Gintautas TAMULEVIČIUS

DEVELOPMENT OF ISOLATED WORD RECOGNITION SYSTEMS

Summary of Doctoral Dissertation
Technological Sciences, Informatics Engineering (07T)

1496-M

VGTU
LEIDYKLA
TECHNIKA

Vilnius 2008

VILNIUS GEDIMINAS TECHNICAL UNIVERSITY
INSTITUTE OF MATHEMATICS AND INFORMATICS

Gintautas TAMULEVIČIUS

DEVELOPMENT OF ISOLATED WORD RECOGNITION SYSTEMS

Summary of Doctoral Dissertation
Technological Sciences, Informatics Engineering (07T)

Vilnius VGTU LEIDYKLA TECHNIKA 2008

Doctoral dissertation was prepared at the Institute of Mathematics and Informatics in 2003–2008.

Scientific Supervisor
Assoc Prof Dr Antanas Leonas LIPEIKA (Institute of Mathematics and Informatics, Technological Sciences, Informatics Engineering – 07T).

The dissertation is being defended at the Council of Scientific Field of Informatics Engineering at Vilnius Gediminas Technical University:

Chairman
Prof Dr Habil Romualdas BAUŠYS (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – 07T).

Members:
Prof Dr Habil Feliksas IVANAUSKAS (Vilnius University, Physical Sciences, Informatics – 09P),
Prof Dr Egidijus KAZANAVIČIUS (Kaunas University of Technology, Technological Sciences, Informatics Engineering – 07T),
Prof Dr Habil Kazys KAZLAUSKAS (Institute of Mathematics and Informatics, Technological Sciences, Informatics Engineering – 07T),
Prof Dr Habil Laimutis TELKSNYS (Institute of Mathematics and Informatics, Technological Sciences, Informatics Engineering – 07T).

Opponents:
Dr Pijus KASPARAITIS (Vilnius University, Technological Sciences, Informatics Engineering – 07T),
Assoc Prof Dr Dalius NAVAKAUSKAS (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – 07T).

The dissertation will be defended at the public meeting of the Council of Scientific Field of Informatics Engineering in the Conference and Seminar Centre of the Institute of Mathematics and Informatics at 2 p. m. on 19 of June 2008.
Address: Goštauto g. 12, LT-01108 Vilnius, Lithuania.
Tel.: +370 5 274 4952, +370 5 274 4956; fax: +370 5 270 0112;
e-mail: doktor@adm.vgtu.lt
The summary of the doctoral dissertation was distributed on 19 May 2008.
A copy of the doctoral dissertation is available for review at the Library of Vilnius Gediminas Technical University (Saulėtekio al. 14, LT-10223 Vilnius, Lithuania) and at the Library of Institute of Mathematics and Informatics (Akademijos g. 4, LT-08663 Vilnius).

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS INSTITUTAS

Gintautas TAMULEVIČIUS

PAVIENIŲ ŽODŽIŲ ATPAŽINIMO SISTEMŲ KŪRIMAS

Daktaro disertacijos santrauka
Technologijos mokslai, informatikos inžinerija (07T)

VGTU
Vilnius LEIDYKLA TECHNIKA 2008

Disertacija rengta 2003–2008 metais Matematikos ir informatikos institute.

Mokslinis vadovas
doc. dr. Antanas Leonas LIPEIKA (Matematikos ir informatikos institutas, technologijos mokslai, informatikos inžinerija – 07T).

Disertacija ginama Vilniaus Gedimino technikos universiteto Informatikos inžinerijos mokslo krypties taryboje:

Pirmininkas
prof. habil. dr. Romualdas BAUŠYS (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07T).

Nariai:
prof. habil. dr. Feliksas IVANAUSKAS (Vilniaus universitetas, fiziniai mokslai, informatika – 09P),
prof. dr. Egidijus KAZANAVIČIUS (Kauno technologijos universitetas, technologijos mokslai, informatikos inžinerija – 07T),
prof. habil. dr. Kazys KAZLAUSKAS (Matematikos ir informatikos institutas, technologijos mokslai, informatikos inžinerija – 07T),
prof. habil. dr. Laimutis TELKSNYS (Matematikos ir informatikos institutas, technologijos mokslai, informatikos inžinerija – 07T).

Oponentai:
dr. Pijus KASPARAITIS (Vilniaus universitetas, technologijos mokslai, informatikos inžinerija – 07T),
doc. dr. Dalius NAVAKAUSKAS (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07T).

Disertacija bus ginama viešame Informatikos inžinerijos mokslo krypties tarybos posėdyje 2008 m. birželio 19 d. 14 val. Matematikos ir informatikos instituto konferencijų ir seminarų centre.
Adresas: Goštauto g. 12, LT-01108 Vilnius, Lietuva.
Tel.: (8 5) 274 4952, (8 5) 274 4956; faksas (8 5) 270 0112;
el. paštas doktor@adm.vgtu.lt
Disertacijos santrauka išsiuntinėta 2008 m. gegužės 19 d.
Disertaciją galima peržiūrėti Vilniaus Gedimino technikos universiteto (Saulėtekio al. 14, LT-10223 Vilnius, Lietuva) ir Matematikos ir informatikos instituto (Akademijos g. 4, LT-08663 Vilnius, Lietuva) bibliotekose.

VGTU leidyklos „Technika” 1496-M mokslo literatūros knyga.

General characteristics of the dissertation

Topicality of the problem. Design and development of speech recognition components require great human and time resources and that causes high costs for this kind of products. Consequently, commercial speech recognition products are designed for wide-used languages such as English, whereas small languages like Lithuanian are ignored. The result of all this situation is speech recognition research for small languages are specialized in adaptation and modification of solutions for wide-used languages.

In order to improve this situation, it is necessary to expand Lithuanian speech recognition research by developing new methodical and technical speech recognition tools, establishing new research trends and applying speech recognition solutions to practical goals.

Aim and tasks of the work: to propose solutions to increase the accuracy and efficiency of the word recognition system without modifying recognition and signal analysis methods. In pursuance of this aim the following issues were dealt with:

1. Increasing the stability and robustness of word endpoint detection.
2. Creation of word references enabling to increase the accuracy and efficiency of the recognition.
3. Segmentation of utterances. Potentiality of recognition of utterance phones.
4. Development of an isolated word recognition system and experimental analysis of proposed procedures and methods.

Scientific novelty

1. The method of the automatic detection of word endpoints.
2. Creation of word reference, using the clustering technique – references are created minimizaing average distances between utterances (distances are calculated using the dynamic time warping approach).
3. Two word segmentation approaches are created: maximum likelihood and minimal prediction error. Words are segmented into phones.
4. Word recognition by phones. A prototype of the word recognition by phones system was created, using the minimal prediction error segmentation approach.

Methodology of research includes mathematical analysis, digital signal processing, and pattern recognition theory. The word recognition system was built using *Microsoft* development environment *Visual Studio 6.0*.

Practical value. Word endpoint detection, clustering-based training procedure and segmentation approaches here have been implemented in the word recognition system *Atpazinimas*, word segmentation system *Segmentacija*, and in the web browser and application control system. The word recognition and segmentation systems were included in the Lithuanian speech recognition research works by the programme "Lithuanian Language in the Information Society" for 2000–2006.

Defended propositions

1. The method of automatic word endpoint detection decreases the level of recognition errors caused by endpoint detection errors.
2. Clustering-based system training procedure increases the recognition accuracy with smaller number of references per word.
3. The developed methodology of word segmentation detects phone boundaries as change moments of the linear prediction model of the speech signal.
4. Experimental results of isolated word recognition and segmentation prove effectiveness of proposed procedures and methods.

Approval of the work. The main results were published in 6 scientific papers: 2 articles in the reviewed periodical publications from the list, approved by the Science Council of Lithuania and 4 papers in the proceedings of conferences. The results were presented in 3 international and 3 national conferences.

The scope of the scientific work. The scientific work consists of six chapters, the list of literature, the list of publications, and one appendix. The total scope of dissertation is 124 pages, 35 pictures, 14 tables, and 1 appendix. The dissertation is written in Lithuanian.

1. Speech recognition systems

The basic speech recognition system consists of three parts: analysis block, a set of references and classification block (Fig 1).

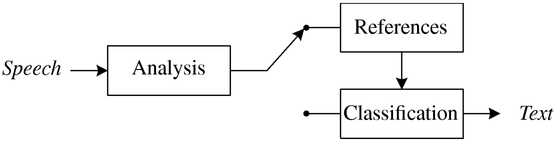

Fig 1. Structure of the speech recognition system

The analysis block analyses speech signal (usually framed) by extracting features, i. e. data characterizing the linguistic content of a signal. A set of a stored utterance features is called a reference. The classification block classifies references according to the distance (similarity) between the reference and analyzed utterance thus making decisions on the best match.

2. Analysis of the speech recognition systems

The dynamic time warping approach has been chosen for word comparison in our recognition system. The choice has been made with our notion that dynamic time warping (DTW) is more applicable to isolated word recognition for its straightforwardness, effectiveness, and lax requirements to training sets.

The chosen DTW approach holds both advantages and disadvantages. In our opinion, advantages are as follows:

- Plainness of algorithm – the DTW algorithm is simple and readily realizable.
- DTW makes an acoustic analysis avoiding the linguistic analysis – a higher level analysis like grammatical, syntactical, etc.
- Straightforward integration of higher level analysis stages. The output of the comparison process could be readily sent for linguistic processing.

However, simplicity of the approach causes some drawbacks. They are:

- Recognition accuracy dependence on the number of references – more word references yield a higher word recognition accuracy.
- Dependence of utterance analysis duration on the utterance length – a longer word means a longer word recognition process.
- Dependence of recognition process duration on the total number of references. Word comparison is performed with all references successively – more references mean a longer recognition process.

3. Implementation of the recognition system

3.1. Potential improvements of the recognition process

We restricted our study to the improvement of isolated word (utterance) recognition avoiding the analysis of feature sets. We have formulated the following trends to improve the DTW approach:

- **Short analysis units.** By choosing shorter units for recognition we can reduce the number of references – shorter segments occur in the spoken speech more often than longer ones. A shorter recognition unit results in a shorter comparison process of references.
- **Optimality of the reference set.** One of the possible ways to increase robustness of a recognition system to noise and speaker is enlargement of the

reference set. An unfavourable outcome of this solution is a longer comparison process. Therefore references must be selected in a manner of an optimal uterrance set.

- **Optimization of the comparison process.** Every reference is completely examined, as usual. Considering some references prospectless for the best match, there should be a possibility to avoid a complete examination of such references. That should shorten the average utterance analysis duration.

We tried to solve reference optimality and prospectless reference rejection problems. Using utterance segmentation into phones, we have proposed a shortened recognition unit and a simplified comparison process thereby.

The practical thesis result is an isolated word recognition and segmentation system KAS. The system has three operation modes: isolated word recognition, word segmentation, and word recognition by phones. It can operate for any discrete speech pattern – phone, syllable, word, or phrase, therefore we use the term "utterance" along with "word" to describe a recognition unit.

3.2. Word recognition

Speech input and processing. There are two speech input modes – by microphone or from the PCM format file. The input signal is processed using a mean subtractor and a first order high-pass filter.

Word endpoint detection. The next stage after speech signal processing is word endpoint detection. There were two endpoint detection methods implemented in the system. The first and simpliest one is the energy threshold method.

The second method is based on detection of abrupt changes in random sequences. After the signal analysis we have obtained a sequence of signal frame energy values. We can write

$$A(k) = \begin{cases} A_1 = N\left(\mu_1, \sigma_1^2\right), & \text{for } k = 1, 2, \ldots, u_1; \\ A_2 = N\left(\mu_2, \sigma_2^2\right), & \text{for } k = u_1 + 1, \ldots, u_2; \\ A_3 = N\left(\mu_3, \sigma_3^2\right), & \text{for } k = u_2 + 1, \ldots, K, \end{cases} \tag{1}$$

here A_1 and A_3 are statistical energy parameters of silence segments before and after the word, A_2 is set of word energy parameters, μ_i and σ_i are the frame energy mean and dispersion, u_1 and u_2 are the moments of abrupt parameter changes satisfying $1 < u_1 < u_2 < K$. The moments of change $\hat{u} = [\hat{u}_1, \hat{u}_2]$ corresponding to the beginning and end of a word we can find by maximizing likelihood function of the set of change moments

$$\hat{u} = \arg\max_u l(u|x). \tag{2}$$

In order to reduce the calculating load, we take a logarithm of likelihood and transform it to the sum of functions of one variable

$$\hat{u} = \arg\max_u \log l(u|x) = \arg\max_u \theta(u|x), \tag{3}$$

where

$$\theta(u|x) = l_1(u_1|x) + l_2(u_2|x). \tag{4}$$

The partial likelihood function $l_i(u_i|x)$ is calculated using the recursive expression

$$\begin{aligned} l_i(k|x) = l_i(k-1|x) - \log b(i) + \log b(i+1) - \\ \frac{1}{2b^2(i)}\left(\sum_{j=0}^{p} a_j(i)x(n-j)\right)^2 + \\ \frac{1}{2b^2(i+1)}\left(\sum_{j=0}^{p} a_j(i+1)x(n-j)\right)^2, \\ \text{for} \quad i=1,2; \quad k=2,3,\ldots,N, \end{aligned} \tag{5}$$

with zero initial conditions.

Since the function $\theta(u|x)$ is a sum of functions of one variable, it could be maximized using the dynamic programming (DP) principle. Bellman functions are defined to this end

$$g_1(u_2|x) = \max_{\substack{u_1 \\ p<u_1<u_2}} l_1(u_1|x), \tag{6}$$

$$g_2(u_3|x) = \max_{\substack{u_2 \\ p+1<u_2<u_3}} \left[l_2(u_2|x) + g_1(u_2|x)\right], \tag{7}$$

$$\text{for} \quad u_3 = p+3, p+4, \ldots, N.$$

Change moments (word endpoints) are determined as minimal arguments of peaks of Bellman functions

$$\hat{u}_k = \min\left[\arg \max_{\substack{n \\ p+k<n<\hat{u}_{k+1}}} g_k(n|x)\right], \tag{8}$$

$$\text{for} \quad k=2,1; \quad \hat{u}_3 = N.$$

At the start of calculations we must have initial parameter values. In real life, they are unknown. Therefore we apply the generalized expectation maximization algorithm in the likelihood function maximization. The first 11 and last 11 frames of a signal are taken as silence segments for parameter evaluation. The rest part of a signal is taken for word. Using the initial parameter values, we calculate the values of likelihood functions, Bellman functions, and determine new values for word boundaries. Calculations are repeated while word boundaries are varying. The algorithm of word endpoint detection is given in Figure 2.

Signal analysis. Two signal analysis methods are implemented in the system: linear prediction coding (LPC) and linear prediction coding cepstrum (LPCC) analyses. An autocorrelation method using the Levinson-Durbin algorithm is applied in the LPC analysis (of 10th order). The LPCC analysis order is variable in the range from 2 to 49. The LPCC analysis with mean subtraction was implemented additionally.

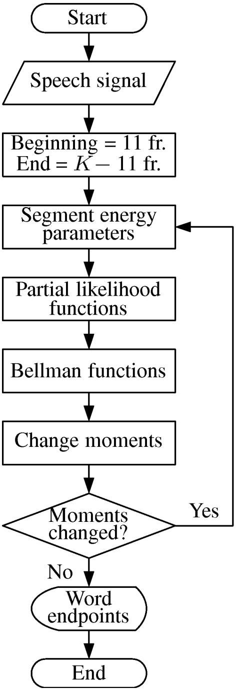

Fig 2. Word endpoint detection algorithm

References. Two reference creation methods are implemented: the direct and clustering-based training. The first one declares utterance as a reference without any additional analysis.

The clustering-based training procedure selects utterances from a set of candidates. The selection criterion is a minimal average distance between the selected utterances and the rest ones

$$\{i_P^m\} = \arg \min_{i \in \{I_P^m\},\ j \neq i} \left[\frac{1}{P-m} \sum_i \min\{D_{ij}\} \right], \qquad (9)$$

$$\text{for} \quad m = 1, 2, \ldots, M.$$

Here m is the number of references, P is the number of utterances-candidates, $\{i_P^m\}$ is the set of m references from P candidates, $\{I_P^m\}$ is the set of all possible reference permutations of m from P, $\{D_{ij}\}$ is the set of distances between all possible reference permutations and the rest ones (distances are calculated using dynamic time warping).

Firstly, all possible single reference variants are analyzed. Utterance having the minimal average distance to others is declared as a reference. Next, all the possible couples of references are analyzed, then all the possible variants of three references, etc. The maximal possible number of references is 5, the maximal number of utterances-candidates is 10.

Comparison. The stage of signal analysis results in a sequence of feature vectors. It must be compared with reference sequences in order to get a numerical assessment of mutual distances (similarities). The sequence of references yielding the smallest distance is treated as a recognized utterance.

Dynamic time warping based on dynamic programming was used to compare sequences. As usual, sequences are completely compared. The fast comparison mode is used. If a partially accumulated distance exceeds the predefined threshold value, the reference is rejected. Later experiments revealed rejection of about 80 % of references.

3.3. Utterance segmentation

The second mode of the KAS system is utterance segmentation.

Speech input, processing, and analysis. Speech input and processing are the same as word recognition mode. A signal is analyzed using the 10-th order LPC analysis.

Segmentation. Just like word endpoint detection, segmentation is also based on detection of abrupt change moments in random sequences. In this case, stationary segments of a signal match phones, and change points are phone boundaries. To this end, we consider a speech signal as a random sequence and describe it as follows

$$x(n) = -a_1(n) \cdot x(n-1) - \cdots - a_p \cdot x(n-p) + b(n) \cdot v(n). \tag{10}$$

The model of signal with M change points can be defined as

$$A(n) = \begin{cases} A_1, & \text{for } n = 1, 2, \ldots, u_1; \\ A_2, & \text{for } n = u_1 + 1, \ldots, u_2; \\ \ldots & \\ A_i, & \text{for } n = u_{i-1} + 1, \ldots, u_i; \\ \ldots & \\ A_M, & \text{for } n = u_{M-1} + 1, \ldots, u_M; \\ A_{M+1}, & \text{for } n = u_M + 1, \ldots, N. \end{cases} \tag{11}$$

Here A is the set of LPC model parameters, M is the number of change points,

$u = [u_1, u_2, \dots, u_M]$ is the set of parameter change moments, satisfying $p < u_1 < u_2 < \cdots < N$.

Two segmentation approaches with a different optimality function have been developed. The first one, called a likelihood function maximization approach, was built by finding maximum likelihood of a set of change points

$$\hat{u} = \arg\max_u l(u|x). \tag{12}$$

Similarly as in endpoint detection, we take a logarithm of the likelihood function and transform it to the sum of functions of one variable, i. e. partial likelihood functions. Again, the dynamic programming principle is applied and Bellman functions are calculated

$$\begin{aligned} &g_i(u_{i+1}|x) = \max\left[g_i(u_{i+1}-1|x), (g_{i-1}(u_{i+1}-1|x) + l_i(u_{i+1}-1|x))\right], \\ &\text{for } i = 1, 2, \dots, M;\ u_{i+1} = p+i+2, p+i+3, \dots, N \end{aligned} \tag{13}$$

under the initial condition

$$\begin{aligned} &g_i(p+i+1|x) = l_i(p+i|x) + g_{i-1}(p+1), \\ &\text{for } i = 1, 2, \dots, M. \end{aligned} \tag{14}$$

Now, change points (phone boundaries) can be found

$$\begin{aligned} &\hat{u}_i = \min\left[\arg \max_{\substack{k \\ p+i \leqslant k \leqslant \hat{u}_{i+1}}} g_i(k|x)\right], \\ &\text{for } i = M, M-1, \dots, 1;\ \hat{u}_{M+1} = N. \end{aligned} \tag{15}$$

In reality, the number of change points and model parameters is unknown. We use the expectation maximization approach again. The number of change points and initial model parameters are defined by the user – he marks the places of likely stationary segments. The number of change points is the number of selections minus 1 and the parameters are computed at the selected places from the analysis frame length segments.
The second developed segmentation approach is the minimal prediction error approach. In this case, the optimality criterion for change points is prediction error

$$\hat{u} = \arg\max_u E(x|u), \tag{16}$$

here $E(x|u)$ is a negative prediction error. This expression is resolved into the sum of partial prediction errors and the dynamic programming principle is applied, Bellman function values are calculated. Change moments are found using (15). The utterance segmentation algorithm is given in Figure 3.

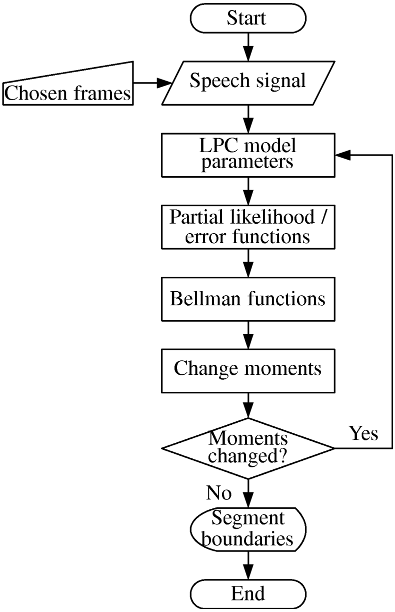

Fig 3. Utterance segmentation algorithm

3.4. Utterance phone recognition

Utterance segments (segments mostly correspond to phones, so hereafter we will call a segment as a phone) can be presented for recognition. By segmenting utterances before recognition we could organize the recognition of utterance phones. That is the idea of the third mode of the system KAS.

The number of phones is finite for any language. Hence, ideal utterance segmentation into phones would result in the number of references equal to the number of phones (or its multiple). In fact, a stable segmentation is hardly achievable, therefore the number of references will be lager than that of phones. But we still can expect it to be smaller than the number of phones in the training set. Beside, there is a theoretical possibility to recognize unknown utterances.

Secondly, a phone is a few times shorter than a word. It means that

- The analysis of one reference would be shorter.
- The memory required for reference storing would be smaller.

We have determined phones as stationary segments. We have followed the same assumption on phone recognition and have chosen the analysis frame of the phone length. This resulted in that

- The classification was reduced to the comparison of mutual vector distances.
- The amount of memory required for reference storing was decreased, every phone is represented by one feature vector.

The algorithm of utterance phone recognition is given in Figure 4.

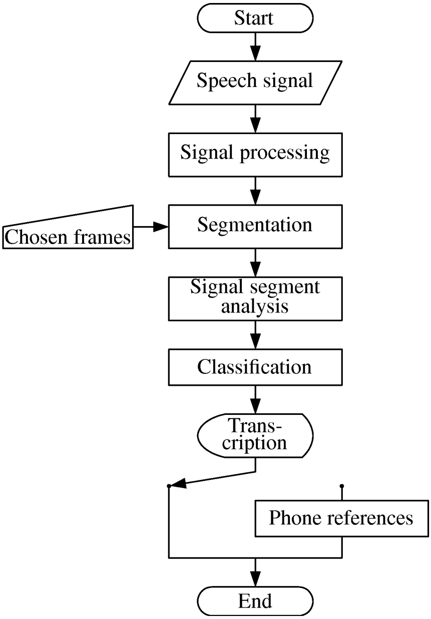

Fig 4. Utterance phone recognition algorithm

4. Experimental test of the system

The aim of experimental tests is actual evaluation of recognition and segmentation accuracy. The following processes have been analyzed: word endpoint detection efficiency, effect of word endpoint detection and system training on the recognition accuracy, segmentation efficiency and utterance phone recognition.

Experimental environment and data. All experiments were done by computer. Optimality of the system working parameters (analysis frame length and shift, analysis order, preemphasis coefficient, threshold values) was not the aim of experiments and they were established according to the author's practical experience. All utterances were recorded using various computers under different acoustic conditions. The vocabulary consisted of 111 words pronounced by 5 men and 5 women. The total number of utterances was 9102. The speakers were labelled with a letter (M – woman, V – man) and one-figure serial number (e. g. 1).

Analysis of word endpoint detection. The target of analysis – accuracy of word endpoint detection using both methods. The energy threshold method was used at the threshold value 0,05. The difference between detected by method and manually marked endpoint larger than 20 % of word length was registered as detection error. The graphical test results are presented in Figure 5 (DP stands for the endpoint detection method proposed by us, and ET for the energy threshold).

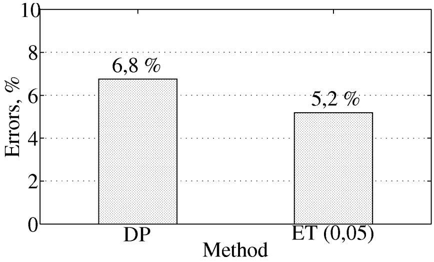

Fig 5. Accuracy of word endpoint detection

As we can see, the accuracy of the energy threshold method was a little bit higher, but the difference was insignificant. In general terms, results of both methods should be qualified as average. However, the DP method has a few advantages over the energy threshold. Firstly, there is no need for parameter adjustment. Secondly, the DP method is more robust to an increase of the signal-to-noise ratio (SNR) and only at high SNR values (20 dB and more) its accuracy is lower.

Word recognition analysis. The aim is to estimate effect of word endpoint detection and system training procedure on the isolated word recognition accuracy.

First of all, recognition was analyzed by different word endpoint detection methods. The recognition threshold for linear prediction coding (LPC) analysis was 0,7, while for linear prediction coding spectrum (LPCC) and LPCC with mean subtraction ($\overline{\text{LPCC}}$) analyses – 0,5. Recognition results are shown in Figure 6. We can see that recognition using automatic endpoint detection was a little bit more accurate despite its lower accuracy in the previous experiment.

When analyzing the training influence on the recognition accuracy, the case of three references per word was chosen. The recognition accuracy, using automatic endpoint detection and training procedure achieved 96–98 %. In comparison with a plain recognition the accuracy has increased from 10 % for LPCC to 15 % for LPC analysis. Furthermore, the accuracy of LPC analysis almost reached that of cepstral analysis. In a speaker-independent experiment an increase in accuracy was about 10 % for all the types of analysis.

Segmentation analysis. The objects of analysis are segmentation approaches. The main criterion for evaluating the segmentation result was the number of distinguished phones regardless of the accuracy of phone boundaries. If it was smaller than the actual number of phones, then it was treated as an error. Experiments were done with 888 utterances of 8 persons. The segmentation results for likelihood maximization (LM) and minimal prediction error (ME) approaches are shown in Figure 7.

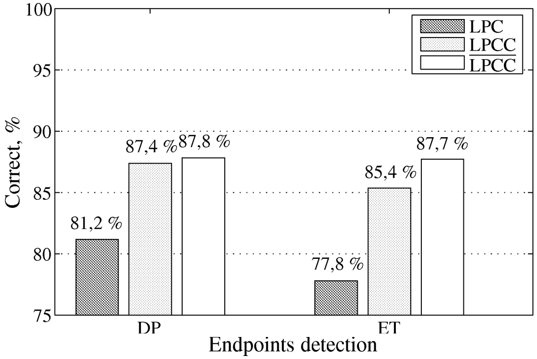

Fig 6. Word recognition accuracy with different endpoints detection methods

As we can see, the ME approach overtook the LM approach nearly by 23 %. Individual speaker results revealed the minimal prediction error approach to be more stable between different speakers (fluctuation of segmentation results was about 15 %) than the likelihood maximization approach (results fluctuated at about 55 %).

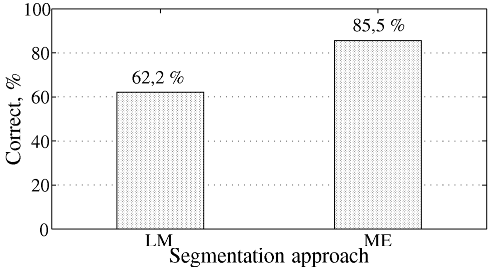

Fig 7. Utterance segmentation results

The number of iterations required for final solution was similar for both approaches: 3–6 iterations. A majority of ME approach errors was one phone error, the difference between the extracted and actual phone number was one. The experiment with segmentation dependance on the recording quality has demonstrated higher robustness of the minimal prediction error approach to the decrease of the signal-to-noise ratio.

Utterance phone recognition. The minimal prediction error segmentation approach was selected for the analysis of the utterance phone recognition. Experiments were carried out with one speaker's utterances (111 words). During training utterances were segmented and phones were labelled using the Lithuanian alphabet. References were created in a direct creation manner, regardless of segmentation correctness. The results of utterance phone recognition are presented in table.

Table. Utterance phone recognition results

Analysis	**Phone unrecognition errors per utterance, %**					
	0	**1**	**2**	**3**	**>3**	**All**
LPCC	5,4	21,6	28,8	15,4	25,2	3,6

As we can see, mostly one phone error occured. The level of complete recognition of the whole utterance was above 5 %. There were 2,5 substitution errors and nearly 0,1 deletion error per word on the average.

In order to decrease one phone error level we applied automatic grammar checking for recognition results. This increased the level of completely correct recognition up to 15,3 % and decreased the level of one phone errors down to 11,7 %.

5. Results and conclusions

Summarizing the results achieved and knowledge obtained the following conclusions are drawn:

1. The method of automatic detection of word endpoint was created. The error level of 6,8 % was a quarter higher than energy threshold error level. Despite this fact word recognition accuracy using automatic method was about 3,5 % higher than using energy threshold. Hence we state that stability is the most important requirement for the word endpoint detection. Stable and robust detection allows decrease the level of recognition errors.

2. The clustering principle was applied in system training. References have been created minimizing the average distance between utterances of the training set (distances are calculated using the dynamic time warping approach). The accuracy of the recognition system trained with 3 references per word (using clustering) was about 2,5 % higher than the acurracy of the system with 5 references per word (created directly). That implies that using selection procedures during training process gives 2–3 % higher recognition accuracy with smaller number of references per word.

3. Proposed word endpoint method and training procedure reduced the error level of the word recognition system. The absolute increase of the accuracy of speaker-dependent word recognition was 10–19 %, in case of speaker-independent recognition increase was 10–11 %. Thus accuracy of the recognition system can be improved optimizing separate stages of the recognition process.
4. The methodology of utterance segmentation was proposed. Two segmentation approaches were created using this methodology: likelihood function maximization and minimal prediction error approach. The likelihood function maximization operated with correct segmentation level of 62,2 %, and the minimal prediction error approach gave the level of 85,5 %. This suggests that phone boundaries in signal can be detected as change moments of linear prediction model parameters of the speech signal.
5. The idea of utterance phone recognition has been proposed. Recognition is performed in two stages: utterance is segmented into phones and then the extracted phones are recognized. The results of experiments were following: all phones of word were recognized correctly in 11,5 % of words, one phone error was made in 11,7 % of words. This way of recognition organization simplifies the comparison process and gives small recognition unit which allows to reduce the number of references.

Results revealed that isolated word recognition researches should be directed towards the improvement of the utterance phone recognition: choice of the recognition unit and signal analysis model, automation of the segmentation process, optimization of the similarity estimation, and application of linguistic processing.

List of publications on the topic of the dissertation

In the reviewed scientific periodical publications from the list approved of the Science Council of Lithuania

1. TAMULEVIČIUS, G.; LIPEIKA, A. Žodžio pradžios ir galo nustatymas atpažįstant atskirai sakomus žodžius. *Elektronika ir elektrotechnika*, 2005, Vol. 2, No. 58, p. 61–64. ISSN 1392-1215 (INSPEC, VINITI).
2. LIPEIKA, A.; TAMULEVIČIUS, G. Segmentation of words into phones. *Electronics and Electrical Engineering*, 2006, Vol. 1, No. 65, p. 11–15. ISSN 1392-1215 (INSPEC, VINITI).

In other editions

3. TAMULEVIČIUS, G.; LIPEIKA, A. Žodžių atpažinimo sistemos kūrimas. *Lietuvos matematikos rinkinys*, 2003, Vol. 43, p. 292–296. ISSN 0132-2818 (CIS, MatSciNet, VINITI, Zentralblatt MATH).

4. TAMULEVIČIUS, G.; LIPEIKA, A. Dynamic time warping based speech recognition system. In *Human Language Technologies. The Baltic Perspective*, 2004, p. 151–161.
5. TAMULEVIČIUS, G.; LIPEIKA, A. Segmentation of nonstationary signals. In *Biomedical Engineering*, 2004, p. 37–40. ISBN 9955-09-739-6.
6. TAMULEVIČIUS, G. Interneto naršyklės valdymas balsu. In *Informacinės technologijos 2007*, 2007, p. 67–70. ISSN 1822-6337.

About the author

Gintautas Tamulevičius was born in 1979.
1997–2001 Studies at the Electronics faculty of Vilnius Gediminas Technical University, bachelor's degree in Electronics and Electrical Engineering.
2001–2003 Studies at the Electronics faculty of Vilnius Gediminas Technical University, master's degree in Electronics and Electrical Engineering.
2003–2007 Doctoral studies at the Institute of Mathematics and Informatics.

Pavienių žodžių atpažinimo sistemų kūrimas

Mokslo problemos aktualumas. Atpažinimo sistemų kūrimas reikalauja didelių laiko ir žmogiškojo darbo resursų, kas sąlygoja didelę tokių produktų savikainą. Todėl komerciniai produktai, kuriuose realizuojamas kalbos atpažinimas, kuriami tik didelėms rinkoms, t. y. plačiai paplitusioms kalboms. Tuo tarpu kalbos, vartojamos nedidelėse srityse, lieka be dėmesio. Toks atsiribojimas lemia tų kalbų atpažinimo tyrimų nykimą – tyrimai apsiriboja didžiosioms kalboms sukurtų metodų ir technologijų pritaikymu ir modifikavimu.

Siekiant sumažinti atskirtį, padidinti praktinę kalbos atpažinimo reikšmę ir įtaką šiuolaikinėms technologijoms, būtina kurti metodines ir technines lietuvių kalbos atpažinimo tyrimų priemones, plėsti atpažinimo klausimų tyrimus, ieškoti ir formuluoti naujas tyrimų kryptis, bandyti pritaikyti atpažinimo sprendimus praktiniams uždaviniams.

Darbo tikslas ir uždaviniai – pasiūlyti sprendimus, kurie leistų padidinti pavienių žodžių atpažinimo sistemos tikslumą bei efektyvumą nemodifikuojant naudojamo atpažinimo ir signalo analizės metodų. Siekiant tikslo buvo sprendžiami šie uždaviniai:

1. Pasiūlyti sprendimą žodžio ribų nustatymo stabilumui ir atsparumui triukšmui didinti.
2. Pasiūlyti žodžių etalonų sudarymo metodą, didinantį atpažinimo tikslumą ir efektyvumą.

3. Pasiūlyti žodžių segmentavimo į garsus metodą. Išnagrinėti žodžio garsų atpažinimo galimybę.
4. Realizuoti pasiūlytuosius metodus. Eksperimentiškai įvertinti žodžių ribų nustatymo ir etalonų kūrimo įtaką atpažinimo tikslumui, segmentavimo metodų tikslumą.

Mokslinis naujumas. Disertacijoje pasiūlyta keletas sprendimų, didinančių pavienių žodžių atpažinimo tikslumą ir efektyvumą. Sukurtas automatinio žodžio ribų nustatymo metodas. Metodas pasižymi atsparumu triukšmui, didesniu nei energijos slenksčio metodas, ir leidžia sumažinti atpažinimo klaidų, kylančių dėl klaidingų žodžio ribų, kiekį. Etalonams kurti pasiūlytas klasterizavimas, minimizuojantis vidutinį atstumą iki klasterių centrų. Sprendimo išskirtinumas – atstumai skaičiuojami naudojant dinaminį laiko skalės kraipymą. Toks sistemos apmokymas (etalonų kūrimas) leidžia padidinti atpažinimo tikslumą su mažesniu etalonų kiekiu.

Sukurti du žodžių segmentavimo į garsus metodai, grindžiami tikėtinumo funkcijos maksimizavimu ir prognozės klaidos minimizavimu. Abiejuose metoduose garsų ribos aptinkamos kaip kalbos signalo modelio parametrų pasikeitimo momentai. Panaudojus mažiausių kvadratų metodą suformuluota ir realizuota žodžių atpažinimo garsais idėja. Žodis atpažįstamas 2 etapais: segmentuojamas į garsus, pastaruosius atpažįstant. Toks atpažinimas leido supaprastinti palyginimo procesą ir sumažinti reikalingų etalonų kiekį. Be to, idėja leidžia formuoti tolimesnio darbo kryptis: tobulinti segmentavimo metodą, taikyti alternatyvius kalbos signalo analizės ir klasifikavimo metodus.

Tyrimų metodika apima matematinės analizės, skaitmeninio signalų apdorojimo, atpažinimo teorijos žinias. Atpažinimo sistema realizuota C++ kalba, naudojant *Microsoft Visual Studio 6.0* programavimo aplinką.

Praktinė vertė. Disertacijoje sukurtieji žodžio ribų nustatymo, klasterizavimu pagrįstas etalonų kūrimo metodas ir segmentacijos metodai buvo panaudoti pavienių žodžių ir frazių atpažinimo sistemoje *Atpazinimas*, žodžių segmentavimo sistemoje *Segmentacija* ir interneto puslapių atidarymo ir programų paleidimo sistemoje. Žodžių atpažinimo ir segmentavimo sistemos įtrauktos į 2000–2006 m. programos „Lietuvių kalba informacinėje visuomenėje“ automatinio lietuvių šnekos atpažinimo tiriamuosius darbus.

Ginamieji teiginiai

1. Automatinis žodžio ribų nustatymo metodas leidžia sumažinti atpažinimo klaidų, kylančių dėl neteisingai nustatytų ribų, lygį.
2. Klasterizavimu pagrįstas etalonų kūrimas leidžia padidinti atpažinimo sistemos tikslumą su mažesniu etalonų skaičiumi.

3. Sukurtieji kalbos signalo segmentavimo metodai leidžia žodžių garsų ribas signale aptikti kaip signalo tiesinės prognozės modelio parametrų pasikeitimo momentus.
4. Atpažinimo sistemos eksperimentinio tyrimo rezultatai patvirtina pasiūlytųjų sprendimų ir metodų efektyvumą.

Darbo apimtis. Darbą sudaro 6 skyriai, literatūros ir autoriaus publikacijų sąrašai bei vienas priedas. Disertacijos aiškinamąjį raštą sudaro 124 teksto puslapiai su 35 iliustracijomis ir 14 lentelių.

Pirmajame skyriuje pristatoma darbo tema, darbo tikslas ir uždaviniai, ginamieji teiginiai bei darbo mokslinis naujumas.

Antrajame skyriuje nagrinėjami kalbos atpažinimo klausimai ir problemos. Apžvelgiama sistemų raida, darbai užsienyje ir Lietuvoje.

Trečiajame skyriuje nagrinėjami kalbos atpažinimo sistemų elementai – signalo analizės metodai, klasifikacijos metodai, jų privalumai ir trūkumai.

Ketvirtajame skyriuje suformuluoti žodžio ribų nustatymo ir žodžių segmentavimo metodai. Etalonams kurti pritaikytas klasterizavimo principas. Suformuluota žodžių atpažinimo garsais idėja.

Penktajame skyriuje pateikti sukurtosios pavienių žodžių sistemos eksperimentinio tyrimo rezultatai. Tirta pasiūlytųjų žodžio ribų nustatymo, segmentavimo metodų darbingumas, mokymo įtaka atpažinimo tikslumui. Atliktas preliminarus žodžių atpažinimo garsais tyrimas.

Šeštajame skyriuje apibendrinami darbo rezultatai, suformuluojamos darbo išvados ir įvardijami ateities darbai vystant žodžių atpažinimą garsais.

Priede pateikiamas sistemai tirti naudotas žodynas.

Bendrosios išvados

Atlikę pavienių žodžių atpažinimo, naudojant dinaminio laiko skalės kraipymo metodą, tyrimą, suformulavę metodo trūkumus ir pasiūlę sprendimus jiems pašalinti, gautus darbo rezultatus apibendriname:

1. Sukurtas automatinis žodžio ribų nustatymo metodas, pasižymintis stabilumu bei atsparumu signalo kokybės kitimui. Eksperimentų metu gautas 6,8 % klaidų lygis buvo maždaug ketvirtadaliu didesnis nei energijos slenksčio metodo. Tačiau atpažinimo klaidų lygis naudojant pasiūlytąjį metodą buvo iki 3,5 % mažesnis nei energijos slenksčio atveju. Todėl teigiame, kad svarbiausia žodžio ribų nustatymo savybė – stabilumas. O stabilus ir triukšmams atsparus žodžių ribų nustatymas leidžia sumažinti atpažinimo klaidų lygį.
2. Etalonams kurti pasiūlytas klasterizavimo principas, minimizuojantis vidutinį atstumą iki klasterių centrų (atstumas skaičiuojamas naudojant dinaminį laiko skalės kraipymą). Sistemos, apmokytos 3 etalonais kiekvienam žodyno žodžiui (naudojant klasterizacijos principą), atpažinimo tikslumas buvo

vidutiniškai 2,5 % didesnis už sistemos su 5 kiekvieno žodžio etalonais (etalonus kuriant tiesiogiai, be jokios atrankos procedūros). Taigi papildomų atrankos procedūrų naudojimas mokyme 2–3 % procentais padidina sistemos atpažinimo tikslumą su mažesniu etalonų skaičiumi.

3. Įdiegus pasiūlytuosius žodžio ribų nustatymo ir etalonų kūrimo metodus nuo kalbėtojo priklausomo pavienių žodžių atpažinimo tikslumo absoliutus padidėjimas buvo 10–19 %, nepriklausomo nuo kalbėtojo atpažinimo – 10–11 %. Taigi kalbos atpažinimo sistemos tikslumas gali būti padidintas ne modifikuojant atpažinimą metodą, o optimizuojant atskirus atpažinimo etapus.
4. Pasiūlyta žodžio garsų ribų nustatymo metodika, kuria remiantis sukurti du metodai žodžiams segmentuoti – tikėtinumo funkcijos maksimizavimo ir prognozės klaidos minimizavimo. Eksperimentų metu tikėtinumo maksimizavimo metodo segmentavimo tikslumas siekė 62,2 %, prognozės klaidos minimizavimo metodo – 85,5 %. Taigi garsų ribos signale gali būti ieškomos kaip kalbos signalo tiesinės prognozės modelio parametrų pasikeitimo momentai.
5. Suformuluota žodžių atpažinimo garsais idėja. Šiuo atveju atpažinimo procesas vykdomas dviem etapais – žodis segmentuojamas į garsus, pastaruosius bandant atpažinti. Eksperimentų metu visi žodžio garsai teisingai atpažinti 15,3 % žodžių, suklysta vienu garsu – 11,7 % žodžių. Pagrindinis tokio atpažinimo organizavimo privalumas – elementarus dviejų pavyzdžių palyginimas ir smulkus atpažinimo vienetas, leidžiantis sumažinti dideliam žodynui reikalingų etalonų kiekį.

Disertacijos darbo rezultatai parodė, kad tolimesni pavienių žodžių atpažinimo tyrimai turėtų būti nukreipti žodžių atpažinimo garsais idėjai tobulinti – parinkti optimalų atpažinimo vienetą, signalo analizės metodą, automatizuoti segmentavimo procesą, optimizuoti panašumo įvertinimo procedūrą lygiagrečiai taikant lingvistinį apdorojimą.

Trumpos žinios apie autorių

Gintautas Tamulevičius gimė 1979 metais.
1997–2001 m. įgijo elektronikos ir elektros inžinerijos bakalauro laipsnį Vilniaus Gedimino technikos universiteto Elektronikos fakultete.
2001–2003 m. įgijo elektronikos ir elektros inžinerijos magistro laipsnį Vilniaus Gedimino technikos universiteto Elektronikos fakultete.
2003–2007 m. studijavo Informatikos inžinerijos doktorantūroje Matematikos ir informatikos institute.

Gintautas Tamulevičius

DEVELOPMENT OF ISOLATED WORD RECOGNITION SYSTEMS

Summary of Doctoral Dissertation
Technological Sciences, Informatics Engineering (07T)

PAVIENIŲ ŽODŽIŲ ATPAŽINIMO SISTEMŲ KŪRIMAS

Daktaro disertacijos santrauka
Technologijos mokslai, Informatikos inžinerija (07T)

2008 05 10. 1,5 sp. l. Tiražas 100 egz.
Vilniaus Gedimino technikos universiteto
leidykla „Technika“, Saulėtekio al. 11, 10223 Vilnius
http://leidykla.vgtu.lt
Spausdino UAB „Baltijos kopija“,
Kareivių g. 13B, 09109 Vilnius
http://www.kopija.lt