ISSN 1392-8619 print/ISSN 1822-3613 online
ŪKIO TECHNOLOGINIS IR EKONOMINIS VYSTYMAS
TECHNOLOGICAL AND ECONOMIC DEVELOPMENT OF ECONOMY
http://www.tede.vgtu.lt
2006, Vol XII, No 1, 50–55

INVESTIGATION OF THE SPEAKER IDENTIFICATION METHOD BASED ON CLUSTERED PSEUDOSTATIONARY SEGMENTS OF VOICED SOUNDS

Povilas Treigys[1], **Antanas Lipeika**[2]

Institute of Mathematics and Informatics,
Akademijos g. 4, LT-08663 Vilnius, Lithuania
E-mail: [1]treigys@ktl.mii.lt, [2]lipeika@ktl.mii.lt

Received 3 October 2005; accepted 4 January 2006

Abstract. The problem of speaker identification is investigated. Basic segments - pseudo stationary intervals of voiced sounds are used for identification. The identification is carried out, comparing average distances between an investigative and comparatives. The coefficients of the linear prediction model (LPC) of a vocal tract are used as features of identification. Such a problem arises in stenographic practice where it is important for speech identification to know who is speaking. Identification should be used in stenography and it has to be fast enough in order not to disturb the stenographer's job. The clustered parameter data will be investigated by providing the performance of the speaker identification method with respect to the computational time and the number of errors.

Keywords: text independent speaker identification, data clustering, vector quantization, time consumption, number of errors, pseudostationary segments of voiced sounds.

1. Introduction

The automatic speaker identification problem is very important to the stenographer. A speaker won't help to relate the recorded speech with a person, who is really speaking. It is clear that a voice phonogram is distorted during recording (influence of environment noise, imperfect recording equipment, etc.). Therefore this investigation field is being developed intensively.

Clustered data are often used in speech transmission systems to reduce the bandwidth of signals [1]. Instead of transmitting all the bits necessary to represent *k* dimensional vector, only the codebook entry number of the centroid closest to the vector would need to be transmitted [2]. Thus, a sequence of codebook entry numbers could be transmitted to represent the entire utterance. At the receiving end, an approximation of the original vector could be constructed by looking up the codebook entry of each number in the sequence. Also, vector quantization (VQ) can be used in many different ways for automatic speaker identification [3]. In some systems VQ is used only to compress data, in others the segregation of vectors is used as a preprocessing step.

The dependence of clustered data of a particular phonogram, calculation time and error rates in speaker identification are investigated in this paper. When pronouncing a voiced sound, a vocal tract is fixed for a short period; therefore there occurs a possibility to "measure" the parameters of a vocal tract and to identify a speaker, using phonograms.

When we want to identify a speaker, we make an assumption that we have a phonogram of an unknown speaker (investigative) and *n* known comparative speakers. Our purpose is to choose from these comparatives the person that is, in some sense, closest to the investigative and then to determine, whether this closest person and the comparative is the same person. So the problem of identification of the closest comparative is investigated in this paper, providing the dependence number of identification errors on clustered phonogram data, and the time consumption needed for the calculation process.

2. Detection of pseudo stationary segments

For the detection of pseudostationary segments [4] in speaker identification, a phonogram is divided into frames (segments) the length of which is *N* digital points of a speech signal and they are moved with respect to each other by *M* points (a step of a frame is *M*). The filtration of a speech signal y_t for all frames is performed according to formula $x_t = y_t - 0.94 y_{t-1}$. This filtration enables us to suppress irregular low frequency components [5]. Afterwards a resulting signal is processed using the Hamming window. The use of the Hamming window and low frequency filtration allows us to get a stable LPC model of a speech signal.

We evaluate the parameters of LPC model by the correlation method, using Durbin's algorithm. Afterwards autocorrelation coefficients of LPC parameters are calculated. Further, using the autocorrelation coefficients of linear prediction parameters of the previous frame and the correlation coefficients of a signal of the next frame, divided by the square LPC model gain coefficient, we calculate the likelihood ratio distance [6] for all neighboring pairs of frames. If a distance between two neighboring frames is less than a preassigned threshold (the threshold is chosen experimentally), we draw a conclusion that moving by a frame step does not change a spectral structure of a signal, i.e. it is pseudostationary. This condition is verified until the likelihood ratio distance exceeds the threshold. Then we consider that a stationary interval is terminated [4].

Since we are not interested in very short pseudo stationary intervals, we compare them with the threshold of the minimal pseudostationary segment and leave for further investigation only those pseudo stationary segments which are longer than this threshold. The likelihood ratio distance has spectral interpretation [7]:

$$d_{LR}(\tilde{S},S) = \int_{-\pi}^{\pi} \frac{\tilde{S}(\theta)/\tilde{b}^2}{S(\theta)/b^2}\frac{d\theta}{2\pi} - 1, \qquad (1)$$

where $S(\theta)$ and $\tilde{S}(\theta)$ are spectral densities of LPC model of the first and the second frame, respectively, $\tilde{b}^2$ and b^2 are square gain coefficients of those models.

Because of the great computation amount, it is not convenient to calculate the likelihood ratio distance in the frequency domain. It is usually calculated in the time domain by:

$$d_{LR}(\tilde{S},S) = \{\frac{r_x(0)}{\tilde{b}^2} r_a(0) + 2\sum_{i=1}^{p} \frac{r_x(i)}{\tilde{b}^2} r_a(i)\} - 1, \qquad (2)$$

where $r_x(i)$ is the autocorrelation function of a signal in the second frame, $r_a(i)$ are autocorrelations of the LPC model parameters of the first frame:

$$r_a(i) = \sum_{k=0}^{p-i} a_{k+i} a_k,\ i = 0,1,2,\ldots,p, \qquad (3)$$

and *p* is the order of LPC model.

3. Identification based on the calculation of the average distance between speakers

Let us have N_x pseudostationary intervals of an investigative speaker (*X*) and N_{A_i} pseudostationary intervals of comparatives (A_i). Let us calculate all the possible distances $d_{ji}(X, A_i)$ between the pseudostationary intervals of investigative *X* and comparatives A_i, *i*=1 ... *n*. Then the average distance between the data of features, describing the investigative *X*, and that describing comparatives A_i, may be calculated as follows [8]:

$$\begin{aligned} D_{X_i} &= \frac{1}{N_X}\sum_{j\in X} \min_{l\in A_i} d_{jl}(X, A_i), \\ D_{Ai} &= \frac{1}{N_{A_i}}\sum_{j\in A_i} \min_{l\in X} d_{jl}(X, A_i), \\ D_X &= D_{X_i} + D_{A_i}. \end{aligned} \qquad (4)$$

Here N_x and N_{A_i} are the numbers of frames in the phonograms of investigative *X* and comparatives A_i; $d_{jl}(X, A_i)$ is the distance between frames.

When detecting the pseudostationary segments we used the likelihood ratio distance. But this measure is not symmetric, i.e.,

$$d_{LR}(\tilde{S},S) \neq d_{LR}(S,\tilde{S}). \qquad (5)$$

It is not a shortcoming in the detection of pseudo stationary segments because a threshold is not high, meanwhile asymmetry appears when the values of distance are high. However, when calculating the average distance, it is desirable that the distance in formula (4) be symmetric. So we make the likelihood ratio distance symmetric [9]:

$$d(\tilde{S},S) = \frac{d_{LR}(\tilde{S},S) + d_{LR}(S,\tilde{S})}{2}. \qquad (6)$$

After calculating the average distance between investigative speaker *X* and all comparatives A_i, we find "the closest" comparative by comparing the average distances:

$$\hat{I} = \min_{1\le i\le n} D_{XA_i}. \qquad (7)$$

The cepstral distance is often used in speaker identification as well [10]. The cepstral coefficients can be calculated from LPC coefficients using formulas [11, 12]:

$$c_0 = \ln b^2,$$

$$c_1 = -a_1,$$

$$c_n = -\sum_{k=1}^{n-1}(1-k/n)a_k c_{n-k} - a_n, n = 2,\ldots,p, \quad (8)$$

$$c_n = -\sum_{k=1}^{n-1}(1-k/n)a_k c_{n-k},\ n = p+1,\ldots,L. \quad (9)$$

The cepstral distance between two frames of a speech signal with the respective coefficients $(a_1,\ldots, a_p, b)$, $(\tilde{a}_1,\ldots, \tilde{a}_p, \tilde{b})$ may be defined as follows:

$$d_{cep}(L) = [u(L)]^2 = (c_0 - \tilde{c}_0)^2 + 2\sum_{k=1}^{L}(c_k - \tilde{c}_k)^2. \quad (10)$$

It is important to know [6] that as L increases, *u(L)* approaches d_2 from below and

$$\lim_{L\to\infty} u(L) = d_2, \quad (11)$$

where d_2^2 has the following spectral interpretation:

$$d_2^2 = \int_{-\pi}^{\pi}\left|\ln\frac{\tilde{S}(\theta)}{S(\theta)}\right|^2 \frac{d\theta}{2\pi}. \quad (12)$$

It should be mentioned that this distance is symmetric and convenient to use for speaker identification.

Usually we desire that a distance would not depend on gain, so we assume that $b = \tilde{b} = 1$. Then (12) may be written as:

$$d_2^2 = \int_{-\pi}^{\pi}\left|\ln\frac{\tilde{S}(\theta)/\tilde{b}^2}{S(\theta)/b}\right| \frac{d\theta}{2\pi}. \quad (13)$$

By substituting the cepstral distance, described by (10), into (4), we can carry out speaker identification by cepstral distance.

4. Speaker identification using VQ

Here, the number of clusters (centroids) is not doubled at every step, but increased by 1. It better shows the dependence of identification quality on the codebook length (a number of clusters). In addition, we use the average distance (4), described in the previous method, for comparing the centers of clusters of an investigative and comparatives.

Further we present the description of the clustering process [13].

Let

$$R_j = \{r_j(0), r_j(1),\ldots, r_j(p),\ b_j^2\},\ j = 1,\ldots,K, \quad (14)$$

be K features vector extracted from pseudostationary speech intervals, where $r_j(0),\ldots, r_j(p)$ are values of the autocorrelation function of j-th pseudo stationary segment. b_j^2 is a square gain of LPC model.

4.1. Calculation of zero centroid

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

$$r^{(0)}(l) = \frac{1}{K}\sum_{j=1}^{K} r_j(l)/b_j^2, l = 0,1,\ldots,p, \quad (15)$$

and estimate the parameters of the linear prediction $A_0 = (a_1^{(0)},\ldots, a_p^{(0)})$ from it. When estimating LPC parameters by Durbin method, we obtain in parallel reflection coefficients $k_1^{(0)},\ldots, k_p^{(0)}$ that correspond to the zero centroid.

4.2. Determination of average distortions while describing features by one reference pattern

When solving this problem, we answer the question what average error we are making, if we describe all the features by one reference pattern

$$D(A_0) = \frac{1}{K}\sum_{j=1}^{K} d(R_j, A_0), \quad (16)$$

where $d(R_j, A_0)$ is the likelihood ratio distance between feature vectors R_j and centroid A_0. The likelihood ratio distance is calculated according to formula (2).

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

4.3. Formation of two new centroids

The 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 0.99 and 1.01, respectively. Thus, from the zero centroid we get two new initial centroids which 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 [14], we may calculate LPC model parameters, corresponding to these initial centroids. LPC model parameters are calculated in such a way:

$$\alpha_i^{(i)}(j) = k_i^{(j)}, \quad (17)$$

$$\alpha_l^{(i)}(j) = \alpha_l^{(i-1)}(j) - k_i^{(j)}\alpha_{i-l}(j), l = 1,\ldots,i-1. \quad (18)$$

When solving (17) and (18) for i=1,...,p, j=1,2, we obtain that

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

The coordinates of these two centroids expressed by LPC model coefficients $(a_1^{(j)},..., a_p^{(j)})$, $j=1, 2$, are used to determine the distance of each feature vector $R_j, j=1,..., K.$ from these centroids, using formula (8). Next, using the nearest neighbor rule, on the basis of the calculated distances we classify features $R_j, j=1,..., K.$ Every feature is attached to a centroid which is closer to this feature. According to (16), the average distortion is assessed, which is caused by the description of $R_j, j=1,..., K,$ by two reference patterns, corresponding to two initial centroids. For that we rewrite (16) in the following way:

$$D(A^{(1)}, A^{(2)}) = \frac{1}{K}\sum_{j=1}^{K} d^*(R_j, A^{(l)}), \qquad (19)$$

where $d^*(R_j, A^{(l)}) = \min\{d(R_j^{\,l}, A^{(1)}), d(R_j, A^{(2)})\}$.

As a result of classification by the nearest neighbor rule we obtain that features $R_j, j=1,..., K$ are divided into two initial clusters. As we have already done in the case of the zero centroid, according to (15) we find the centers of gravity of these clusters or the so – called improved initial centroids and their representation by the LPC parameters. Further, in new of the same formulas, we calculate the distances of features $R_j, j=1,..., K,$ from the improved initial centroids once more, and classify the features according to the nearest neighbor rule.

On the basis of classification results the average distortion is calculated according to (19) which is due to the replacement of two reference patterns, describing features $R_j, j=1,..., K,$ by LPC parameters, corresponding to two initial centroids. If the average distortion decreases more than given threshold ε, further specification of the centroid position is continued. If it decreases less than ε, the iterative process is terminated. At the same time the procedure of LPC parameter estimation is stopped as well. If the average distortion is more than given threshold 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 given quantity δ or when the number of centroids is the same as the largest given number of centroids. All calculations are carried out by the same formulas as in the case of two centroids.

5. Test results

Testing data are comprised of 76 different speakers' phonograms. Later on from these phonograms LPC parameters of order 10 are extracted. The identification process is performed by calculating the average distance between the comparative and investigative data of features. The computer used for calculations had 1200 MHz CPU, and 384Mb of RAM. First of all, the identification was accomplished by calculating the distances of the nonclustered data using asymmetric (4) distance calculation. The numbers of misidentification errors were 4 with the time consumption 6.01 min. To reduce the computation time all further computations (in Figs 1, 2) were performed using the asymmetric distance calculations (4). Also, the symmetric distance calculations (5) among clusters are shown in Figs 3, 4. Figs 5, 6 show the comparison results of the number of errors and the computational time consumption.

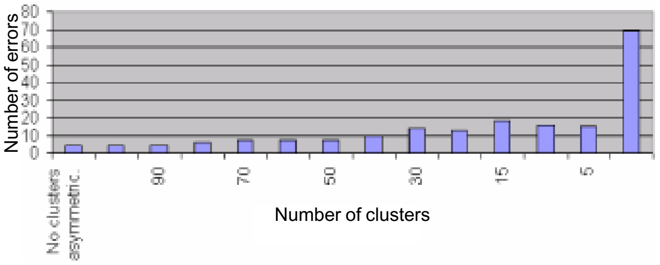

Fig 1. Misidentification dependence on the clustered data in asymmetric distance calculations

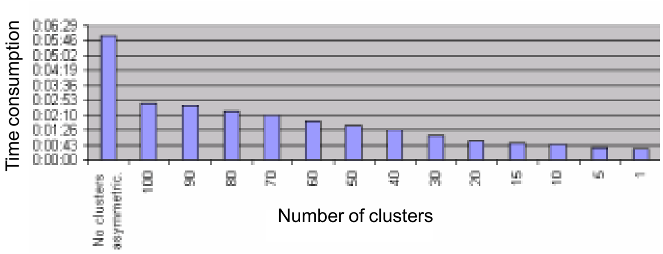

Fig 2. Dependence of identification time consumption on the clustered data in asymmetric distance calculations

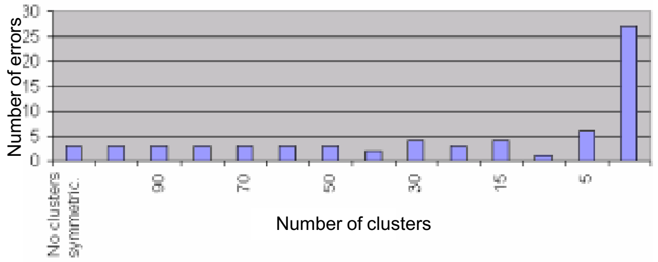

Fig 3. Misidentification dependence on the clustered data in symmetric distance calculations

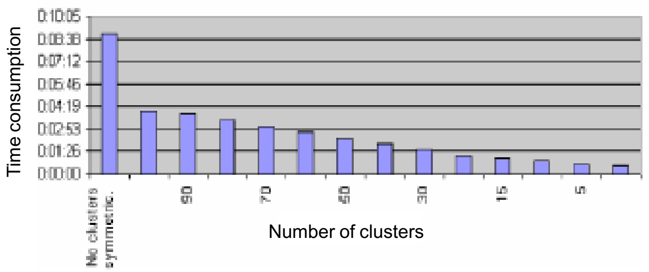

Fig 4. Dependence of identification time consumption on the clustered data in symmetric distance calculations

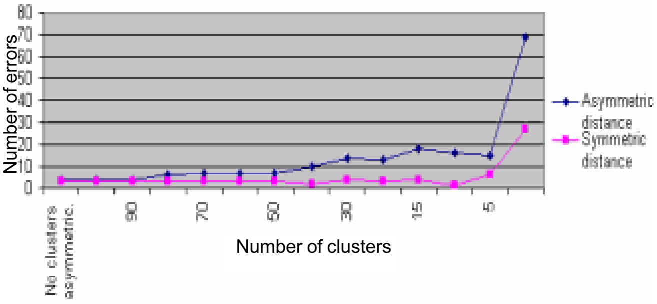

Fig 5. Compared number of errors in symmetric and asymmetric distance calculation methods

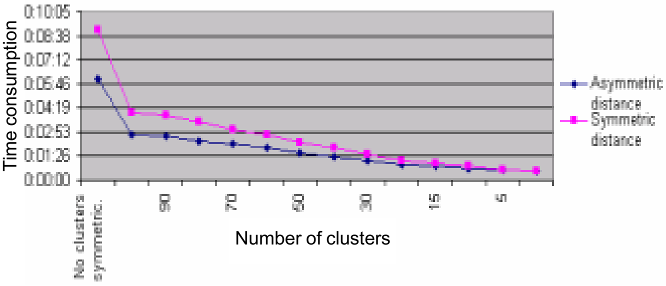

Fig 6. Time consumption compared in symmetric and asymmetric distance calculation methods

Fig 1 shows the identification number of errors depending on a different number of clusters. The minimal error for nonclustered data using asymmetric distance computations (4) is 4. Beginning with the vector data of 30 clusters the figure shows that the misidentification error of each next cluster is not growing gradually up, as we should expect. Finally, when computations are performed with a single vector which consists of only one feature, describing investigative, misidentification grows straight up to 69.

Fig 2 represents time consumption during the identification process on the clustered data. It shows that time consumption is falling down whenever the number of clusters decreases. Time consumption on the nonclustered data identification process is 6.01 min. and the identification of one cluster took – 32 seconds. A considerable time consumption jump (3.17min) is observed on the nonclustered data with asymmetric distance (4) as compared to the data of cluster size of 100 feature vectors (see Fig 2).

All further figures present the symmetric distance calculation (5) results.

This figure shows that symmetric distance calculation is more stable than asymmetric. The number of errors seen in the identification remains the same (3) up to the use of the cluster of 40 data of features. Then identification instability occurs.

Here, a jump of time consumption compared to asymmetric distance calculation is more significant. Time consumption needed for the nonclustered data and the cluster of 100 data of features decreases by 5min.

In Fig 5 speaker's identification instability is observed in both methods, starting from cluster size 50. Fig 6 shows that the time consumption needed for identification processes has a straight fall for the nonclustered data and cluster of size 100. Then time consumption decreases (as cluster size decreases) smoothly.

6. Conclusions

A text independent speaker identification algorithm is presented in this paper. The influence of clustered data on the identification performance using asymmetric and symmetric distance calculation methods is investigated. The influence was investigated in two ways: the computational time consumption and misidentification number, using a different number of clusters in the identification process. By comparing misidentification and time consumption jumps over the clustered data used in the identification process, the lowest number of error and time consumption is obtained using 60 and 90 clusters in symmetric and asymmetric distance calculation methods. The time needed for those calculation processes is 2.42min and 2.44min. pro rata. Using 60 and 90 clusters, the number of errors in the symmetric distance is stable and remains 3 while in asymmetric it is 4. The time consumption needed for computational processes decreases from 8.59min to 2.42min and 6.01 min to 2.36 min in both methods, accordingly, over all 76 investigative persons (preserving the lowest error number). Using symmetric distance calculation method between clusters, time consumption has decreased 4 times. The results also show that for the same misidentification (3), two times less data of features were used.

References

1. Kevans, R. L.; Rodman, R. D. Voice Recognition. Artech House, Boston, London, 1996.
2. Furui, S. Digital Speech Processing, Synthesis and Recognition. Marcel Dekker, Inc., 2001.
3. Lipeika, A.; Lipeikienė, J. Speaker identification using vector quantization. *Informatica*, Vol 6, No 2, 1995, p. 167–180.
4. Lipeika, A.; Lipeikiene, J. Speaker identification methods based on pseudo stationary segments of voiced sounds. *Informatica*, Vol 7, No 4, 1996, p. 469–484.
5. Rabiner, L.; Juang, B. H. Fundamentals of Speech Recognition. Prentice Hall, 1993.
6. Gray, A. H.; Markel., J. D. Distance Measures for Speech Processing. *IEEE Trans. on Acoustic Speech and Signal Processing*, Vol ASSP-24, No 5, 1976, p. 380–391.
7. Gray, R. M.; Buzo A.; Gray, A. H.; Matsuyama, Y. Distortion measures for speech processing. *IEEE Trans. on Acoustic Speech and Signal Processing*, Vol ASSP-28, No 4, 1980, p. 367–376.
8. Lipeika, A.; Lipeikienė, J. The use of pseudostationary segments for speaker identification. In: Proceedings of the 3rd

European Conference on Speech Communication and Technology, Berlin, Germany, 21-23 September, Vol 3, 1993, p. 2303–2306.

9. De Souza, P., Thomson, P. J. LPC distance measures and statistical tests with particular reference to the likelihood ratio. *IEEE Trans. on Acoustic Speech and Signal Processing,* Vol ASSP-30, No 2, 1982, p. 304–315.
10. Assaleh, K.; Mammone, R. Robust cepstral features for speaker identification. In: Proc. of the International Conference on Acoustics Speech and Signal Processing, Vol I, 1994, p. 129–132.
11. Atal, B. S. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal Acoust. Soc. Amer.* Vol 55, No 6, 1974, p.1304–1312.
12. Atal, B. S. Automatic recognition of speakers from their voices. *Proc. IEEE*, Vol 64, No 4, 1976, p. 460–465.
13. Juang, B. H.; Wang, D. Y. and Gray, A. H. Distortion performance of vector quantization for LPC voice coding. *IEEE Trans. on Acoustic Speech and Signal Processing*, Vol ASSP-30, No 2, 1982, p. 294–304.
14. Rabiner, L. R.; R. Schafer, W. Digital Processing of Speech Signals. Prentice Hall, Inc., Englewood Cliffs, New Jersey, 1978.

KALBĖTOJO IDENTIFIKAVIMO METODO, GRĮSTO VOKALIZUOTŲ GARSŲ KLASTERIZUOTAIS PSEUDOSTACIONARIAIS SEGMENTAIS, TYRIMAS

P. Treigys, A. Lipeika

Santrauka

Nagrinėjama kalbėtojo identifikavimo problema. Identifikuoti naudojami vokalizuotų garsų pseudostacionarūs intervalai. Identifikuojama lyginant vidutinius atstumus tarp tiriamojo ir lyginamųjų kalbėtojų. Požymius sudaro balso trakto tiesinės prognozės modelio koeficientai, gauti iš fonogramos pseudostacionarių intervalų. Nagrinėjama problema aktuali stenografuotojams, kada kalbėtojų yra daug ir kiekvieną jų atpažinti pagal balsą yra gana sunku. Straipsnyje nagrinėjama sugrupuotų duomenų įtaka identifikavimo metodo klaidoms ir skaičiavimo laikui.

Pagrindiniai žodžiai: nepriklausomas nuo teksto identifikavimas, duomenų grupavimas, vektorinis kvantavimas, laiko sąnaudos, vokalizuotų garsų pseudostacionarūs segmentai, identifikavimo klaidos.

Antanas LIPEIKA. Doctor of Science, senior researcher, Recognition Processes Department of the Institute of Mathematics and Informatics, associated professor Radio-electronics and Fundamental Sciences Departments of Vilnius Gediminas Technical University and a professor of the Mathematics and Informatics Department of Vilnius Pedagogical University. Research interests: processing and recognition of random processes, detection of changes in the properties of random processes, signal processing, speech processing, speech and speaker recognition.

Povilas TREIGYS. Doctoral student of the System Analysis Department of the Institute of Mathematics and Informatics, a lecturer of the Vilnius College of Higher Education. Research interests: image analysis, detection and object's feature extraction in image processing, medical image automated objects segmentation, optimization methods, software engineering.