INFORMATICA, 1995, Vol. 6, No. 2, 167–180

SPEAKER IDENTIFICATION USING VECTOR QUANTIZATION

Antanas LIPEIKA and Joana LIPEIKIENĖ

Institute of Mathematics and Informatics
2600 Vilnius, Akademijos St. 4, Lithuania

Abstract. The use of vector quantization for speaker identification is investigated. This method differs from the known methods in that the number of centroids is not doubled but increases by 1 at every step. This enables us to obtain identification results at any number of centroids. This method is compared experimentally with the method (Lipeika and Lipeikienė, 1993a, 1993b), where feature vectors of investigative and comparative speakers are compared directly.

Key words: speaker identification, likelihood ratio distance, vector quantization.

1. Introduction. In speaker identification we come upon some problems that must be solved. First of all it is a selection of features, a choice of a distance measure for comparing two feature vectors and a decission rule.

In Lipeika and Lipeikienė (1993a, 1993b) the speaker identification method is presented, where feature vectors are extracted from pseudostationary intervals of voiced sounds. Coefficients of the linear prediction model (LPC) and cepstral coefficients computed from the LPC coefficients are used as feature parameters. The LPC coefficients being taken as feature vectors, the likelihood ratio distance was used for comparing features, whereas the cepstral coefficients being as feature vectors, the Euclidean distance was used.

For comparison of two speakers the average distance

$$D_{xy} = \frac{1}{N_x} \sum_{j \in X} \min_{i \in Y} d_{ji}(X, Y) + \frac{1}{N_y} \sum_{j \in Y} \min_{i \in X} d_{ji}(X, Y). \qquad (1)$$

was calculated. Since symmetric $d_{ji}(X, Y)$ were used, D_{xy} is symmetric, too.

We can imagine that the feature vectors of a speaker X in a multivariate feature space form one cluster, that consists of subclusters. Similarly the feature vectors of a speaker Y form another cluster, that is also noncompact, but consists

of subclusters. Using the distance formula (1) for comparison of two speakers, each element of cluster X is compared with each element of cluster Y.

But if we look upon the problem of speaker identification as the problem of cluster analysis, we can calculate the distance between two clusters by other ways. One of such ways is comparing not every element of cluster X with every element of cluster Y, but comparing the centres of subclusters X and Y. Speaker identification methods using vector quantization (VQ) are based on this principle.

According to Song *et al.* (1985), Burton (1987), Zinke (1993), Irvine and Owens (1993), Xu *et al.* (1989), Rosenberg and Soong (1986), Soong and Rosenberg (1988), Buck *et al.* (1985) good results are obtained using vector quantization for speaker identification. As a rule, in vector quantization a cluster is devided into $M = 2^n$ subclusters, i.e., starting from one cluster (calculation of the zero centroid) the number of subclusters is doubled at every step as long as it reaches the number M. It is not convenient because we obtain a very large number of subclusters, what is often not necessary and does not correspond to the real number of subclusters present in features to be clustered. Therefore we modified the clusterizing algorithm of Juang *et al.* (1982) so that the number of subclusters increases by 1, but is not doubled.

2. Description of the clustering algorithm. Let

$$R_j = \{r_j(0), r_j(1), \ldots, r_j(p), b_j^2\}, \quad j = 1, \ldots, K,$$

be the vector of K features, where $r_j(0), \ldots, r_j(p)$ are values of the autocorrelation function of the j-th pseudostationary interval of the voiced sound of a speaker to be clustered; b_j^2 is a gain of the LPC model, calculated from the autocorrelation function $r_j(0), \ldots, r_j(p)$; p is the order of the LPC model.

Calculation of the zero centroid. We may calculate a "gravity centre" or the so called zero centroid of a cluster, that consists of feature vectors R_j. We update the zero centroid calculating the statistics

$$r^{(0)}(l) = \frac{1}{K} \sum_{j=1}^{K} r_j(l)/b_j^2, \quad l = 0, 1, \ldots, p \tag{2}$$

and estimating the LPC model parameters $A_0 = (a_1^{(0)}, \ldots, a_p^{(0)})$ from it. When estimating LPC model parameters according to the Durbin method (Rabiner and

Schafer, 1978) we obtain in paralell the reflection coefficients $k_1^{(0)}, \ldots, k_p^{(0)}$ that correspond to the zero centroid.

Determination of the average distortion while describing features by one reference pattern. When solving this problem we answer the question what an average error we are making if we describe all features by one reference pattern

$$D(A_0) = \frac{1}{K} \sum_{j=1}^{K} d(R_j, A_0), \tag{3}$$

where $d(R_j, A_0)$ is the likelihood ratio distance between the feature vectors R_j and the centroid A_0.

The likelihood ratio distance (Gray *et al.*, 1980) is calculated according to the formula

$$d(R_j, A_0) = \frac{1}{b_j^2} \sum_{k=0}^{p} r_j(k) r_k(A_0) - 1, \tag{4}$$

where $r_k(A_0)$, $k = 0, 1, \ldots, p$ are autocorrelation coefficients of LPC parameters of the zero centroid, which are calculated using the following relationships:

$$r_0(A_0) = \sum_{k=0}^{p} \left(a_k^{(0)}\right)^2, \quad a_0 = 1; \tag{5}$$

$$r_l(A_0) = 2 \sum_{k=0}^{p-1} a_{k+l}^{(0)} a_k^{(0)}, \quad l = 1, 2, \ldots, p. \tag{6}$$

The likelihood ratio distance has a spectral interpretation:

$$d(R_j, A_0) = \frac{1}{2\pi} \int_0^{2\pi} \frac{\widetilde{S}(\omega)/b_j^2}{S(\omega)/(b^{(0)})^2} d\omega - 1, \tag{7}$$

where $\widetilde{S}(\omega)$ is the spectral density, calculated from the LPC model, corresponding to the autocorrelation function $r_j(0), \ldots, r_j(p)$; $S(\omega)$ is the spectral density, calculated from the LPC parameters A_0 of the zero centroid.

If the LPC model parameters are $(a_0, a_1, \ldots, a_p, b)$, then the spectral density is calculated in a such way (Box and Jenkins, 1970):

$$S(\omega) = \frac{2b^2}{|1 + a_1 e^{-j\omega} + a_2 e^{-j2\omega} + \ldots + a_p e^{-jp\omega}|^2}, \quad 0 \leqslant \omega \leqslant \pi. \tag{8}$$

If the average distortion $D(A_0)$ exceeds the given threshold δ, then we must form two centroids from the zero centroid, which would represent the feature vectors R_j, $j = 1, \ldots, k$, more exactly to make the average distortion less.

Formation of two new centroids. Formation of new centroids is an iterative procedure. The initial point of this process is the reflection coefficients, corresponding to the zero centroid. We "distort" the reflection coefficients, multiplying them by multipliers 0.99 and 1.01, respectively. Thus, from the zero centroid we get two new initial centroids, whose coordinates determine two collections of the reflection coefficients $(k_1^{(1)}, \ldots, k_p^{(1)})$ and $(k_1^{(2)}, \ldots, k_p^{(2)})$. From the latter, using the recurrent relation (Rabiner and Schafer, 1978), one may calculate LPC model parameters, corresponding to these initial centroids. The LPC model parameters are calculated in such a way:

$$a_i^{(i)}(j) = k_i^{(j)}, \tag{9}$$

$$a_l^{(i)}(j) = a_l^{(i-1)}(j) - k_i^{(j)} a_{i-1}^{(i-1)}(j), \quad l = 1, \ldots, i-1. \tag{10}$$

When solving (9) ir (10) for $i = 1, \ldots, p$; $j = 1, 2$, we obtain that

$$a_l^{(j)} = a_l^{(p)}(j), \quad l = 1, \ldots, p; \; j = 1, 2.$$

The coordinates of these two centroids, expressed by the LPC model coefficients $(a_1^{(j)}, \ldots, a_p^{(j)})$, $j = 1, 2$, are used to determine the distance of each feature vector R_j, $j = 1, \ldots, K$ from these centroids (4)–(6). Further, using the nearest neighbour rule, on the basis of calculated distances we classify the features R_j, $j = 1, \ldots, K$. Every feature is attached to a centroid which is closer to this feature.

According to (3), the average distortion is assesed, which caused by the description of R_j, $j = 1, \ldots, K$ by two reference patterns, corresponding to the two initial centroids. For that we rewrite (3) in the following way:

$$D\left(A^{(1)}, A^{(2)}\right) = \frac{1}{K} \sum_{j=1}^{K} d^*\left(R_j, A^{(l)}\right), \tag{11}$$

where for each j smaller of two values of $d(R_j, A^{(1)})$, $d(R_j, A^{(2)})$ is taken as $d^*(R_j, A^{(l)})$.

As a result of classification by the nearest neighbour rule we obtain that features R_j, $j = 1, \ldots, K$, are devided into two initial clusters. As we have

done in the case of the zero centroid (2), we find centres of gravity of these clusters or the so called two improved initial centroids and their representation by LPC parameters. Further, according to the same formulas, we again calculate the distances of features R_j, $j = 1, \ldots, K$, from the improved initial centroids and classify the features according to the nearest neighbour rule. On the basis of classification results, the average distortion is calculated according to (11), which is due to the replacement of two reference patterns, describing the features R_j, $j = 1, \ldots, K$ by the LPC parameters, corresponding to the two initial centroids. If the average distortion decreases more than the given threshold ε, a further specification of centroid position is continued. If it increases less than ε, the iterative procedure is terminated. At the same time the procedure of LPC parameter estimation is stopped too. If the average distortion is less than the given threshold δ, the clustering process stops too. If it is more than δ, the cluster, which caused the largest average distortion, is divided into two clusters and the clustering process continues. It terminates only when the average distortion is less than the given quantity δ or when the number of centroids coincides with the largest given number of centroids. All the calculations are carried out according to the same formulas as in the case of two centroids.

Logical diagram of the clustering algorithm. The process of the clustering algorithm may be divided into such stages:

1) determine the zero centroid;

2) determine a distance of each feature vector from the zero centroid;

3) determine the average distortion $D(R_j, A_0)$ caused when describing the features R_j, $j = 1, \ldots, K$ by one reference pattern;

4) if the average distortion exceeds the given threshold δ, then two new centroids are formed from the zero centroid; otherwise the clustering process terminates;

5) $D = D(R_j, A_0)$; for each centroid determine a distance from the features R_j, $j = 1, \ldots, K$; classify the features by the nearest neighbour rule; calculate the average distortion; find the centroid with the largest average distortion;

6) determine the decrease of average distortion, caused describing the features by these centroids;

7) on the basis of classification improve the position of centroids;

8) repeat 5), 6), 7) as long as the decrease of the average distortion exceeds

the given threshold ε; if it becomes less than ε, go over to 9);

9) if the average distortion is less than the given in advance quantity δ, the clustering process terminates. Otherwise, go over to 10);

10) for the centroid with the largest average distortion calculate the reflection coefficients; with the aid of multipliers 0.99 and 1.01 obtain two sets of the reflection coefficients; from these coefficients calculate LPC model parameters, corresponding to the initial position of the two new centroids; afterwards, go over to 5).

3. Experiments. After calculating the centres of clusters (centroids) by the described clustering algorithm, the identification was carried out comparing codebooks of investigatives and comparatives by algorithm (1) which was used in the previous identification method.

The "reliability reserve" (Lipeika and Lipeikienė, 1993a, 1993b) was used as a criterion of identification quality. It enables us to compare different identification methods when there are no identification errors or their number is the same. The greater the reliability reserve, the better the identification method.

For experiments we have designed the software in C language. Speech signals used in the experiments were digitized by a 12 bits A/C converter at the rate of 10 KHz. Feature vectors were calculated from the pseudostationary intervals of voiced sounds (Lipeika and Lipeikienė, 1993a, 1993b) under such conditions:

- frame length 25 ms;
- frame step 5 cm;
- threshold for detecting of pseudostationary segments 0.07;
- threshold for a pseudostationary segment length – 25 ms;
- order of the LPC model 10.

A. Identification by a keyword. Phonograms of five speakers were recorded. Each speaker in two sessions repeated a Lithuanian word "namas" in two sessions for 10 times. The aim of the experiment was to investigate how the reliability reserve depends on a number of reference patterns when the identification is carried out by vector quantization.

The experimental results are given in Table 1. The reliability reserve as a function of a number of reference patterns is presented in Fig. 1. Negative values of the reliability reserve are restricted to -0.01.

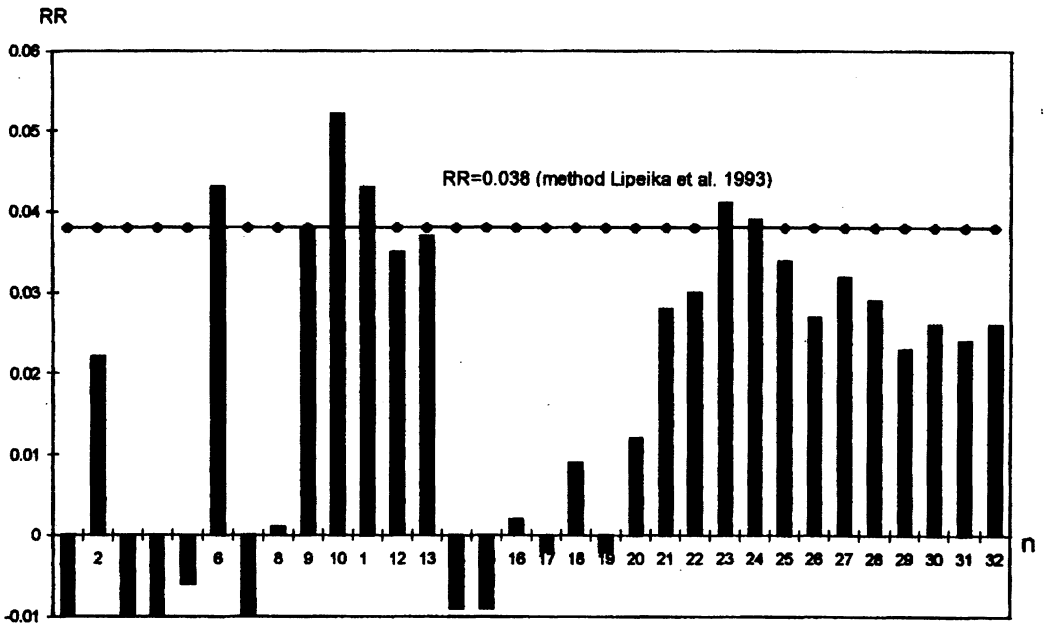

Fig. 1. Reliability reserve (RR) as function of number of reference patterns

Table 1. Dependence of reliability reserve on the number of reference patterns when identifying by the vector quantization method

Number of ref. patt.	Reliabil. reserve	Number of iden. errors	Number of ref. patt.	Reliabil. reserve	Number of iden. errors hfil
1	−0.588	2	17	−0.002	1
2	0.022	0	18	0.009	0
3	−0.072	2	19	−0.002	1
4	−0.087	1	20	0.012	0
5	−0.006	1	21	0.028	0
6	0.043	0	22	0.030	0
7	−0.041	1	23	0.041	0
8	0.001	0	24	0.039	0
9	0.038	0	25	0.034	0
10	0.052	0	26	0.027	0
11	0.043	0	27	0.032	0
12	0.035	0	28	0.029	0
13	0.037	0	29	0.023	0
14	−0.009	1	30	0.026	0
15	−0.009	1	31	0.024	0
16	0.002	0	32	0.026	0

The results show that when the number of reference patterns is less than 10, the reliability reserve varies in a wide range, but identification errors are more frequent. When the number of reference patterns varies from 9 to 13, there are no identification errors at all. Further there is an interval of unstable decisions again and the reliability reserve fluctuates about zero. Only when the number of reference patterns is over 20, the identification results become stable and there is quite a large reliability reserve.

If we compare these results with the identification results of the method (Lipeika and Lipeikienė, 1993a, 1993b) where the reliability reserve is 0.038, we see that only when the number of reference patterns is 6, 9, 10, 11, 23, 24, the reliability reserve exceeds 0.038. Since the reliability reserve becomes

Table 2. Dependence of reliability reserve on the number of reference patterns when identifying by the vector quantization method (text independent)

Number of ref. patt.	Reliabil. reserve	Number of iden. errors	Number of ref. patt.	Reliabil. reserve	Number of iden. errors hfil
1	−0.585	10	17	0.048	0
2	−0.462	10	18	0.035	0
3	−0.143	1	19	0.019	0
4	−0.029	1	20	0.047	0
5	−0.008	1	21	0.060	0
6	−0.071	1	22	0.040	0
7	−0.009	1	23	0.005	0
8	−0.050	1	24	0.029	0
9	0.019	0	25	0.038	0
10	−0.076	1	26	0.034	0
11	−0.010	1	27	0.015	0
12	0.006	0	28	0.041	0
13	0.012	0	29	0.036	0
14	0.005	0	30	0.048	0
15	0.059	0	31	0.052	0
16	0.055	0	32	0.044	0

stable only if the number of reference patterns is large and is less than of method (Lipeika and Lipeikienė, 1993a, 1993b), we make a conclusion that using of vector quantization for speaker identification by a keyword we don't achieve better results.

B. Text independent identification. Analogously phonograms of five women were recorded and studied. The difference was only that text independent phonograms were recorded during two sessions. The identification results are presented in Table 2. The reliability reserve as a function of a number of reference patterns is presented in Fig. 2. Negative values of the reliability reserve are restricted to −0.01.

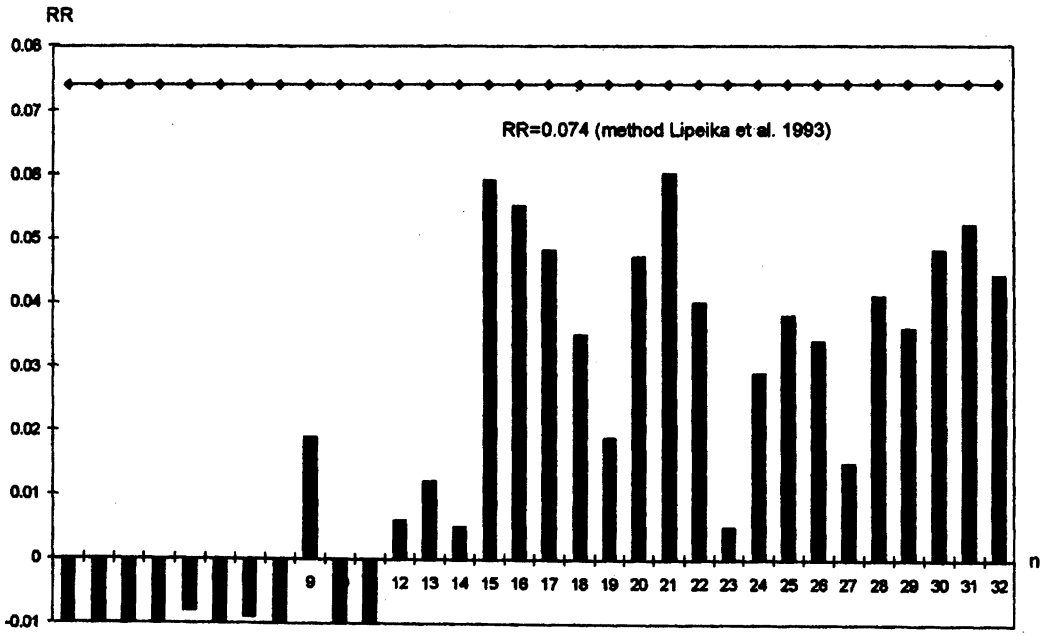

Fig. 2. Reliability reserve (RR) as function of number of reference patterns

Table 3. Dependence of reliability reserve on the number of reference patterns when identifying by the vector quantization method (text dependent)

Number of ref. patt.	Reliabil. reserve	Number of iden. errors	Number of ref. patt.	Reliabil. reserve	Number of iden. errors hfil
1	−2.555	7	17	−0.005	1
2	−0.191	5	18	−0.017	1
3	−0.063	3	19	−0.023	1
4	−0.221	4	20	−0.012	1
5	−0.047	5	21	−0.026	1
6	−0.072	3	22	−0.024	1
7	−0.150	3	23	−0.003	1
8	−0.073	1	24	0.000	0
9	−0.013	2	25	0.017	0
10	−0.004	1	26	0.017	0
11	−0.015	2	27	0.010	0
12	−0.001	1	28	0.011	0
13	−0.007	1	29	0.014	0
14	−0.025	1	30	0.011	0
15	−0.006	1	31	0.014	0
16	−0.006	1	32	0.019	0

As we see, the results are similar as in the case of identifying by a keyword. When the number of reference patterns is small, we obtain identification errors. When this number is over 11, there are no identification errors at all. But the reliability reserve is less than that of method (Lipeika and Lipeikienė, 1993a, 1993b), where the reliability reserve is 0.074. Note that when the number of reference patterns is 1,2,3, we obtain 10 identification errors. At any number of reference patterns the reliability reserve does not exceed 0.074 (method Lipeika and Lipeikienė, 1993a, 1993b).

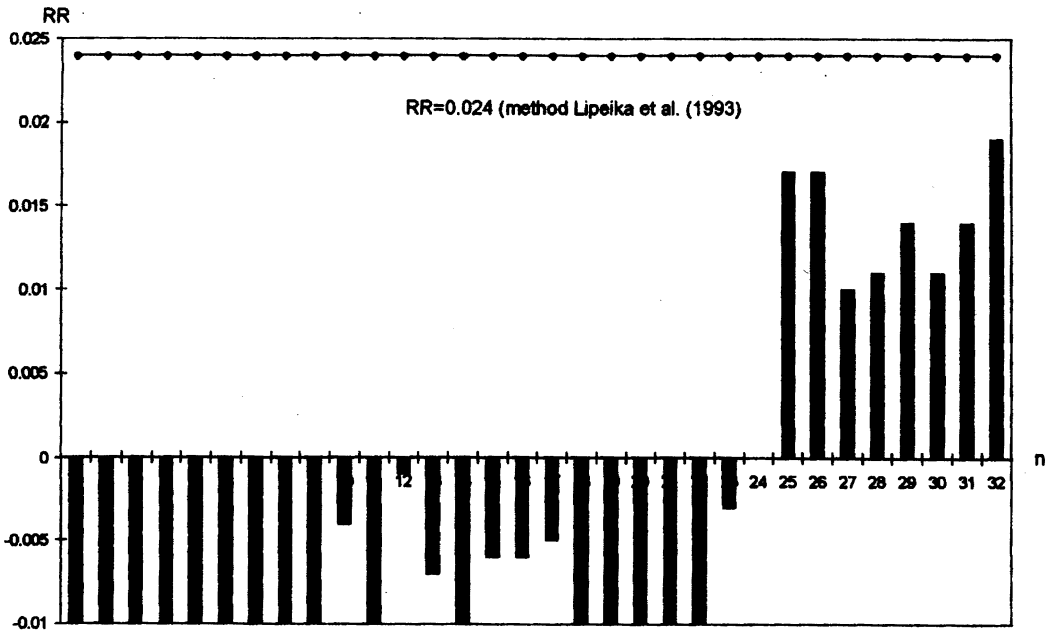

Fig. 3. Reliability reserve (RR) as function of number of reference patterns

C. Text dependent identification of telephone speech. Text dependent phonograms of four men and one woman have been recorded over a telephone. The identification results are presented in Table 3. The reliability reserve as a function of a number of reference patterns is presented in Fig. 3. Negative values of the reliability reserve are restricted to -0.01.

In this experiment errors are obtained all the time, as long as the number of reference patterns varies from 1 to 24. Starting from 25 reference patterns, there are no more identification errors, but the reliability reserve is less than that of method (Lipeika and Lipeikienė, 1993a, 1993b) where the reliability reserve is 0.024).

4. Conclusions. The speaker identification method has been developed based on feature vector quantization. This method differs from known methods in that the number of reference patterns is not doubled but increases by 1 at every step. This enables us to obtain identification results at any number of reference patterns not only at number $M = 2^n$, where n is an integer. The experimental investigation showed that by this method the identification results obtained were no better than by the method Lipeika and Lipeikienė (1993a, 1993b).

REFERENCES

Box, J., and G.Jenkins (1970). *Time Series Analysis, Forecasting and Control.* San Francisco, Cambridge, London, Amsterdam.

Buck, T., D.Burton, and J.Shore (1985). Text dependent speaker recognition using vector quantization. *Proc. of the ICASSP-85,* 391–394.

Burton, D. (1987). Text-dependent speaker verification using vector quantization source coding. *IEEE trans. on ASSP-35,* **2**, 133–143.

Gray, R., A.Buzo, A.Gray, and Y.Matsuyama (1980). Distortion measures for speech processing. *IEEE trans. on ASSP-28,* **4**, 367–376.

Irvine, D., and F.Owens (1993). A comparison of speaker recognition techniques for telephone speech. *Proc. of the EUROSPEECH'93,* 2275–2278.

Juang, B., D.Wong, and A.Gray (1982). Distortion performance of vector quantization for LPC voice coding. *IEEE trans. on ASSP-30,* **2**, 294–304.

Lipeika, A., and J.Lipeikienė (1993a). The use of pseudostationary segments for speaker identification. *Proc. of the EUROSPEECH'93,* 2303–2306.

Lipeika, A., and J.Lipeikienė (1993b). Speaker identification. *INFORMATICA,* **4**(1–2), 45–56.

Rabiner, L., and R.Schafer (1978). *Digital Processing of Speech Signals.* Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Rosenberg, A., and F.Soong (1986). Evaluation of a vector quantization talker recognition system in text independent and text dependent modes. *Proc. of the ICASSP-86,* 873–876.

Soong, F.K., and A.Rosenberg (1988). On the use of instantaneous and transitional spectral information in speaker recognition. *IEEE trans. on ASSP-36,* **6**, 871–879.

Soong, F.K., A.E.Rosenberg, L.R.Rabiner, and B.H.Juang (1985). A vector quantization approach to speaker recognition. *Proc. of the ICASSP-85,* 387–390.

Xu, L., J.Oglesby and J.Mason (1989). The optimization of perceptually based features for speaker identification. *Proc. of the ICASSP-89,* 520–523.

Zinke, J. (1993). Influence of pattern compression on speaker verification. *Proc. of the EUROSPEECH'93,* 2267–2270.

Received January 1995

A. Lipeika is a Doctor of Technical Sciences, a senior researcher of the Recognition Processes Department at the Institute of Mathematics and Informatics. Scientific interests include: processing and recognition of random processes, detection of changes in the properties of random processes, digital signal processing, speaker identification.

J. Lipeikienė is a Doctor of Technical Sciences, a senior researcher of the Recognition Processes Department at the Institute of Mathematics and Informatics. Scientific interests include: processing of random signals, robust methods for determination of change-points in the properties of random processes, data compression.

KALBANČIOJO IDENTIFIKAVIMAS NAUDOJANT VEKTORINIO KVANTAVIMO METODĄ

Antanas LIPEIKA, Joana LIPEIKIENĖ

Darbe nagrinėjamas vektorinio kvantavimo metodo taikymas kalbančiojo identifikavimui. Šis metodas skiriasi nuo žinomų metodų tuo, kad kiekviename vektorinio kvantavimo žingsnyje centroidų skaičius ne dvigubinamas, o didinamas vienetu. Taip gaunami rezultatai, esant bet kokiam centroidų skaičiui.

Šis metodas yra eksperimentiškai sulygintas su metodu [1, 2], kur tiriamojo ir lyginamojo kalbančiųjų požymių vektoriai yra sulyginami betarpiškai.