INFORMATICA, 1999, Vol. 10, No. 4, 377–388

Speaker Recognition Based on the Use of Vocal Tract and Residue Signal LPC Parameters

Antanas LIPEIKA, Joana LIPEIKIENĖ
Institute of Mathematics and Informatics
Akademijos 4, 2600 Vilnius, Lithuania
e-mail: lipeika@ktl.mii.lt

Received: October 1999

Abstract. The problem of text-independent speaker recognition based on the use of vocal tract and residue signal LPC parameters is investigated. Pseudostationary segments of voiced sounds are used for feature selection. Parameters of the linear prediction model (LPC) of vocal tract and residue signal or LPC derived cepstral parameters are used as features for speaker recognition. Speaker identification is performed by applying nearest neighbour rule to average distance between speakers. Comparison of distributions of intraindividual and interindividual distortions is used for speaker verification. Speaker recognition performance is investigated. Results of experiments demonstrate speaker recognition performance.

Key words: speaker recognition, speaker identification, speaker verification, LPC features, cepstral features, intraindividual and interindividual distortions, vocal tract and residue signal features.

1. Introduction

Automatic speaker recognition by voice problem can be divided into speaker identification and speaker verification (Furui, 1994). Speaker identification is the process of determining from which of the registered speakers a given utterance comes. Speaker verification is the process of accepting or rejecting the identity claim of a speaker. Speaker recognition methods also can be divided into text-dependent and text-independent methods. Text-dependent methods require the speaker to provide utterance having the same text as training utterance. Text-independent methods do not require a specific text being spoken.

Speaker recognition performance mainly depends on features used for recognition and a decision rule. A number of speaker specific features are used in speaker recognition. The most popular features are (Campbell, 1997) perceptual linear prediction model (PLP), linear prediction coding (LPC), LPC derived cepstrum (LPC-Cep), complex cepstrum, Mel scale cepstrum, normalised cepstrum, d cepstrum, dd cepstrum, pitch, filter banks and others. These features, with the exception of pitch, represent mostly vocal tract parameters. In (Thevenaz *et al.*, 1995) results of speaker verification experiments show the usefulness of the residue signal when used alone, even if it proves to be less efficient than the synthesis filter. When features of synthesis filter and the residue signal were combined, reduction of the error rate was achieved. There the complex cepstrum was used as features of synthesis filter (to represent vocal tract) and real cepstrum – to represent features of residue signal. The joint use of these features was obtained through a weighted sum of the distances observed individually. Similar results were obtained for speaker identification by He *et al.* (1995) and Liu *et al.* (1996). LPC derived cepstral features were used as features of synthesis filter and FFT cepstrum parameters as features of residue signal.

We investigated this problem in a slightly different way. We used LPC model to represent vocal tract as well as residue signal. Features were extracted from pseudostationary segments of voiced sounds. After detection of a pseudostationary segment of voiced sound we estimated synthesis filter LPC parameters (which mainly correspond to vocal tract) and calculated the residue signal by inverse filtering. This residue signal in short segments can be regarded as a pseudoperiodic random signal. Its energy density spectrum is concentrated in the region of low frequencies. Consequently, after lowpass filtering we can decimate this residue signal and evaluate the residue signal LPC model parameters. So we have feature vector corresponding to the pseudostationary segment of voiced sound, which consists of synthesis filter and residue signal LPC parameters. The first set of LPC parameters represents mainly the vocal tract, the second – the residue signal.

Two distance measures are used for speaker recognition. Likelihood ratio distance is used as distance measure between LPC features and cepstral distance is used as distance measure between LPC derived cepstral features.

For identification an average distance between investigative and comparative speakers (Lipeika *et al.*, 1996) was calculated by assigning some weight β to the distance between residue signal parameters and $(1-\beta)$ to the synthesis filter parameters.

For verification coincidence of distributions of intraindividual and interindividual distortions estimates for investigative and comparative speakers was calculated also by assigning weight β to the distance between residue signal parameters and $(1-\beta)$ to the synthesis filter parameters. The likelihood ratio distance was used as a distance measure. Similar investigations were performed by using LPC derived cepstral parameters as feature vectors.

Recognition experiments were performed with the two databases, formed at the Lithuanian Institute of Forensic Examination. One consists of speech phonogramms of 29 men, other – of 11 women. Results of experiments show, that proper use of residue signal in speaker recognition improves considerably speaker recognition performance.

2. Speaker Verification

Let we have investigative speech record X and comparative speech record Y. From pseudostationary segments of voiced sounds of the investigative speech record we can estimate synthesis filter LPC features

$$A_x(i) = \{a_1^x(i), \dots, a_p^x(i), b^x(i)\}, \quad i = 1, 2, \dots, N_x,$$

corresponding mainly to the vocal tract and LPC features

$$C_x(i) = \{c_1^x(i), \dots, c_p^x(i), d^x(i)\}, \quad i = 1, 2, \dots, N_x,$$

corresponding to the residue signal. Here $b^x(i), d^x(i)$ are gain parameters of the LPC model. Similarly, for the comparative speech record we can estimate synthesis filter LPC parameters $A_y(j),\ j = 1, 2, \dots, N_y$ and residue signal LPC features $C_y(j),\ j = 1, 2, \dots, N_y$.

Let us denote distance between ith feature vector of investigative speaker and jth feature vector of comparative speaker corresponding mainly to a vocal tract parameters as $d_{ij}(A_x, A_y)$ and distance between corresponding residue signal features as $d_{ij}(C_x, C_y)$. Let us assign weight $\beta,\ 0 \leqslant \beta \leqslant 1$, which defines influence of the vocal tract and residue signal features to the common distance. So, common distance between ith feature vector of investigative speaker and jth feature vector of comparative speaker can be defined (Lipeika *et al.*, 1996) as

$$d_{ij}(\beta, X, Y) = (1 - \beta) d_{ij}(A_x, A_y) + \beta d_{ij}(C_x, C_y). \tag{1}$$

When $\beta = 0$, common distance depends only on the vocal tract parameters, when $\beta = 1$ – only on residue signal parameters. In any other case this distance depends on vocal tract and residue signal parameters.

If for every feature vector of investigative speaker we find “nearest” feature vector j^* of comparative speaker and for every feature vector of comparative speaker we find “nearest” feature vector i^* of investigative speaker, we can regard distances $d_{ij^*}(\beta, X, Y)$ and $d_{i^*j}(\beta, X, Y)$ as estimation of interindividual distortions. If we divide speech record of comparative speaker into two parts, such kind of distances between feature vectors of these parts can be regarded as estimation of intraindividual distortions. So, we can compare estimates of distributions of intraindividual and interindividual distortions. If investigative and comparative speech records are of the same speaker, distributions of intraindividual and interindividual distortions should coincide. If investigative and comparative speech records are of different speakers, distributions of intraindividual and interindividual distortions should differ significantly.

In our method, histograms of intraindividual and interindividual distortions were used as estimates of distributions of intraindividual and interindividual distortions. When histograms of intraindividual and interindividual distortions coincide more than preassigned threshold, it is regarded that investigative and comparative speech records belong to the same speaker. If not, it is regarded that investigative and comparative speech records belong to the different speakers.

3. Speaker Identification

Let we have speech record of investigative speaker X and M speech records of comparative speakers $Y(l),\ l = 1, \dots, M$. The problem is to find speech record $I(\beta)$ “nearest” to X, that $D(\beta, X, Y(I(\beta))) < D(\beta, X, Y(l))$ when $l \neq I(\beta)$.

Suppose we have extracted vocal tract features

$$A_X(i),\ i = 1, 2, \ldots, N_X, \quad A_{Y(l)}(i),\ i = 1, 2, \ldots, N_{Y(l)}$$

and residue signal features

$$C_X(i),\ i = 1, 2, \ldots, N_X, \quad C_{Y(l)}(i),\ i = 1, 2, \ldots, N_{Y(l)}$$

from these speech records.

Using distance measure (1) defined for two feature vectors, we can calculate the average distance between investigative X and comparative $Y(l)$ speech records as function of the weight β in the following way

$$D(\beta, X, Y(l)) = \frac{1}{N_X} \sum_{i=1}^{N_X} d_{ij^*}(\beta, X, Y(l)) + \frac{1}{N_{Y(l)}} \sum_{i=1}^{N_{Y(l)}} d_{i^*j}(\beta, X, Y(l)). \qquad (2)$$

In the sense of interindividual distortions, this average distance is estimation of the mean of interindividual distortions between speech records X and $Y(l)$.

Then speech record $I(\beta)$ "nearest" to the investigative record X is

$$I(\beta) = \arg\min D(\beta, X, Y(l)) \qquad 1 \leqslant l \leqslant M. \qquad (3)$$

When we use LPC features for speaker recognition, likelihood ratio distance is used in expressions (1)–(3). For LPC derived cepstral features – cepstral distance is used (Lipeika *et al.*, 1996a).

4. Feature Extraction

As it was mentioned in the introduction, features for speaker recognition were extracted from pseudostationary segments of voiced sounds. Selection of pseudostationary segments and extraction of LPC synthesis filter features (corresponding mainly to vocal tract) is described in Lipeika *et al.* (1993).

Let us consider feature extraction from the residue signal (corresponding mainly to excitation signal of the vocal tract). When we have the estimated LPC synthesis filter features we can use these LPC model parameters to perform inverse filtering. If synthesis filter system function is of the form

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_p z^{-p}}, \qquad (4)$$

the inverse filter system function will be $A(z)$. So we can easily perform inverse filtering of the speech signal and output of the inverse filter will be the residue signal.

Residue signal also can be described by LPC model. But energy of the residue signal in the frequency domain is concentrated mainly at low frequencies and the residue signal

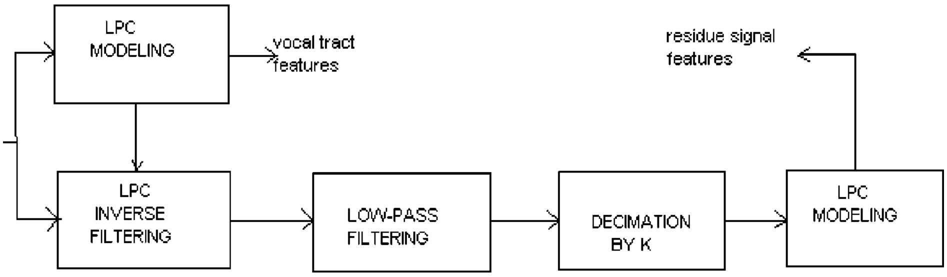

Fig. 1. Block diagram of feature extraction.

must be decimated before LPC modelling. To avoid aliasing in the frequency domain, before decimation we use low-pass filtering to exclude high frequency components of the residue signal. Then for decimated residue signal LPC modelling is used once more and we get another vector of the LPC model parameters (features) corresponding mainly to vocal tract excitation signal. The block diagram of feature extraction is depicted in Fig. 1. As it was mentioned earlier, LPC derived cepstral features also are used for speaker recognition in our study. Calculation of cepstral features from LPC features is described in Atal (1976).

5. Experiments

Two speech data bases of free speech were used for experiments. These data bases were created at the Lithuanian Institute of Forensic Examination. The first database consists of speech records of 11 female speakers: 5 women – tape recorded speech and 6 women – telephone speech. The second database consists of speech records of 29 male speakers: 14 men – tape recorded speech and 15 men – telephone speech. Each speaker in the data base is represented by two speech records – comparative and investigative. Duration of each speech record is about 6–10 minutes. These records were sampled by 16 bit A/D converter. Speech Interactive System (SIS) produced by Speech Technology Centre (St. Petersburg) was used for sampling and editing. The sampling frequency used was 10417 Hz (this sampling frequency is usually used at the Lithuanian institute of Forensic Expertise for phonoscopic examination). Using feature extraction program designed for this purpose about 400 feature vectors representing vocal tract and excitation signal were extracted for each speech record. Decimation factor was chosen 10 for men and 5 for women in feature extraction (accordingly cut-off frequency of the low-pass filter 0.1 (for men) and 0.2 (for women).

5.1. *Speaker Verification Experiments*

In speaker verification experiments after preliminary investigation the threshold of coincidence of intraindividual and interindividual distortions was chosen 0.8. It means that if histograms of intraindividual and interindividual distortions coincide more than 0.8

Table 1

Average coincidence values of intraindividual and interindividual distortions for five women tape recorded speech. LPC features.

	m001–2	m002–2	m003–2	m007–2	m008–2
m001–1	**0.879**	0.057	0.380	0.279	0.359
m002–1	0.072	**0.876**	0.044	0.124	0.221
m003–1	0.397	0.033	**0.895**	0.122	0.216
m007–1	0.293	0.095	0.251	**0.826**	0.371
m008–1	0.398	0.163	0.311	0.425	**0.778** *!!!*

Table 2

Average coincidence values of intraindividual and interindividual distortions for five women tape recorded speech. LPC derived cepstral features.

	m001–2	m002–2	m003–2	m007–2	m008–2
m001–1	**0.891**	0.053	0.372	0.279	0.324
m002–1	0.033	**0.867**	0.041	0.062	0.155
m003–1	0.365	0.055	**0.898**	0.131	0.206
m007–1	0.281	0.083	0.248	**0.827**	0.332
m008–1	0.379	0.173	0.302	0.421	**0.732** *!!!*

(80%), decision is made that investigative and comparative speech records are of the same speaker. If not, decision is made that investigative and comparative speech records belong to different speakers. We calculated the coincidence of intraindividual and interindividual distortions at $\beta = 0$ (coincidence of vocal tract parameters), $\beta = 0.5$ (equal influence of vocal tract and excitation signal parameters) and $\beta = 1.0$ (coincidence of excitation signal parameters) and then the average value was calculated. After that the average value of coincidence was compared with the threshold and final decision was made.

In Table 1 speaker verification results for five women (tape recorded speech) are presented. LPC features were used in this experiment. Numbers in the table means average coincidence value for given investigative and comparative speech records. Names of investigative records are in rows, comparative – in columns.

This table shows that one verification error (4.0%) was made. Speech records m008–1 and m008–2 are not accepted as records of the same speaker – false rejection.

Verification results for the same speech records obtained using LPC derived cepstral features are presented in Table 2. It's seen, it is no significant difference in results (the same error rate) obtained for LPC and LPC derived cepstral features.

Speaker verification results for six women (telephone speech, LPC features) are presented in Table 3. In this case there are four verification errors (11.11%, false acceptance). In four cases records of different speakers are regarded as records of the same speaker.

Table 3

Average coincidence values of intraindividual and interindividual distortions for six women telephone speech. LPC features.

	m004–2	m005–2	m006–2	m009–2	m010–2	m011–2
m004–1	**0.844**	0.772	0.435	0.527	0.502	0.680
m005–1	0.837 *!!!*	**0.890**	0.308	0.416	0.352	0.460
m006–1	0.483	0.0313	**0.869**	0.577	0.807*!!!*	0.786
m009–1	0.650	0.460	0.731	**0.807**	0.721	0.818 *!!!*
m010–1	0.547	0.312	0.691	0.615	**0.888**	0.828 *!!!*
m011–1	0.642	0.380	0.711	0.759	0.717	**0.855**

Table 4

Average coincidence values of intraindividual and interindividual distortions for six women telephone speech. LPC derived cepstral features.

	m004–2	m005–2	m006–2	m009–2	m010–2	m011–2
m004–1	**0.831**	0.750	0.402	0.519	0.506	0.673
m005–1	0.818 *!!!*	**0.881**	0.291	0.395	0.349	0.444
m006–1	0.454	0.249	**0.878**	0.594	0.821*!!!*	0.752
m009–1	0.620	0.398	0.709	**0.840**	0.724	0.808 *!!!*
m010–1	0.533	0.258	0.685	0.618	**0.847**	0.830 *!!!*
m011–1	0.644	0.336	0.672	0.745	0.698	**0.825**

Verification results for the same speech records obtained using LPC derived cepstral features are presented in Table 4. Again, it was no significant difference in results (the same error rate). Similar experiments were performed with speech records of male speakers. For 14 men tape recorded speech using LPC features there were obtained four verification errors (2.04%, 2 – false acceptance and 2 – false rejection). For the same speech records using LPC derived cepstral features there were obtained five verification errors (2.55%, 2 – false acceptance and 3 – false rejection).

For 15 men telephone speech using LPC features there were obtained nineteen errors (8.44%, 3 – false rejection and 16 – false acceptance). Using LPC derived cepstral features there were obtained fourteen verification errors (6.22% : 10 – false acceptance and 4 – false rejection).

From experiments with female speech records we can conclude that there is no difference in performance between LPC and LPC derived cepstral features. Verification of telephone speech records provides higher error rate.

Experiments with the male speech records also demonstrate that verification of telephone speech records provides higher error rate. But LPC derived cepstral features are better than LPC features for telephone speech and a little worse for tape recorded speech.

5.2. *Speaker Identification Experiments*

In speaker identification experiments we used the same speech data bases as in speaker verification. Every speech record in turn was regarded as investigative and compared with all records contained in the database.

First of all we would like to illustrate how much it is possible to improve speaker identification performance by common use (with some weight β) vocal tract and excitation signal features. For this purpose both speech databases of male speakers were linked up and identification was performed at β values: 0.0, 0.1, 0.2, ..., 1.0. For every speech record and β value a reliability reserve was calculated. LPC features were used in this experiment.

The reliability reserve was introduced in (Lipeika *et al.*, 1993). Reliability reserve for every investigative speech record means at what extent further is the "nearest" comparative speech record of different speakers than comparative speech record of the same speaker. If it is possible to increase reliability reserve by changing weight β, it means that it is possible to improve identification performance.

When β is changing from zero to one for the particular speaker, the reliability reserve as usual is changing continuously and at some β value reaches its maximum (Lipeika *et al.*, 1996a). In Fig. 2 a difference between maximum reliability reserve and reliability reserve at $\beta = 0$ (when only vocal tract features are used) for every speech record of linked up speech databases of male speakers is shown. Bars marked with letter "t" correspond to telephone speech records.

We can notice, that not for every speech record it is possible to improve speaker identification performance by common use of vocal tract and residue signal features. Also, from this figure we can not conclude that for tape recorded speech the reliability reserve increases more than for telephone speech. It rather depends on a particular speaker.

Let's regard choosing of the weight β. For solving of this problem, the following experiment was performed. For the male speech database it was evaluated how long the

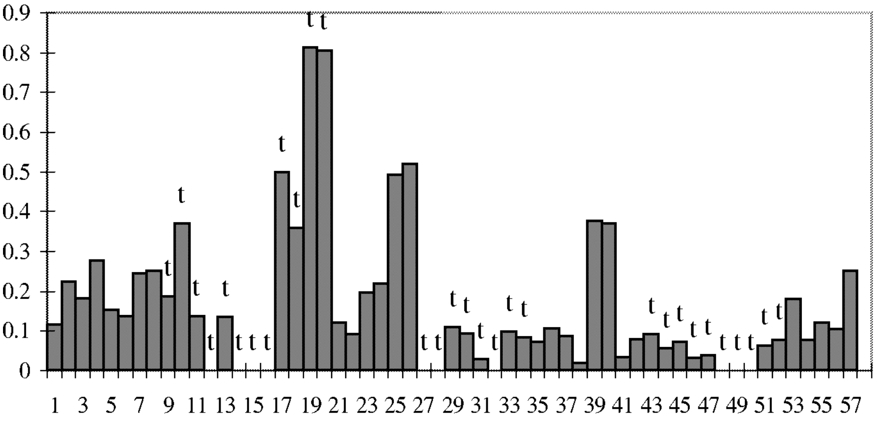

Fig. 2. Difference between maximum reliability reserve and reliability reserve at $\beta = 0$ for every speech record. Male speakers. Bars marked with letter "t" correspond to telephone speech records.

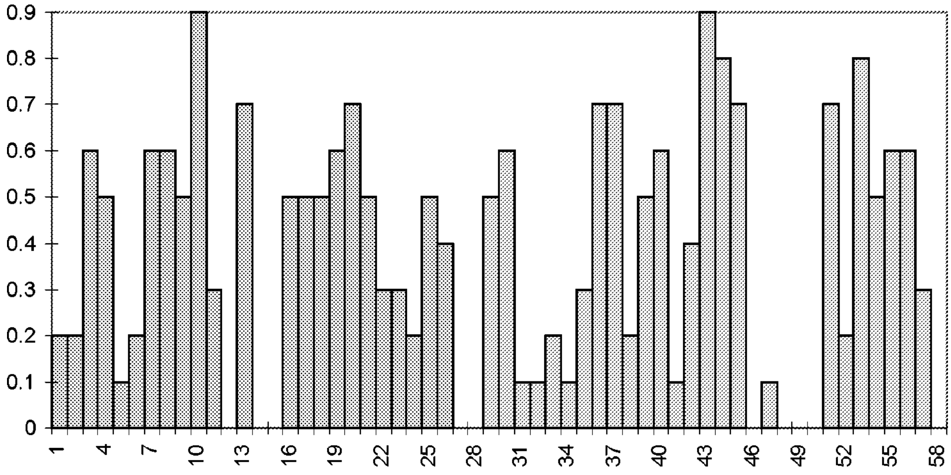

Fig. 3. Values of β at which reliability reserve reaches its maximum for each speech record. Male speakers, LPC features.

reliability reserve increases increasing β from 0 to 1.0. In Fig. 3 it is shown at which β value the reliability reserve reaches its maximum for different speech records. From this chart we can see that for most records maxima are reached at β values in the range from 0 to 0.5. So $\beta = 0.0,\ 0.3$ and 0.5 was chosen and an average distance between speech records was calculated. The results of speaker identification experiments are presented in Table 5.

From the speaker identification experiments we can conclude that there is no identification errors for female and male tape recorded speech records. For male telephone speech using LPC features about 20% identification errors were obtained. Slightly better results were obtained using LPC derived cepstral features.

Table 5

Results of speaker identification experiments for both speech databases and different types of features. Identification at β values 0.0, 0.3, 0.5.

Speech database	type of features	error rate
5 female speakers, tape recorded speech	LPC	0
6 female speakers, telephone speech	LPC	0
14 male speakers, tape recorded speech	LPC	0
15 male speakers, telephone speech	LPC	6 (20.68%)
5 female speakers, tape recorded speech	LPC derived cepstral	0
6 female speakers, telephone speech	LPC derived cepstral	0
14 male speakers, tape recorded speech	LPC derived cepstral	0
15 male speakers, telephone speech	LPC derived cepstral	5 (17.24%)

6. Conclusions

The problem of text-independent speaker recognition based on the use of vocal tract and residue signal LPC and LPC derived cepstral features was investigated. Speaker verification was based on comparison of distributions of intraindividual and interindividual distortions. Average distance between speech records was used for speaker identification. Speech databases of 11 female and 29 male speakers were used for speaker recognition experiments.

From speaker verification experiments with female speech records we can conclude that there is no difference in a performance between LPC and LPC derived cepstral features. Verification of telephone speech records provides a higher error rate.

Experiments with male speech records also demonstrate that verification of telephone speech records provide a higher error rate. But LPC derived cepstral features are better than LPC features for telephone speech and a little worse for tape recorded speech.

From speaker identification experiments we can conclude that there is no identification errors for female and male tape recorded speech records. For male telephone speech using LPC features about 20% identification errors were obtained. Slightly better results were obtained using LPC derived cepstral features.

Acknowledgements

The authors would like to thank Dr. Bernardas Šalna, head of the Department of Phonoscopic Expertise, Lithuanian institute of Forensic Expertise for encouragement and support of our research and research associate Viktor Makarevič for preparation of the speech databases.

References

Atal, B. (1976). Automatic recognition of speakers from their voices. *Proc. of the IEEE*, **64**(4), 460–475.

Campbell, J.P. (1997). Speaker recognition: A tutorial. *Proc. of the IEEE*, **85**(9), 1437–1462.

He, J., L. Liu, G. Palm (1995). On the use of features from prediction residual signals in speaker identification. In *Proc. of the 3rd European Conference on Speech Communication and Technology*. Madrid. pp. 313–316.

Furui, S. (1994). An overview of speaker recognition technology. In *Proc. of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*. Martigny. pp. 1–9.

Lipeika A., J. Lipeikienė. (1996). Recent advances in speaker identification. In *Proc. of the Elsnet Goes East and IMAC Workshop "Integration of Language and Speech*. Moscow. pp. 97–106.

Lipeika A., J. Lipeikienė (1996a). Speaker identification methods based on pseudostationary segments of voiced sounds. *Informatica*, **7**(4), 469–484.

Lipeika A., J. Lipeikienė (1993). The Use of Pseudostationary Segments for Speaker Identification. In *Proc of the 3rd European Conference on Speech Communication and Technology*. Berlin, Germany. pp. 2303–2306.

Liu, L., J. He, G. Palm (1996). Signal modeling for speaker identification. In *ICAASP'96*. pp. 665–668.

Thevenaz, P., H. Hugli (1995). Usefulness of the LPC-residue in text-independent speaker verification, *Speech Communication*, **17**, 145–157.

A. Lipeika is a doctor of technical sciences, a senior researcher of the Recognition Processes Department at the Institute of Mathematics and Informatics and an associate professor at the Radioelectronics Department of Vilnius Technical University. Scientific interests include: processing and recognition of random processes, detection of changes in the properties of random processes, signal processing and speaker recognition.

J. Lipeikienė is a doctor of technical sciences, a senior researcher of the Recognition Processes Department at the Institute of Mathematics and Informatics and an Associate Professor at the Informatics Department of Vilnius Pedagogical University. Scientific interests include: processing of random signals, including speech signals, robust methods for determination of change-points in the properties of random processes, data compression.

Kalbančiojo atpažinimas, panaudojant balso trakto ir liekanos signalo LPC parametrus

Antanas LIPEIKA, Joana LIPEIKIENĖ

Darbe yra nagrinėjamas nepriklausančio nuo teksto kalbančiojo atpažinimo, panaudojant balso trakto ir liekanos signalo LPC parametrus, uždavinys. Požymių išskyrimui yra naudojami vokalizuotų garsų pseudostacionarūs intervalai. Balso trakto ir liekanos signalo tiesinės prognozės (LPC) modelio parametrai arba iš LPC parametrų paskaičiuoti kepstro koeficientai yra naudojami kaip požymiai kalbančiojo atpažinimui. Kalbančiojo identifikavimas yra atliekamas pritaikant vidutiniam atstumui tarp kalbančiųjų artimiausio kaimyno taisyklę. Kalbančiojo verifikavimui yra naudojamas intraindividualių ir interindividualių iškraipymų pasiskirstymų įverčių sulyginimas. Atliktas kalbančiojo atpažinimo algoritmų darbingumo eksperimentinis tyrimas. Eksperimentų rezultatai demonstruoja kalbančiojo atpažinimo metodo darbingumą.