, it does not satisfy the requirement of system simplicity. Other parameters are based on the method of zero-crossings analysis. The simplicity of parameter extraction at high effectiveness is their fundamental advantage [2, 7, 16].

Ranges of effectiveness changes versus number of words are presented in graphical form in Fig. 9. It is characteristic that if for $M = 40$ the spread between sets of parameters is considerable, s_r values become more concentrated when the size of the vocabulary is decreased. This means that it is possible to apply $\varrho_0(p)$, $x(p)$, $Pf_2(p)$, $F_{oct}(p)$, $F_{ter}(p)$ to small vocabularies. The $x(p)$ set would be most preferred, because of its high effectiveness, small dimension of the parameter vector and calculation simplicity. $\varrho_0(p)$ was found to be most stable for all vocabularies, what can be explained by the time structure for this set of parameters. In spite of normalization significant differences in the current signal structure of words are retained. From a comparison of results for various equipotent subsets of vocabularies we can see (Table 2) that the smallest diversification is achieved from parameters $\varrho_0(p)$ and $x(p)$. The $Pf_1(r)$ set turned out to be positively the worst parameter. In its construction it took advantage of the same $\varrho_0(p)$ information twice on both axii of coordinates and acquiring a simple form it "lost" its force of discriminating parameters of zero-crossings density. The effectiveness of $Pf_2(r)$ parameters is somewhat better. It may be explained by the fact that the second axis of ordinates was calculated on the basis of the signals derivative. A division into two groups of parameters can be clearly noticed. $\varrho_0(p)$ and $x(p)$ belong to the first group, while the rest to the second one. The application of the traditional spectral analysis of octave bands ($P = 6$), as well as tertiary bands ($P = 16$), gave worse than expected results. Performed trials of reducing the parameter space resulted in the worsening of the results, while trials of vectors' concatenation resulted in a slight effectiveness improvement. For example, the concatenation of $\varrho_0(p)$ and $x(p)$ for $M = 40$ resulted in an effectiveness increase from 73.15% for $\varrho_0(p)$ to 76.25%. Similar results for voice identification were obtained by MAJEWSKI [13]. It results from Table 4 that most parameters could be

used as a parametrization method, especially for an adequately selected small vocabulary ($M \leqslant 5$). For slightly larger sets ($M \leqslant 10$), only $\varrho_0(p)$ and $x(p)$ could assure satisfactory discrimination force.

It is difficult to determine the typical effectiveness threshold of parameters, because it depends on the actual operating conditions of the ASR system. It seems that if it would be possible to provide a certain protection, e.g. two repetitions of the same word or a back confirmation, then ASR systems could fulfill a fundamental role in man-machine communication with effective recognition of isolated words beginning from 70%–80% [1, 11]. In a contrary case the required level of correct recognition for simple systems should exceed 90% [12]. It should be noted here that recognition does not exceed 95% in a case of listeners listening to so-called word lists, even when transmission and surrounding conditions are very favourable.

Many of the recognition systems of limited word vocabularies, described in literature, can collaborate with many operators speakers because of the segmental approach used to describe words. Thus, the simplicity of the global approach is connected with its smaller versatility.

6. Summary and conclusions

Most detailed analysis and comments can be found in paragraph 5.3. Thus, here we will only formulate general thesis and conclusions.

1. All three blocks of the system, i.e. "source", "measurement" and "classifier" are important when creating simple ASR systems.

2. Research has proved that in the "source" block the correct selection of the vocabulary is important – not only with respect to size, but also with respect to adequate phonetical diversification. This way the recognition effectiveness can be improved by over ten per cent.

3. The proper selection of parameters in the "measurement" block is traditionally the most important criterion affecting the systems effectiveness. The parameters of the time structure, i.e. $\varrho_0(p)$ and $x(p)$, satisfy the effectiveness and measurement simplicity criteria at the same time, best of all investigated sets of parameters.

4. A simple decision principle and adequate similarity function should be established in the classification block. Performed research and other authors papers [3, 18] indicate that the *NN* algorithm is the best classifier in simple systems. When parameters are normalized, it is convenient to use the Euclidean or Hamming distance function, while for not normalized parameters – the Camberra distance function is best.

5. In the global discription of words with single vectors in an observation space with dimension p, small vocabularies under 10 freely selected words (Table 3) or vocabularies containing 20 words adequately chosen on the basis of phonetical selection Table 5 have a chance of being implementated. The recognition effectiveness or the size of the vocabulary may probably be increased by introducing the global frequency-time description, for example [11, 18].

References

[1] Cz. BASZTURA, *Acoustic sources, signals and images* (in Polish) WKiŁ 1988.

[2] Cz. BASZTURA, *Automatic speaker recognition by zero-crossing analysis of the speech signal*, in: Speech Analysis and Synthesis, [Ed.] W. Jassem, 5, PWN, Warszawa 1980, pp. 5–40.

[3] Cz. BASZTURA, J. ZUK, *Similarity functions of images in non-parametric voice identification algorithms*, Archives of Acoustics, **16**, 2, (1991).

[4] Cz. BASZTURA, T. SAWCZYN, *Automatic segmentatation of a speech signal with the application of selected parametrization methods*, submitted for publication in Archives of Acoustics.

[5] M. K. BROWN, L. R. RABINER, *On the use of energy in LPC-based recognition of isolated words*, Bells System Techn. Journal, **61**, 10, 2971–2987 (1982).

[6] K. H. DAVIS, R. BIDDULPH, S. BALASHEK, *Automatic recognition of spoken digits*, J.A.S.A., **24**, 6, 637–642 (1952).

[7] W. DAXTER, E. ZWICKER, *On-line isolated word recognition using a microprocessor system*, Speech Communication **1**, 1, 21–27 (1982).

[8] R. GUBRYNOWICZ, *Application of zero-crossing method in speech signal analysis and automatic word recognition* (in Polish) Prace IPPT PAN 31, (1974).

[9] S. HUANG, R. GRAY, *Spellmode recognition based on vector quantization*, Speech Communication, **7**, 1, 41–53 (1988).

[10] A. ICHIKAWA, Y. NAKAO, K. NAKATA, *Evaluation of various parameter sets in spoken digits recognition*, IEEE Transactions on Audio and Electroacustics, AU–21, 3, 202–209 (1973).

[11] H. KUBZDELA et al., *Speech visualization* (in Polish) Ed. W. Jassem Prace IPPT PAN Warszawa 1987.

[12] W. A. LEA, *Trends in speech recognition*, Prentice-Hall Inc., Englewood Cliffs 1980.

[13] W. MAJEWSKI, *Speaker identification by means of averaged spectra of key words* (in Polish) Proc. XXXVI Open Seminar on Acoustics, Szczyrk-Biła September 1989, pp. 85–90.

[14] L. R. RABINER, J. G. WILPON, *Speaker-indpendent isolated word recognition for a moderate size 54 word vocabulary*, IEEE Trans. Acoust. Speech and Signal Processing, ASSP-27, 6, (1979) 583–587.

[15] L. R. RABINER, S. E. LEVINSON, M. M. SONDHI, *On the application of vector quantization and hidden Markov models to speaker-independent isolated word recognition*, Bell System Techn. Journal **62**, 4, 1075–1105 (1983).

[16] M. R. SAMBUR, L. R. RABINER, *A speaker-independent digit recognition system*, Bell System Techn. Journal, 54, 1, 81–102 (1975).

[17] R. W. SCHAFER, L. R. RABINER, *Digital representations of speech signals*, Proc. the IEEE, **64**, 4, 662–677 (1975).

[18] R. TADEUSIEWICZ, Speech signal (in Polish), WKiŁ Warszawa 1989.

Received November 28, 1989