VILNIUS GEDIMINAS TECHNICAL UNIVERSITY
INSTITUTE OF MATHEMATICS AND INFORMATICS

Juozas KAMARAUSKAS

SPEAKER RECOGNITION BY VOICE

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

Vilnius LEIDYKLA TECHNIKA 2009

Doctoral dissertation was prepared at the Institute of Mathematics and Informatics in 2004–2009.

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 Gintautas DZEMYDA (Institute of Mathematics and Informatics, Technological Sciences, Informatics Engineering – 07T).

Members:

Assoc Prof Dr Algirdas BASTYS (Vilnius University, Physical Sciences, Informatics – 09P),

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

Prof Dr Habil Rimantas ŠEINAUSKAS (Kaunas University of Technology, Technological Sciences, Informatics Engineering – 07T),

Prof Dr Habil Laimutis TELKSNYS (Institute of Mathematics and Informatics, Technological Sciences, Informatics Engineering – 07T).

Opponents:

Prof Dr Dalius NAVAKAUSKAS (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – 07T),

Dr Algimantas Aleksandras RUDŽIONIS (Kaunas University of Technology, Technological Sciences, Informatics Engineering – 07T).

The dissertation will be defended at the public meeting of the Council of Scientific Field of Informatics Engineering at the Institute of Mathematics and Informatics, Room 203, at 2 p. m. on 27 May 2009.

Address: Akademijos g. 4, LT-08663 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 24 April 2009.

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, Lithuania).

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS INSTITUTAS

Juozas KAMARAUSKAS

ASMENS ATPAŽINIMAS PAGAL BALSĄ

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

Vilnius LEIDYKLA TECHNIKA 2009

Disertacija rengta 2004–2009 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. Gintautas DZEMYDA (Matematikos ir informatikos institutas, technologijos mokslai, informatikos inžinerija – 07T).
Nariai:
doc. dr. Algirdas BASTYS (Vilniaus universitetas, fiziniai mokslai, informatika – 09P),
prof. habil. dr. Romualdas BAUŠYS (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07T),
prof. habil. dr. Rimantas ŠEINAUSKAS (Kauno technologijos universitetas, technologijos mokslai, informatikos inžinerija – 07T),
prof. habil. dr. Laimutis TELKSNYS (Matematikos ir informatikos institutas, technologijos mokslai, informatikos inžinerija – 07T).
Oponentai:
prof. dr. Dalius NAVAKAUSKAS (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07T),
dr. Algimantas Aleksandras RUDŽIONIS (Kauno technologijos universitetas, technologijos mokslai, informatikos inžinerija – 07T).

Disertacija bus ginama viešame Informatikos inžinerijos mokslo krypties tarybos posėdyje 2009 m. gegužės 27 d. 14 val. Matematikos ir informatikos institute, 203 auditorijoje.
Adresas: Akademijos g. 4, LT-08663 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 2009 m. balandžio 24 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“ 1611-M mokslo literatūros knyga.

General characteristic of the dissertation

Relevance of the problem. Problems of speaker recognition become more and more relevant all over the world. These problems arise in criminology, information protection; it can be used in entrance control systems, mobile banking and e-commerce. A big attention is paid to speaker's recognition all over the world, both intellectual and material resources are allocated, various testing centres have been established. If other kinds of biometrics need special expensive equipment, voice biometrics does not need it.

In spite of great achievements in speaker recognition technology there is no theory created on how does human separate one voice from the other and there is no system of features created that would let separate two voices having different phrases, speaking environment, sound recording channels and so on.

Voice biometrics gives worse results compared to other kinds of biometrics but it could be widely used. Therefore investigations in that field should be made.

Aim of the work – to perform analysis of speaker recognition systems and to propose solutions to increase the accuracy and efficiency of speaker recognition system.

Tasks of the work. In pursuance with this aim the following issues were dealt with:

1. Algorithm of automatic speech activity detection.
2. New and effective system of features that should increase recognition accuracy and reduce amount of calculation.
3. Method of calculation of initial parameters when creating speakers models.
4. Experimental results of proposed methods and compare it with baseline methods.

Scientific novelty

- Automatic method of speech activity detection that is fast and does not require any additional actions from the user.
- System of features that combines vocal tract parameters and excitation parameters. As excitation (source) parameter pitch was used. Four formants and three antiformants were used as vocal tract parameters.
- Method of evaluation of initial GMM parameters. They are calculated using modified LBG Vector Quantization method.

Methodology of research includes mathematical analysis, probability theory and statistics, digital signal processing and pattern recognition theory. The speaker recognition system was built using *Borland* development environment *Turbo C++ 2006.*

Defended propositions

- Proposed system of features that consist of excitation source and vocal tract parameters.
- Proposed speech activity detection (SAD) method.
- Proposed method for estimation of initial GMM parameters.
- Created software for automatic speaker recognition.

The scope of the scientific work. The scientific work consists of an introduction, four chapters, conclusions, references, list of publications. The total scope of the dissertation – 124 pages, 58 pictures, 8 tables. The dissertation is written in Lithuanian.

1. Automatic speaker recognition systems

Abstraction of automatic speaker recognition system is shown in Fig. 1.

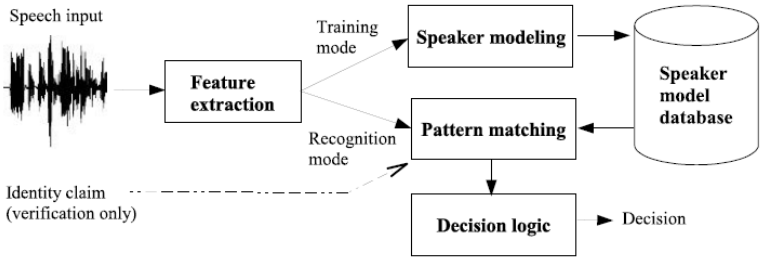

Fig. 1. Structure of automatic speaker recognition system

This system operates in two different modes: *training* and *recognition*. In *training mode* new speaker is enrolled to the system. Model of the speaker is created and stored in system‘s database. In *recognition mode* unknown speaker gives speech input and system makes the decision about speaker’s identity. In both modes *feature extraction* is performed first. Feature extraction converts speech signal into some numerical descriptors, so called feature vectors, that represent speaker‘s individuality. During training phrase speakers model is created from the feature vectors. There are lots of methods of speaker modeling.

In the recognition phase feature vectors are calculated from the unknown speaker's voice sample. After that in the pattern matching similarity score is calculated between unknown speaker's speech vectors and models stored in the database. The last step is decision making. Decision module makes decision about speakers' identity according to similarity scores.

2. Analysis of the speaker recognition system

The Gaussian mixture model (GMM) approach was used for speaker modeling and pattern matching in our recognition system. The choise have been made with notion that a linear combination of Gaussian basis functions is capable of representing a large class of sample distributions. Distribution of components of feature vectors cannot be precisely approximated with functions of simple standard distribution. Also this statistical method is text-independent.

The main drawback of GMM method is big amount of calculations especially in parameter estimation procedure using standard expectation – maximization (EM) algorithm.

3. Implementation of the recognition system

3.1. Main tasks in design of automatic speaker recognition systems

There are three main tasks that must be solved in designing of automatic speaker recognition system:

- Voice activity detection algorithm.
- Design of system of features.
- Speaker modeling and pattern matching algorithm.

3.2. Voice activity detection algorithm

Voice activity detection (VAD) is very important stage in speech/speaker recognition process. „Speech detector" is used in speech/speaker recognition systems and it's task is to find frames of the signal that corresponds to the speech and separate it out of noise for further processing. Feature vectors should be calculated from the signal frames that correspond to the speech. „Speech detectors" differ according to features and classification method they are using.

Simple traditional methods like energy threshold or zero crossing rate do not provide desired results especially in the case of bad recording conditions. Complicated methods, like using HMM and so on are not fast, besides often

they are not fully automated and require patterns of speech and noise for system training.

We propose fast and fully automated algorithm of voice activity detection. Algorithm of proposed method is shown in Fig. 2.

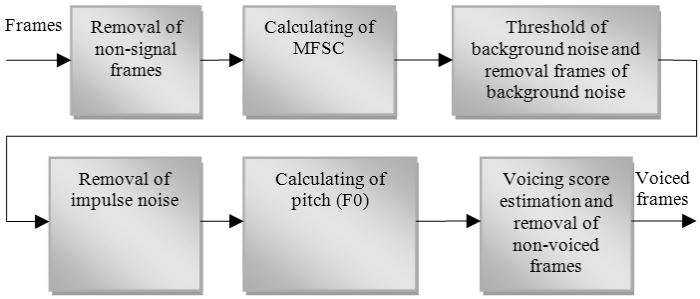

Fig. 2. Voice activity detection algorithm

The first step is *removal of non-signal frames*. Sometimes in digital recordings there are parts of the signal, where zero values are written or there is quantization noise of analog/digital converters. These parts should not be analized because there is no background noise or speech signal. Maximum of signal amplitude is calculated in the frame and compared with threshold, equal to 130. If this value is less than threshold, frame is eliminated from further calculations.

Second step is calculation of mel-frequency spectrum (MFSC). The Fast Fourier transform of the signal frame is calculated first. Then MFSC is calculated using triangular overlapping filters, formed by mel-frequency scale. Number of filters is 33.

$$E(m,i) = \sum_{k=1}^{512} \left| X_F(m,k) \right| \mathrm{H}(i,k), \tag{1}$$

where $X_F(m,k)$ – Fourier transform of m-th frame, $1 \le i \le 33$, $1 \le m \le M-1$, M – count of frames, $\mathrm{H}(i,k)$ – function of triangular filters.

Third step is calculating threshold of background noise and removal frames of background noise. Average energy of MFSC for every frame m is calculated first:

$$E_{av}(m) = \frac{1}{33} \sum_{i=1}^{33} E(m,i) \cdot \tag{2}$$

Then 10 frames with minimal values of E_{av} are found. These frames correspond to background noise. Then mean value of 10 frames is calculated for every component of mel-frequency spectrum:

$$E_n(i) = \frac{\sum_{m=1}^{10} E(m,i)}{10} \cdot \tag{3}$$

And threshold for the background noise can be expressed:

$$Thr = 2 \cdot \frac{1}{33} \sum_{i=1}^{33} E_n(i) \cdot \tag{4}$$

Then the average energy of MFSC of every frame m is compared against threshold Thr. If it is less than threshold, frame is considered as background noise and is not used in further calculations.

Next step is *removal of impulse noise*. Sounds that are shorter than 15 ms are removed and do not used in further calculations.

Next step is calculation of pitch. Frequency – domain method is used there. Every frame is multiplied by Hamming window, then the filtering using band-pass filter is applied. Frequency range of filter is 60–3 300 Hz. Then LPC analysis of 8-th order is performed and inverse filtering using LPC parameters is applied. After that we get excitation signal. This signal is filtered with 32 order low – pass filter with cut-off frequency at 2 000 Hz. Then Fast Fourier transform is applied to the filtered excitation signal and we get spectrum of this signal. Correlation function of the spectrum is calculated. The distance between two peaks of the correlation function corresponds to the pitch.

The last step is voicing score estimation. If pitch value is in the range 60–500 Hz, frame is considered as voiced and is used for further calculations in the feature extraction.

Fig. 3 shows operation of proposed voice activity detection algorithm is shown. Signalogramm and segmented parts of the signal after removal of non-signal frames, background noise and impulse noise are shown above. These parts are marked with rectangles. Pitch contours are shown below. Parts of the signal where pitch value is equal to 0, were segmented using algorithm, mentioned above, but there values of the pitch were not found, so these frames of signal are discarded too in feature extraction phase.

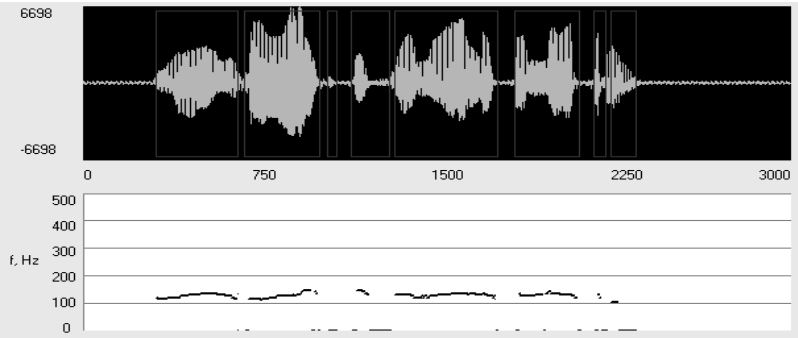

Fig. 3. Illustration of the segmentation algorithm

3.3. Design of system of features

Feature extraction is very important phase in speaker‘s recognition. There are lots of features, that are used in speech/speaker recognition systems: LPC parameters, LPC cepstrum (LPCC), mel-frequency cepstrum (MFCC), formants and so on. We realized two systems of features in our speaker recognition system:

- Standard MFCC (baseline in speaker recognition).
- Proposed system of features: four formants, three antiformants and pitch.

MFCC are calculated in standard way. Hamming window is applied to the frame of signal. Then spectrum of the frame is calculated using FFT. Size of FFT – 512 points. Filter bank of overlapping triangular filters, allocatel by mel frequency is formed and Fourier spectrum in frequency domain is multiplied by these filters. Thus we get mel-frequency spectrum coefficients (MFSC). To get MFCC we apply discrete cosine transform to the MFSC. We used 25 overlapped triangular filters and order of MFCC is 13 (these parameters can be changed).

If we look at the Fourier spectrum of the signal frame we will see there some peaks, that are called formants. In frequency range of 200–5 000 Hz we can see 3–5 maximas. Each formant corresponds to a resonance in the vocal tract.

Positions of the formants are well seen if we look at transfer function of the vocal tract. We can calculate transfer function from the LPC parameters, that corresponds to the vocal tract.

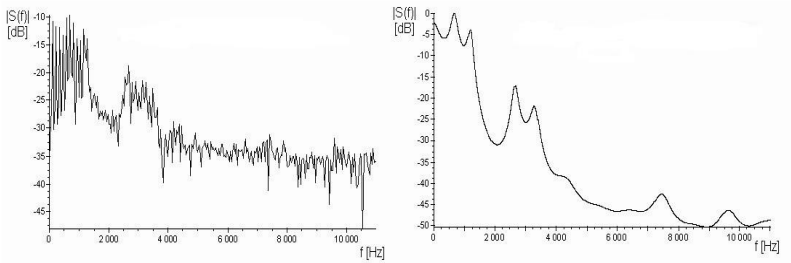

Fig. 4. Fourier transform of signal frame and transfer function calculated from the LPC parameters

In the left side of Fig. 4 Fourier transform of the signal frame of the vowel A is shown. In the right side transfer function calculated from the LPC parameters of this frame is shown, where positions of the formants are seen visibly.

Calculation of the formants is a trivial task. This is because maximas of the spectrum disappear in certain conditions and their calculation from the envelope of the spectrum becomes imposible. Method of the line spectral pairs was used for this purpose.

In the Fig. 5 signalogramm, spectrogramm and line spectral pairs of the phonemes A E and I are shown.

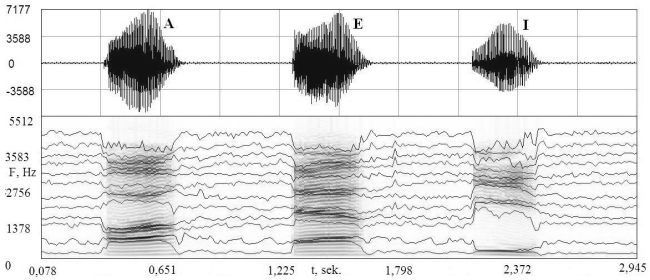

Fig. 5. Signalogramm spectrogramm and line spectral pairs

As we can see in Fig. 5, spectral pairs enshroud formants, so frequency of spectral pair can be assigned to corresponding formant. Lets denote frequency of *N*-th spectral pair as *LSF*(*N*), frequency of *M*-th formant as *F*(*M*), frequency of *K*-th antiformant as *ANF*(*K*). We used such evaluation of formants and antiformants:

$$\begin{cases} F(1) = LSF(2); \\ F(2) = LSF(5); \\ F(3) = LSF(8); \\ F(4) = LSF(11); \\ ANF(1) = (LSF(2)+LSF(3))/2; \\ ANF(2) = (LSF(5)+LSF(6))/2; \\ ANF(3) = (LSF(8)+LSF(9))/2. \end{cases} \quad (5)$$

Dispersion of higher formants and antiformants is bigger, so to "equalize" dispersions of all formants and antiformants, we calculate them in mel-frequency scale.

Our proposed system of features consists of four formants, three antiformants and pitch value.

3.4 Speaker modeling and pattern matching

The Gaussian Mixture Models (GMM) approach was used for speaker modeling and pattern matching. One of the problems is calculating of initial parameters of GMM. There are no good theoretical ideas to solve this problem.

Parameter initialization in recognition system is performed dividing space of feature vectors into the non-overlapped clusters. Statistical parameters of the clusters like mean and standard deviation are assigned as initial parameters of GMM. Mixture weight is calculated as ratio between count of feature vectors in the corresponding cluster with count of all feature vectors. Count of clusters is equal to count of components of GMM.

Assuming T_i – size of i-th cluster (count of feature vectors in this cluster), T – count of all feature vectors, $\vec{x}_{i,t}$ – t-th feature vector in cluster i. Then initial parameters of i-th component of GMM can be expressed:

$$\vec{\mu}_i = \frac{1}{T_i}\sum_{t=1}^{T_i}\vec{x}_{i,t}, \quad (6)$$

$$\vec{\sigma}_i = \sqrt{\frac{1}{T_i}\sum_{t=1}^{T_i}(\vec{x}_{i,t} - \vec{\mu}_i)^2}, \quad (7)$$

$$p_i = \frac{T_i}{T}. \quad (8)$$

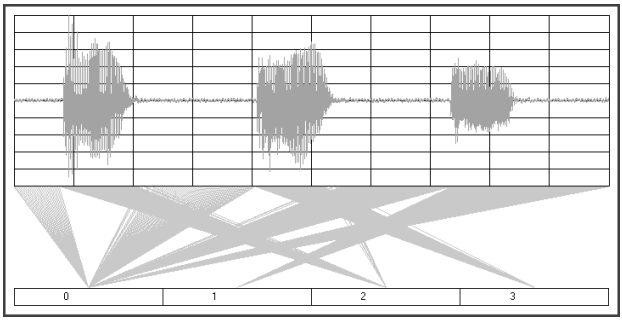

Fig. 6. Clusterization process

We propose to use modified vector quantization (VQ) approach for GMM parameters initialization. This approach is similar to standard LBG method, difference is that count of clusters (centroids) increases by 1 in every iteration "dividing" centroid with maximum distortion. Initial space of feature vectors is dividing in clusters where feature vectors of similar parts of signal are grouped. Fig. 6 shows clusterization process.

There are three phonemes recorded – A E and I. Count of clusters is equal to 4. As seen in Fig. 6, after clusterization, frames of noise are assigned to the 0 cluster. Most of frames of phoneme A are assigned to the 2-nd cluster. Frames of phoneme E are assigned to the 3-rd cluster and frames of phoneme I are assigned to the 1-st cluster. So, vector quantization groups similar parts of the signal and every component of feature vector of similar sound can be approximated by individual Gaussian distribution.

Parameter estimates are obtained iteratively using a special case of the expectation-maximization (EM) algorithm.

4. Experimental test of the system

Purpose of experimental test is to evaluate accuracy of speaker recognition using proposed system of features and compare it with baseline methods. Experimental tests were performed using various combinations of features:

- Using only excitation source parameters (pitch).
- Using only vocal tract parameters: four formants.
- Using vocal tract parameters: four formants and three antiformants.
- Combining vocal tract and excitation source parameters: four formants and pitch.

- Combining vocal tract and excitation source parameters: four formants, three antiformants and pitch.

Experiments were performed using standard features: MFCC. Recognition accuracy dependence on count of GMM components was investigated too.

Experimental conditions and data. All experiments were done by personal computer. System working parameters (analysis frame length and shift, order of analysis and so on) were set according to authors experience and were not changed.

Experiments were performed using Speech Technology Center (STC) database, recorded in 1996–1998. The main purpose of the database is to investigate individual speaker variability and to validate speaker recognition algorithms. The database was recorded through a 16-bit Creative Labs sound card with 11 025 Hz sampling rate. This database is distributed by ELRA.

The same phrases of 41 male speakers, pronounced approximately 15 times were taken. 3 recordings were used to build speaker's models, others for recognition.

Speaker recognition using excitation source parameters. Experiments of speaker recognition accuracy were done using as features pitch only and with different count of GMM components. In the Fig. 7 FAR – FRR curves are shown, when 10 GMM components were used.
EER value 24.6% was obtained in that case.

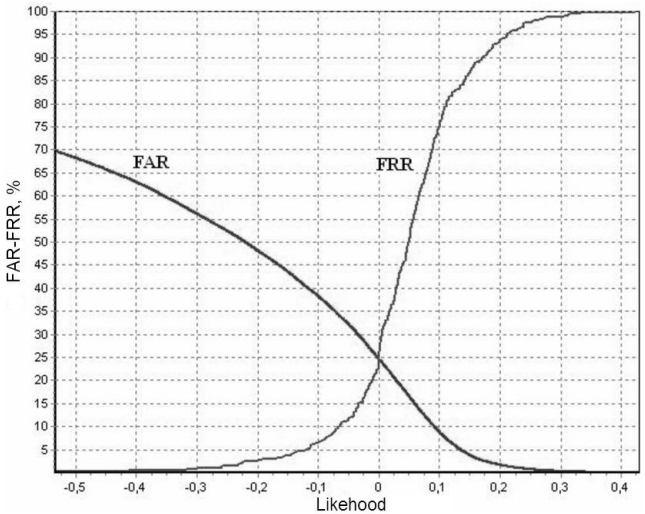

Fig. 7. FAR-FRR curves using pitch as features and 10 GMM components

Speaker recognition using vocal tract parameters: four formants. Experiments of speaker recognition using four formants as features and different count of components of GMM were performed. In the Fig. 8 below FAR – FRR curves are shown, when 15 components of GMM were used. EER value 12.26% was obtained in that case.

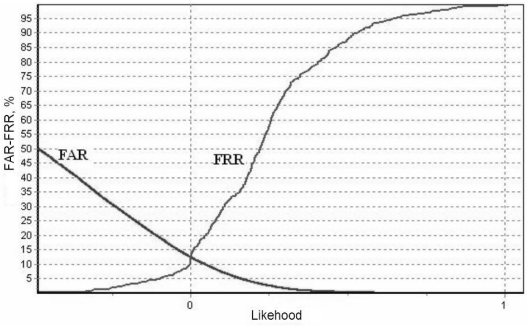

Fig. 8. FAR-FRR curves using four formants as features and 15 GMM components

Speaker recognition using vocal tract parameters: four formants and three antiformants. Experiments of speaker recognition using four formants and three antiformants as a features and different count of components of GMM were performed. In the Fig. 9 below FAR – FRR curves are shown, when 20 components of GMM were used. EER value 7.62% was obtained in that case.

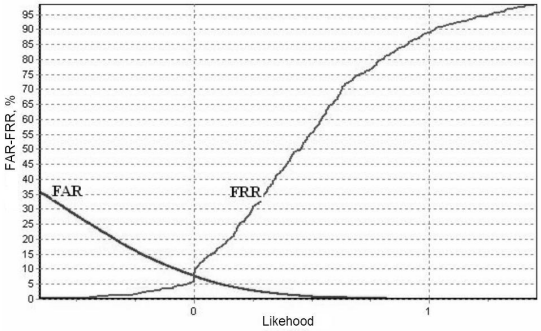

Fig. 9. FAR-FRR curves using four formants and three antiformants as features and 20 GMM components

Speaker recognition using four formants and pitch. Experiments of speaker recognition using four formants with pitch as features and different

count of components of GMM were performed. FAR-FRR curves, when 20 components of GMM were used are shown in Fig. 10. EER value 8.17% was obtained in that case.

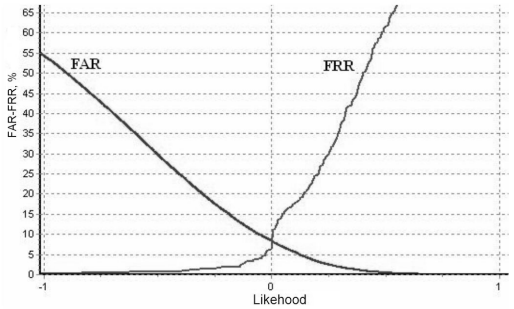

Fig. 10. FAR-FRR curves using four formants and pitch as features and 20 GMM components

Speaker recognition using four formants three antiformants and pitch. Experiments of speaker recognition using four formants three antiformants with pitch as a features and different count of components of GMM were performed. FAR-FRR curves, when 20 components of GMM were used are shown in Fig. 11. EER value 5.17% was obtained in that case.

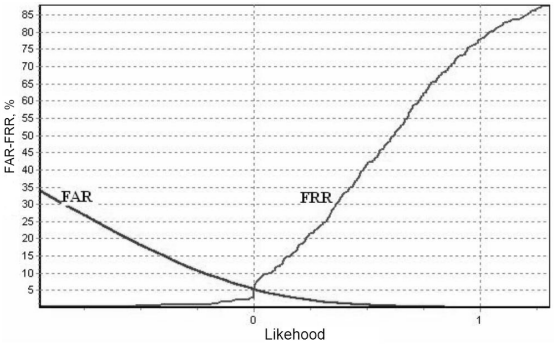

Fig. 11. FAR-FRR curves using four formants three antiformants and pitch as features and 20 GMM components

Speaker recognition using MFCC. Experiments of speaker recognition using standard MFCC as a features and different count of components of GMM

were performed. FAR-FRR curves, when 20 components of GMM were used are shown in Fig. 12. EER value 5.86% was obtained in that case.

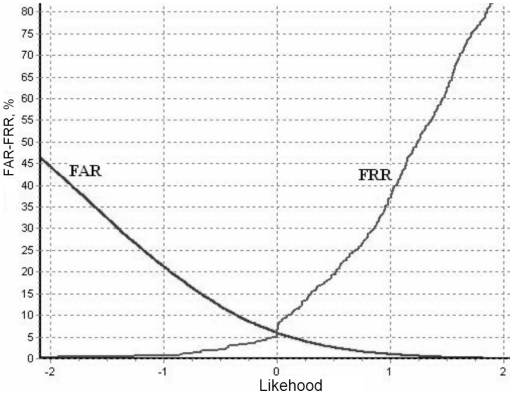

Fig. 12. FAR-FRR curves using standard MFCC as features and 20 GMM components

Summary of experimental results. Experimental results of speaker recognition using different features and different count of GMM components are shown in Table, where EER value is given. There F0 denoted system of features, that consist of pitch, 4F – system of features, that consist of 4 formants, 4F3A – system of features, that consist of 4 formants and 3 antiformants, 4F3AF0 – system of features, that consist of 4 formants 3 antiformants and pitch.

Summary of experimental results of speaker recognition

Count of GMM components	F0	4F	4F3A	4FF0	4F3AF0	MFCC
1	26.65%	-	-	-	-	-
3	24.62%	-	-	-	-	-
5	24.71%	13.8%	10.35%	9.22%	7.44%	7.35%
10	24.6%	12.59%	8.33%	8.25%	5.94%	6.27%
15	-	12.26%	7.99%	8.2%	**5.45%**	**6.15%**
20	-	12.4%	7.62%	8.17%	**5.17%**	**5.86%**
Adaptive	-	12.13%	8.01%	8.13%	**5.66%**	**5.89%**

Results and conclusions

After developing speaker recognition system as well as after performing a computer experiments the following conclusions were made:

1. The automatic method for voice activity detection was proposed. This method is fast and does not require any additional actions from the user, such as indicating patterns of the speech signal and noise.
2. System of features has been proposed, that consist of vocal tract parameters and excitation source parameters. Four formants and three antiformants calculated in mel-frequency scale were used as parameters of the vocal tract and pitch was used as parameter of excitation source. Proposed system of features outperformed standard MFCC features in all parameters of accuracy of recognition and EER value 5.17% was obtained using proposed features, EER value 5.86% was obtained using MFCC.
3. Proposed system of features is of lower dimension and consists of 8 components, like MFCC of 13-th order consists of 13 components. Therefore we need to implement less operations of calculation during EM algorithm, when we create speakers GMM.
4. To equalize dispersions of all formants and antiformants, they were calculated in mel-frequency scale.
5. The method of estimation of initial GMM parameters was proposed too. Vector quantization approach was proposed for this case. Experiments made had shown that vector quantization approach provided best results of accuracy in this case and outperformed other methods of forming of clusters. Method of random forming was outperformed (EER value) by 0.71% and method of linear division – by 0.88%, yet not reducing count of iterations necessary to build speaker's model.

List of Published Works on the Topic of the Dissertation

In the reviewed scientific periodical publications

1. Kamarauskas, J. 2006. Automatic Segmentation of the Phonemes using Artificial Neural Networks, *Electronics and Electrical Engineering* 8(72): 39–42. ISSN 1392–1215.
2. Kamarauskas, J. 2008. Speaker recognition using Gaussian Mixture Models, *Electronics and Electrical Engineering* 5(85): 29–32. ISSN 1392–1215 (Thomson ISI Master Journal List).

In the other editions

3. Kamarauskas, J. 2007. Kalbančiojo atpažinimas taikant vektorinį kvantavimą, *iš Informacinės technologijos 2007*, 42–46. ISSN 1822-6337.
4. Šalna, B., Kamarauskas, J. 2005. Automatinio asmens atpažinimo iš balso problemos ir perspektyvos kriminalistikoje. *Jurisprudencija* 66(58): 140–145. ISSN 1392-6195.

About the author

Juozas Kamarauskas was born in Vilnius, on 11 of July 1980.

First degree in Electronics Engineering, Faculty of Electronics, Vilnius Gediminas Technical University, 2002. Master of Science in Informatics Engineering, Faculty of Fundamental Sciences, Vilnius Gediminas Technical University, 2004. In 2004–2008 – PhD student of Institute of Mathematics and Informatics.

ASMENS ATPAŽINIMAS PAGAL BALSĄ

Mokslo problemos aktualumas – šiuolaikiniame pasaulyje vis aktualesnės tampa asmens atpažinimo pagal balsą problemos. Šios problemos atsiranda kriminalistikoje, informacijos apsaugoje, tai gali būti taikoma įėjimo kontrolės sistemose, mobiliojoje bankininkystėje, internetinėje prekyboje ir t. t. Joms yra skiriamas didelis dėmesys, materialiniai bei intelektualiniai ištekliai, sukurti įvairūs testavimo centrai. Jei kitos biometrijos rūšys reikalauja specialios, dažnai brangiai kainuojančios įrangos, asmens atpažinimo pagal balsą sistemos to nereikalauja.

Nepaisant nemažų pasiekimų šioje srityje, iki šiol nėra sukurtos teorijos kaip žmogus atskiria vieną balsą nuo kito akustiniame lygyje, o taip pat nėra sukurtos universalios požymių sistemos, leidžiančios laisvai atskirti skirtingus balsus, esant skirtingoms frazėms, skirtingai kalbėjimo aplinkai, skirtingiems garso įrašymo kanalams, triukšmams ir t. t.

Iš visų biometrijos rūšių, naudojant balso biometriją kol kas gaunami vieni iš prasčiausių rezultatų, tačiau galėtų turėti labai platų pritaikymą. Dėl šios priežasties reikėtų atlikti daugiau tyrimų šioje srityje.

Darbo tikslas – atlikti kalbančiojo atpažinimo sistemų analizę, pasiūlyti sprendimus, didinančius kalbančiojo atpažinimo sistemų veikimo tikslumą bei darbo efektyvumą.

Darbo uždaviniai. Darbo tikslui pasiekti darbe reikia spręsti šiuos uždavinius:

1. Pasiūlyti automatinį vokalizuotų garsų išskyrimo iš įrašyto kalbos signalo algoritmą.
2. Pasiūlyti naują efektyvią požymių sistemą, didinančią asmens atpažinimo tikslumą bei mažinančią reikalingų skaičiavimo operacijų skaičių.
3. Pasiūlyti efektyvų metodą kalbėtojų modelių pradinių parametrų vertinimui.
4. Realizuoti pasiūlytus metodus. Eksperimentiškai įvertinti sukurtos atpažinimo sistemos tikslumą, lyginant pasiūlytus požymius su šiuo metu vienais iš plačiausiai naudojamų pasaulyje.

Mokslinis naujumas. Disertacijoje pasiūlyta keletas sprendimų, didinančių atpažinimo sistemos veikimo tikslumą bei darbo efektyvumą. Sukurtas automatinis vokalizuotų garsų išrinkimo iš kalbos bei triukšmo signalų metodas, veikiantis tiksliau nei, pavyzdžiui, energijos slenksčio metodas. Pasiūlytas metodas yra kompleksinis, susidedantis iš kelių atskirų algoritmų: signalo kadrų su nulinėmis ir labai žemomis signalo reikšmėmis atmetimo, foninio triukšmo radimo bei melų skalės spektro slenksčio nustatymo, žadinimo signalo radimo bei nevokalizuotų garsų atmetimo. Šis metodas taip pat pašalina įvairius pavienius impulsinius trikdžius. Kalbančiųjų modelių kūrimui bei atpažinimui parenkami tik vokalizuoti garsai.

Pasiūlyta nauja požymių vektorių, turinčių nedaug komponenčių, sistema. Kaip žinoma, kalbos signalas generuojamas žadinimo signalui veikiant balso traktą. Pasiūlyti požymių vektoriai susideda tiek iš žadinimo signalo parametrų, tiek ir iš balso trakto parametrų. Kaip žadinimo signalo parametras yra naudojamas žadinimo signalo pagrindinis dažnis (F0), kaip balso trakto parametrai – formantės (kalbos signalo kadro Furjė spektro gaubtinės maksimumų dažniai) bei antiformantės (kalbos signalo kadro Furjė spektro gaubtinės minimumų dažniai). Siekiant sumažinti aukštesnių formančių bei antiformančių dispersiją, jos skaičiuojamos melų skalėje. Kadangi pasiūlytų požymių vektorių komponenčių skaičius yra nedidelis (nuo penkių iki aštuonių vektoriaus komponenčių), lyginant su tradicinėmis (trylika arba trisdešimt devynios komponentės), dėl to gerokai pagreitėja skaičiavimai, ypač pakartotiniame parametrų vertinime, matematinės vilties maksimizavimo algoritme, kuriant kalbėtojų Gauso mišinių modelius.

Pasiūlytas metodas pradiniam kalbančiųjų modelių parametrų vertinimui. Tam tikslui panaudotas modifikuotas LBG vektorinio kvantavimo algoritmas.

Tyrimų metodika apima matematikos, taip pat tikimybių teorijos bei matematinės statistikos, skaitmeninio signalų apdorojimo bei atpažinimo teorijos žinios. Asmens atpažinimo pagal balsą sistema realizuota C++ kalba, panaudojant *TURBO C++ 2006* integruotą programų kūrimo aplinką.

Ginamieji teiginiai

- Pasiūlytoji požymių sistema, susidedanti iš žadinimo signalo bei balso trakto parametrų.
- Pasiūlytas vokalizuotų garsų išskyrimo metodas.
- Pasiūlytas pradinių GMM parametrų vertinimo metodas.
- Sukurtoji automatinio kalbančiojo atpažinimo programinė įranga.

Darbo apimtis. Šią disertaciją sudaro: įvadas, 4 skyriai, bendrosios išvados, literatūros sąrašas ir autoriaus publikacijų sąrašas. Disertacijos aiškinamąjį raštą sudaro 124 teksto puslapiai, su 58 paveikslais ir 8 lentelėmis. Literatūros sąraše 120 šaltinių.

Įvade suformuluojama tiriamoji problema, aptariamas temos aktualumas, darbo tikslas, metodai ir priemonės, mokslinis naujumas, ginamieji teiginiai.

Pirmame skyriuje bendrai aptariamos kalbančiojo atpažinimo sistemos, jų klasifikacija, taip pat pagrindinės sąvokos. Čia taip pat aptariama kalbančiojo atpažinimo sistemų raida, biometrinių sistemų darbingumo vertinimo parametrai, automatinės kalbančiojo atpažinimo sistemos, o taip pat ir pagrindinės problemos, su kuriomis susiduriama kalbančiojo atpažinime.

Antrame skyriuje nagrinėjami kalbos generavimo bei modeliavimo klausimai, detaliai nagrinėjami automatinių kalbančiojo atpažinimo sistemų elementai, kalbos signalų apdorojimo klausimai, požymių sistemos, naudojamos kalbos ir kalbančiojo atpažinime, kalbos signalų segmentavimo klausimai ir kalbėtojų modelių kūrimo bei požymių palyginimo būdai, naudojami tiek priklausomame, tiek ir nepriklausomame nuo ištartos frazės kalbančiojo atpažinime.

Trečiasis skyrius skirtas kalbančiojo atpažinimo sistemos realizacijai. Čia aptariamas pasiūlytas automatinis vokalizuotų garsų išrinkimo iš kalbos signalų bei triukšmo metodas, šiek tiek modifikuotas žadinimo signalo pagrindinio dažnio radimo metodas, pasiūlyta požymių vektorių sistema bei pradinio GMM parametrų vertinimo metodas.

Ketvirtajame skyriuje pateikti sukurtos atpažinimo sistemos eksperimentinio tyrimo rezultatai. Eksperimentais tirtas kalbančiojo asmens atpažinimo tikslumas, panaudojant įvairias požymių sistemas: pasiūlytą požymių vektorių sistemą, susidedančią iš formančių, antiformančių ir žadinimo signalo pagrindinio dažnio, taip pat ir atskiras šios požymių sistemos

dalis. Rezultatų palyginimui atlikti eksperimentai ir su standartiniais požymiais – melų skalės kepstro koeficientais. Eksperimentų metu tirta ir atpažinimo tikslumo priklausomybė nuo Gauso mišinių komponenčių skaičiaus, pradinio GMM parametrų vertinimo įtaka atpažinimo tikslumui ir t. t.

Paskutiniame skyriuje apibendrinami darbo rezultatai ir suformuluojamos išvados, aptariamos tolesnės atpažinimo sistemos vystymo galimybės.

Bendrosios išvados

Darbo metu buvo atlikta kalbančiojo atpažinimo sistemų analizė, atliktas kalbančiojo atpažinimo, panaudojant pasiūlytus sprendimus, tyrimas. Gautus darbo rezultatus apibendriname ir pateikiame išvadas:

1. Pasiūlytas automatinis vokalizuotų garsų išrinkimo (segmentavimo) metodas, nereikalaujantis iš vartotojo jokių triukšmo ir kalbos signalo pavyzdžių nurodymo. Foninio triukšmo parametrai automatiškai nustatomi atmetus visus kadrus su nulinėmis ir labai žemomis signalo reikšmėmis ir po to randant tam tikrą kadrų skaičių su minimaliomis melų skalės spektro energijos reikšmėmis.
2. Pasiūlyta požymių sistema, skirta asmens atpažinimui pagal balsą, susidedanti iš žadinimo signalo parametrų bei balso trakto parametrų. Kaip žadinimo signalo parametrą panaudojome balso stygų virpėjimo dažnį – žadinimo signalo pagrindinį dažnį F0. Kaip balso trakto parametrus panaudojome keturias formantes (kalbos signalo spektro gaubtinės maksimumų dažnius) bei tris antiformantes (kalbos signalo spektro gaubtinės minimumų dažnius). Gauti atpažinimo rezultatai pagal visus tikslumo parametrus pralenkė šiuo metu vienus iš plačiausiai naudojamų pasaulyje spektrinių požymių – melų skalės kepstro koeficientus (MFCC), naudojamus kalbos bei asmens atpažinime. Gautas lygių klaidų lygis panaudojant pasiūlytą požymių vektorių sistemą – LKL=5,17 %, tuo tarpu, panaudojus MFCC, LKL=5,86 %.
3. Pasiūlytos požymių sistemos vektoriai turi mažesnį komponenčių skaičių, vektorius susideda iš 8 komponenčių, tuo tarpu, panaudojant 13 eilės MFCC, vektorius susideda iš 13 komponenčių. Dėl šių priežasčių tiek kuriant kalbėtojų modelius, tiek ir atpažinimo metu, naudojant pasiūlytą požymių vektorių sistemą reikia atlikti apie 1,6 karto mažiau skaičiavimo operacijų. Dėl to teigiame, kad požymių sistema, susidedanti iš 4 formančių, 3 antiformančių bei žadinimo signalo pagrindinio dažnio F0 gali būti panaudota kalbančiojo atpažinimui pagal balsą ir tai yra efektyvesnė požymių sistema, nei standartinė – MFCC.
4. Kadangi aukštesnių formančių bei antiformančių dispersija yra didesnė nei žemesnių, kad „suvienodinti“ dispersijas, pasiūlyta formantes bei

antiformantes skaičiuoti melų skalėje. Atlikus eksperimentus, paaiškėjo, kad formančių bei antiformančių skaičiavimas melų skalėje šiek tiek pagerino (LKL reikšmė sumažėjo 0,11–0,26 %) atpažinimo tikslumą, nei naudojant jas tiesinėje skalėje. Todėl teigiame, kad formantes bei antiformantes geriau skaičiuoti melų skalėje.

5. Pasiūlytas metodas pradinių GMM parametrų vertinimui. Tam tikslui panaudotas modifikuotas LBG vektorinio kvantavimo (VK) algoritmas. Atlikus eksperimentus paaiškėjo, kad panaudojus pradinių parametrų vertinimui VK metodą, gautas didžiausias atpažinimo tikslumas (lygių klaidų lygis sumažėjo 0,71–0,88 %), nei panaudojant atsitiktinio klasterių formavimo ar tiesiško požymių vektorių dalijimo į klasterius metodus. Tačiau VK metodas nesumažino iteracijų skaičiaus, reikalingo tikslinant GMM parametrus. Todėl galima teigti, kad vertinant pradinius GMM parametrus geriau naudoti vektorinio kvantavimo metodą, nei parinkti atsitiktines parametrų vertes.

Disertacijos darbo rezultatai parodė tolimesnę atpažinimo sistemos vystymo kryptį – esamos požymių sistemos gerinimą bei jos papildymą panaudojant kitus spektrinius bei žadinimo šaltinio požymius. Papildomais kalbančiojo požymiais gali būti panaudotas balso tembras, pirmųjų formančių amplitudžių santykis ir t. t.

Trumpos žinios apie autorių

Juozas Kamarauskas gimė 1980 m. liepos 11 d. Vilniuje.

2002 m. įgijo elektronikos inžinerijos bakalauro laipsnį Vilniaus Gedimino technikos universiteto Elektronikos fakultete. 2004 m. įgijo informatikos inžinerijos mokslo magistro laipsnį Vilniaus Gedimino technikos universiteto Fundamentinių mokslų fakultete. 2004–2008 m. – Matematikos ir informatikos instituto doktorantas.

Juozas Kamarauskas

SPEAKER RECOGNITION BY VOICE

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

ASMENS ATPAŽINIMAS PAGAL BALSĄ

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

2009 04 08. 1,5 sp. l. Tiražas 100 egz.
Vilniaus Gedimino technikos universiteto
leidykla „Technika“, Saulėtekio al. 11, 10223 Vilnius
http://leidykla.vgtu.lt
Spausdino UAB „Biznio mašinų kompanija“,
J. Jasinskio g. 16A, 01112 Vilnius
http://www.bmk.lt