INFORMATICA, 2002, Vol. 13, No. 1, 37–46

Development of Isolated Word Speech Recognition System

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

Received: September 2001

Abstract. The isolated word speech recognition system based on dynamic time warping (DTW) has been developed. Speaker adaptation is performed using speaker recognition techniques. Vector quantization is used to create reference templates for speaker recognition. Linear predictive coding (LPC) parameters are used as features for recognition. Performance is evaluated using 12 words of Lithuanian language pronounced ten times by ten speakers.

Key words: speech recognition, speaker recognition, dynamic time warping, vector quantization, LPC features.

1. Introduction

Speech recognition is the process of automatic extracting and determining linguistic information conveyed by a speech wave using computers (Furui, 2001). Automatic speech recognition methods have been investigated for many years aimed at realizing transcription and human-computer interaction systems. The first technical paper to appear on speech recognition was published in 1952. It described Bell Labs spoken digit recognizer Audrey (Davis *et al.*, 1952). The system relied on measuring spectral resonances during the vowel region of each digit. In the 1960s several fundamental ideas, such as filter bank spectrum analysis, zero crossing analysis, time-normalization methods, in speech recognition were published (Rabiner *et al.*, 1993). Dynamic programming method for time aligning of a pair of speech utterances was proposed (Vintsyuk, 1968). In the 1970s isolated word recognition became advanced technology due to fundamental studies (Velichko *et al.*, 1970; Sakoe *et al.*, 1978; Itakura, 1975). Pattern recognition, dynamic programming, linear predictive coding (LPC) ideas were applied to speech recognition. Speech recognition systems were made truly speaker independent (Rabiner *et al.*, 1979). In the 1980s a focus of research was the problem of connected word recognition. Speech research was shifted from template based approaches to statistical modeling methods, hidden Markov model (HMM) approach (Rabiner, 1989; Jelinek, 1999) and neural network methods (Weibel, 1989). In the 1990s main focus of research was large vocabulary continuous speech recognition (Lee *et al.*, 1996), robust speech recognition (Junqua *et al.*, 1996), including syntax, semantics, pragmatics into speech recognition higher level

processing (Jurafsky *et al.*, 2000). Speech recognition systems have been developed for a wide variety of applications, ranging from small vocabulary word recognition to large vocabulary speech dictation.

Our aim is to develop isolated word speech recognition system for Lithuanian language. We use the rich experience of speech recognition system developers in implementation of our system: LPC features and dynamic time warping. To make the system speaker independent we use our experience in speaker recognition system development (Lipeika *et al.*, 1993, 1996). Performance of the system was evaluated.

2. Feature Extraction

Well known and widely used (Rabiner *et al.*, 1993) framework was applied to feature extraction from speech signals. Fig. 1 shows a block diagram of this framework, LPC processor.

Endpoint detection. First step in feature extraction is speech endpoint detection. Endpoint detection is based on signal energy evaluation. Background noise energy level is evaluated at the beginning and the end of speech signal and energy thresholds are applied to find speech beginning and end points. Simple logic is used to make a decision.

Preemphasis. Preemphasis is used to flatten speech signal spectrum and to make speech signal less sensitive to finite precision effects later in the speech signal processing. Simple a first order infinite impulse response filter was used for preemphasis:

$$y'(n) = y(n) - 0.95y(n-1). \tag{1}$$

Frame blocking. The preemphasized speech signal $y'(n)$ is blocked into frames of $N = 256$ samples, with adjacent frames separated by $M = 128$ samples.

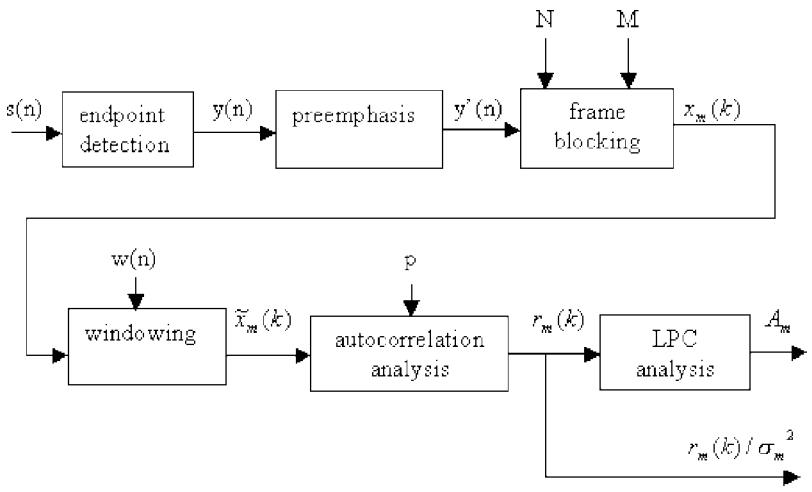

Fig. 1. Block diagram of feature extraction framework.

Windowing. Windowing is used to minimize the speech signal discontinuities at the beginning and end of each analysis frame. The Hamming window was used for windowing:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi\, n}{N-1}\right), \quad 0 \leqslant n \leqslant N-1. \tag{2}$$

Autocorrelation analysis. Each frame of windowed signal was autocorrelated using expression

$$r_m(k) = \sum_{n=0}^{N-1-k} \tilde{x}_m(n)\tilde{x}_m(n+k), \quad k = 0, 1, \ldots, p \tag{3}$$

where $p = 10$ is the order of the LPC analysis.

LPC analysis. Durbin's method was used to solve recurrently LPC analysis equations:

$$E_0 = r_m(0);$$

$$a_i(i) = k_i = \left\{ r_m(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} r_m\big(|i-j|\big) \right\} / E^{(i-1)}, \quad i = 1, 2, \ldots, p; \tag{4}$$

$$a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)}; \tag{5}$$

$$E^{(i)} = (1 - k_i^2) E^{(i-1)}. \tag{6}$$

The LPC analysis results are linear prediction model coefficients $A_m = \{a_1, \ldots, a_p\}$, where

$$a_k = a_k^{(p)}, \quad k = 1, \ldots, p,$$

and LPC model gain term

$$\sigma_m^2 = E^{(p)}.$$

LPC coefficients A_m and gain normalized autocorrelation coefficients $r_m(k)/\sigma_m^2$, $k = 0, 1, \ldots, p$ are stored as feature vectors of each analysis frame. This form of storage is convenient for distance calculation at the recognition stage.

3. Speech Recognition

Our development of isolated word speech recognition system is based on a use of dynamic time warping (DTW) for speech pattern matching (Itakura, 1975). The DTW process nonlinearly expands or contracts the time axis to match the same phoneme positions between the input speech and reference templates. Block diagram of our speech recognition system is shown in Fig. 2. Test and reference templates are time sequences of feature

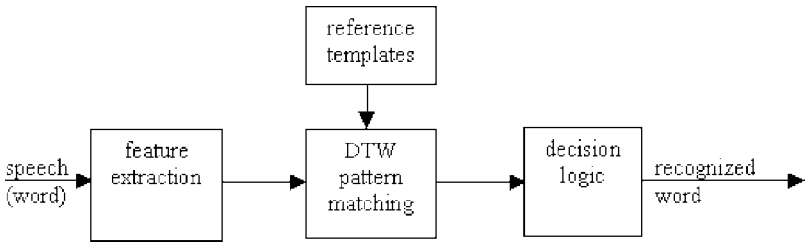

Fig. 2. Block diagram of isolated word speech recognition system.

vectors

$$\tilde{S}_m = \{\tilde{A}_m, \tilde{r}_m(k)/\tilde{\sigma}_m^2,\ k = 0, 1, \ldots, p\}, \quad m = 0, 1, \ldots, \tilde{T},$$

and

$$S_n = \{A_n, r_n(k)/\sigma_n^2,\ k = 0, 1, \ldots, p\}, \quad n = 0, 1, \ldots, T,$$

corresponding to test and reference utterances.

Distance measure. Symmetrisized likelihood ratio distance was used to compare a pair of test and reference feature vectors. Likelihood ratio distance can be expressed in the time domain as

$$d_{LR}(\tilde{S}, S) = \left\{ \frac{\tilde{r}(0)}{\tilde{\sigma}^2} r_a(0) + 2 \sum_{i=1}^{p} \frac{\tilde{r}(i)}{\tilde{\sigma}^2} r_a(i) \right\} - 1, \tag{7}$$

where $r_a(i), i = 0, 1, \ldots, p$ are autocorrelations of LPC model parameters of the reference feature vector

$$r_a(i) = \sum_{k=0}^{p-i} a_{k+i} a_k, \quad i = 0,\ 1,\ 2, \ldots, p. \tag{8}$$

Time index is omitted in these expressions for convenience.

The likelihood ratio distance has spectral interpretation (Juang *et al.*, 1982)

$$d_{LR}(\tilde{S}, S) = \int_{-\pi}^{\pi} \frac{\tilde{S}(\theta)/\tilde{\sigma}^2}{S(\theta)/\sigma^2} \frac{d\theta}{2\pi} - 1, \tag{9}$$

where $\tilde{S}(\theta)$ and $S(\theta)$ are spectral densities corresponding to test and reference LPC model. Likelihood ratio distance measure is not symmetric:

$$d_{LR}(\tilde{S}, S) \neq d_{LR}(S, \tilde{S}). \tag{10}$$

This distance measure can be symmetrisized as follows:

$$d(\tilde{S}, S) = \frac{d_{LR}(\tilde{S}, S) + d_{LR}(S, \tilde{S})}{2}. \tag{11}$$

Dynamic Time Warping. Dynamic time warping is minimization of a global test and reference speech pattern dissimilarity measure (Rabiner *et al.*, 1993). Let we have test template $X = \{\tilde{S}_m,\ m = 0, 1, \ldots, T_x\}$ and the reference template $Y = \{S_n,\ n = 0, 1, \ldots, T_y\}$. We can define two warping functions Φ_x and Φ_y, which relate the indices of the two speech patterns, i_x and i_y to a common time axis k, i.e.,

$$i_x = \Phi_x(k), \quad k = 1, 2, \ldots, T,$$

and

$$i_y = \Phi_y(k), \quad k = 1, 2, \ldots, T.$$

Then a global speech pattern dissimilarity measure $d_\Phi(X, Y)$ can be defined based on the warping fuuctions as the accumulated distortion over the entire pattern

$$d_\Phi(X, Y) = \sum_{k=1}^{T} d\big(\Phi_x(k), \Phi_y(k)\big) m(k)/M_\Phi, \tag{12}$$

where $d(\Phi_x(k), \Phi_y(k)) = d(\tilde{S}_{i_x}, S_{i_y}), m(k)$ is path weighting coefficient and M_Φ is a path normalizing factor.

Dynamic programming method (Bellman, 1957) is used for global dissimilarity measure minimization. Dynamic programming solution depends on a number of meaningful constraints (endpoint, monotonicity, local continuity, global path, slope weighting) and is out of scope of this paper. Constraints, used in our system, were defined by (Itakura, 1975). Local continuity constraints are shown in Fig. 3.
Local continuity constraints define from which points on the search grid we can reach predefined point on a grid.

During DTW pattern matching test pattern is matched to each reference template and global dissimilarity measure is calculated. A reference template providing minimal global dissimilarity measure is accepted as recognized word by the decision logic.

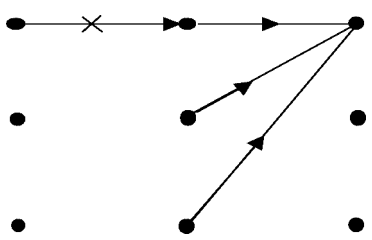

Fig. 3. Local continuity constraints, defined by Itakura. Two consecutive horizontal moves are not allowed.

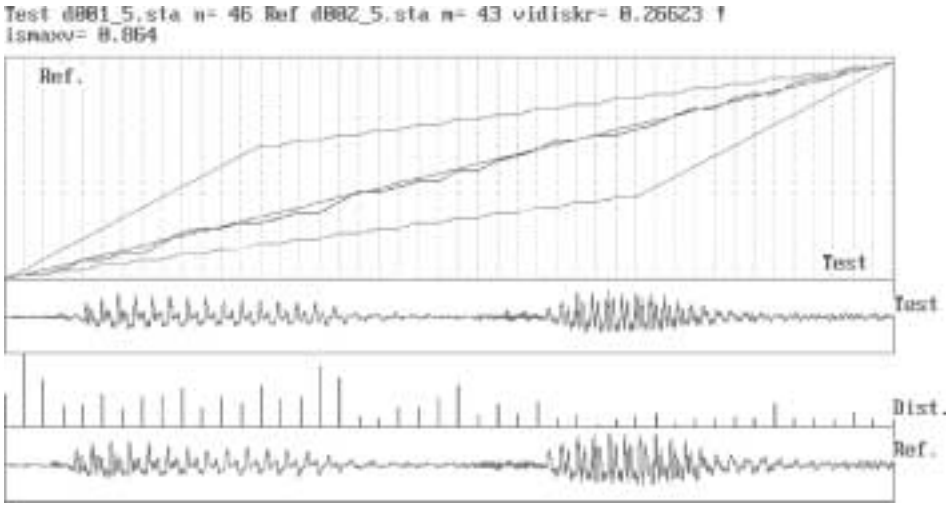

Fig. 4. Illustration of dynamic time warping process. The two utterances of the Lithuanian word “penki”.

Illustration of dynamic time warping process is presented in Fig. 4.

Dynamic time warping of two utterances of the Lithuanian word “penki” pronounced by the same speaker is presented in the picture. Global path constraints, linear time alignment path and dynamic time warping paths are displayed on the top of the picture. The two utterances and distances on the optimal path grid points are displayed below. Obtained dissimilarity measure for these utterances is 0.26623.

4. Speaker Dependent Isolated Word Recognition Experiments

Twelve Lithuanian words, digits 0–9 and “pradžia”, “pabaiga” pronounced by 10 speakers were used for recognition. Each word was pronounced ten times by each male speaker in noise-free conditions, yielding 120 utterances per speaker, total 1200 utterances. Each utterance was sampled at a rate of 11025 Hz. Other analysis and recognition parameters are presented in description of the algorithms. The first utterance of each word of each speaker was used as a reference template for speaker dependent speech recognition. Speech recognition results are shown in Table 1.

The main source of errors is that only one word of each speaker was used as a reference template for recognition.

Table 1

Results of speaker dependent speech recognition

Speakers	D0	D1	D2	D3	D4	D5	D6	D7	D8	D9
Errors %	0	0	0	2.8	0	4.6	0	0	0	0.9

Total error %: 0.83

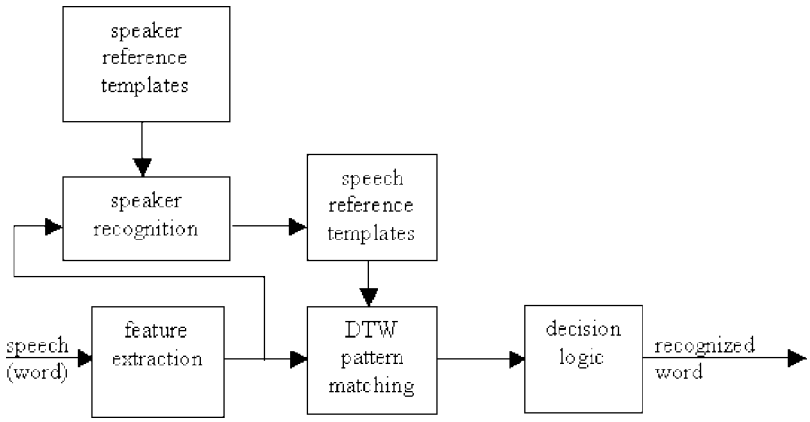

Fig. 5. Block diagram of a speaker-independent speech recognition system.

5. Speaker Independent Isolated Word Recognition Experiments

Speech recognition systems should be speaker independent, i.e., inter-speaker variability should be eliminated. Various adaptation methods are used to deal with inter-speaker variability (Junqua, 1996). Our approach to solution of this problem is to use speaker recognition to deal with inter-speaker variability. We use text-independent speaker recognition method (Lipeika, *et al.*, 1993) to find reference templates most suitable for a word provided for recognition. Block diagram of our speaker-independent speech recognition system is shown in Fig. 5.

Speaker reference templates were created from the same speech utterances as those for speech recognition. Distance between a word provided for recognition and speaker reference template was calculated using full search algorithm

$$D_{XA_i} = \frac{1}{N_X} \sum_{j \in X} \min_{l \in A_i} d_{j\,l}(X, A_i). \tag{13}$$

Here N_x is the number of feature vectors in a word X provided for recognition; A_i is a speaker reference template; $d_{jl}(X, A_i)$ is the distance between j-th frame of X and l-th frame of A_i. Speaker "nearest" to the word X is defined as

$$\hat{I} = \min_{1 \leqslant i \leqslant n} D_{XA_i}, \tag{14}$$

where n is the number of speakers. Then we select speech reference templates of the speaker $\tilde{I}$ as the most suitable for recognition of the word X. Speech recognition results are shown in Table 2.

As we see from Table 1, total recognition error increased from 0.83% for speaker dependent speech recognition to 1.94% for speaker independent speech recognition.

Table 2

Results of speaker independent speech recognition

Speakers	D0	D1	D2	D3	D4	D5	D6	D7	D8	D9
Errors %	0	0.9	3.7	3.7	3.7	3.7	0	2.8	0	0.9

Total error %: 1.94

6. Speech Recognition Using Vector Quantization

One problem arises in our speaker independent speech recognition approach. Due to a large number of feature vectors (about 400) in speaker reference templates amount of computation significantly increases. Solution of this problem is reducing a number of feature vectors in speaker reference templates. We used clusterization to solve this problem. Two clusterization – vector quantization methods were used to reduce a number of feature vectors. One (VQ1) was developed by Juang *et al.* (1982) based on splitting every cluster into two clusters, another (VQ2) – by Lipeika *et al.* (1995) based on splitting a cluster with largest average distortions into two clusters.

We carried out speaker independent speech recognition experiments to evaluate performance of these methods and to show how performance depends on the codebook size. Codebook sizes were chosen 32, 64 and 128. Speech recognition results are shown in Table 3.

As we see from Table 3 better results were obtained when we used vector quantization based on splitting a cluster with largest average distortions into two clusters. In this case only 32 feature vectors are needed to create a speaker reference template, so using this method we significantly reduce a computation amount. Unfortunately, reduction of a computation amount increases speech recognition error from 1.94% to 2.5%.

Table 3

Results of speaker independent speech recognition using vector quantization

Speakers	Speaker depend. recogn.	Speaker independ recogn.	VQ1 32 ref. patt.	VQ1 64 ref. patt.	VQ1 128 ref. patt.	VQ2 32 ref. patt.	VQ2 64 ref. patt.	VQ2 128 ref. patt.
D0	0	0	2.8	1.8	20.3	2.8	4.6	5.5
D1	0	0.9	0.9	0.9	2.8	0.9	0.9	0.9
D2	0	3.7	2.8	0.9	1.8	1.8	1.8	1.8
D3	2.8	3.7	2.8	3.7	10.2	2.8	2.8	2.8
D4	0	3.7	5.5	8.3	20.3	6.5	4.6	5.5
D5	4.6	3.7	4.6	3.7	21.3	4.6	6.5	4.6
D6	0	0	2.8	1.8	4.6	0	2.8	1.8
D7	0	2.8	2.8	1.8	6.5	2.8	0.9	1.8
D8	0	0	0	0.9	2.8	0.9	0.9	0
D9	0.9	0.9	1.8	1.8	3.7	1.8	0.9	0.9
Total error %	0.83	1.94	2.68	2.59	9.44	2.5	2.68	2.59

7. Conclusions

Isolated word speech recognition system based on dynamic time warping (DTW) and linear predictive coding features (LPC) has been developed. Speaker adaptation is performed using speaker recognition techniques. Vector quantization is used to create reference templates for speaker recognition.

Twelve Lithuanian words, digits 0–9 and "pradžia", "pabaiga" pronounced 10 times by 10 speakers were used for performance evaluation. Results of experiments showed that:

- recognition error rate in speaker dependent mode is 0.83%;
- recognition error rate in speaker independent mode is 1.94%;
- using vector quantization in speaker independent mode best result was obtained when vector quantization was based on splitting a cluster with largest average distortions into two clusters and codebook size was chosen 32. Recognition error rate was obtained 2.5%. Computation amount was significantly reduced on account of slightly increased error rate.

In the future performance of the system should be tested on large amount of speech data.

Acknowledgments

The authors would like to thank PhD student Bronislovas Balvočius for his contribution in preparing the speech database and MSc student Sigita Laurinčiukaitė for performing tadeous recognition experiments.

References

R.E. Bellman (1957). *Dynamic Programming*, Princeton University Press, Princeton, New Jersey, USA.

Davis K.H., Biddulp R., S. Balashek (1952). Automatic recognition of spoken digits. *J. Acoust. Soc.Amer.,* **24**(6), 637–642.

Furui S. (2001). *Digital Speech Processing, Synthesis and Recognition,* Marcel Dekker, Inc.

Itakura F. (1975). Minimum prediction residual applied to speech recognition. *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-23 (1), 67–72.

Jelinek F. (1999). *Statistical Methods to Speech Recognition.* MIT Press.

Juang B.H., D.Y. Wang, A.H. Gray (1982). Distortion performance of vector quantization for LPC voice coding. *IEEE Tranc. on Acoustic Speech and Signal Processing*, ASSP-30 (2), 294–304.

Junqua J.-C., J.-P. Haton (1996). *Robustness in Automatic Speech Recognition, Fundamentals and Applications*. Kluwer Academic Publishers.

Jurafsky D., J.H. Martin (2000). *Speech and Language Processing*. Prentice Hall.

Lee Ch.-H., Soong F.K., K.K. Palival (1996). *Automatic Speech and Speaker Recognition.* Kluwer Academic Publishers.

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, 21–23 September, pp. 2303–2306.

Lipeika A., J. Lipeikienė (1995). Speaker identification using vector quantization. *Informatica*, **6**(2), 167–180.
Lipeika A., J. Lipeikienė (1996). Speaker identification methods based on pseudostationary segments of voiced sounds. *Informatica*, **7**(4), 469–484.
Rabiner L.R., S.E. Levinson, A.E. Rosenberg, J.G. Wilpon (1979). Speaker independent recognition of isolated words using clustering techniques. *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-27, 336–349.
Rabiner L.R. (1989). A Tutorial on hidden markov models and selected applications in speech recognition, *Proc. IEEE*, **77**(2), 257–289.
Rabiner L., B.-H. Juang (1993). *Fundamentals of Speech Recognition.* Prentice Hall.
Sakoe H., S. Chiba (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-26 (1), 43–49.
Vintsyuk T.K. (1968). Speech Discrimination by Dynamic Programming. *Kibernetika*, **4**(2), 81–88.
Velichko V.M., N.G. Zagoruyko (1970). Automatic Recognition of 200 Words. *Int. J. Man-Machine Studies*, **2**, 223.
Weibel A., T. Hanazawa, G. Hinton, K. Shikano, K. Lang (1989). Phoneme recognition using time-delay neural networks. *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-37, 393–404.

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, speech processing, speech 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, e-learning.

L. Telksnys is a professor, doctor habilitatis in informatics, Doctor Honoris Causa of the Kaunas University of Technology, a head of the Recognition Processes Department at the Institute of Mathematics and Informatics. He is the author of an original theory of detecting changes in random processes and the developer of a computerized system for statistical analysis and recognition of random signals. His current research interests are: recognition of random processes, speech processing and computerized multimedia systems.

Atskirai pasakytų žodžių atpažinimo sistemos kūrimas

Antanas LIPEIKA, Joana LIPEIKIENĖ, Laimutis TELKSNYS

Atskirai pasakytų žodžių atpažinimo sistema, besiremianti dinaminiu laiko kraipymu, yra kuriama. Sistemos prisitaikymas prie kalbėtojo yra atliekamas naudojant kalbančiojo atpažinimo metodus. Atpažįstant kalbantįjį, etalonų sudarymui yra naudojamas vektorinis kvantavimas. Kaip požymiai atpažinimui yra naudojami tiesinės prognozės modelio parametrai. Sistemos darbingumas yra įvertintas atpažįstant 12 lietuvių kalbos žodžių, kuriuos po 10 kartų ištarė 10 skirtingų kalbėtojų.