Sigita LAURINČIUKAITĖ

ACOUSTIC MODELLING OF LITHUANIAN SPEECH RECOGNITION

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

1489-M

Vilnius LEIDYKLA TECHNIKA 2008

VILNIUS GEDIMINAS TECHNICAL UNIVERSITY
INSTITUTE OF MATHEMATICS AND INFORMATICS

Sigita LAURINČIUKAITĖ

ACOUSTIC MODELLING OF LITHUANIAN SPEECH RECOGNITION

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

Vilnius LEIDYKLA TECHNIKA 2008

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

Scientific Supervisor

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

Prof Dr Habil Feliksas IVANAUSKAS (Vilnius University, Physical Sciences, Informatics – 09P),

Prof Dr Habil Kazys KAZLAUSKAS (Institute of Mathematics and Informatics, Physical Sciences, Informatics – 09P),

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

Opponents:

Asoc Prof Dr Dalius NAVAKAUSKAS (Vilnius Gediminas Technical University, Technological Sciences, Electrical Engineering and Electronics – 01T),

Dr Pijus KASPARAITIS (Vilnius 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 Seminars Centre of the Institute of Mathematics and Informatics at 3 p. m. on 17 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 16 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, Lithuania).

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS INSTITUTAS

Sigita LAURINČIUKAITĖ

LIETUVIŲ ŠNEKOS ATPAŽINIMO AKUSTINIS MODELIAVIMAS

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

Vilnius VGTU 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 – 07 T).

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. Gintautas DZEMYDA (Matematikos ir informatikos institutas, technologijos mokslai, informatikos inžinerija – 07T),

prof. habil. dr. Feliksas IVANAUSKAS (Vilniaus universitetas, fiziniai mokslai, informatika – 09P),

prof. habil. dr. Kazys KAZLAUSKAS (Matematikos ir informatikos institutas, fiziniai mokslai, informatika – 09P),

dr. Algimantas Aleksandras RUDŽIONIS (Kauno technologijos universitetas, technologijos mokslai, informatikos inžinerija – 07T).

Oponentai:

doc. dr. Dalius NAVAKAUSKAS (Vilniaus Gedimino technikos universitetas, technologijos mokslai, elektros ir elektronikos inžinerija – 01T),

dr. Pijus KASPARAITIS (Vilniaus universitetas, technologijos mokslai, informatikos inžinerija – 07T).

Disertacija bus ginama viešame Informatikos inžinerijos mokslo krypties tarybos posėdyje 2008 m. birželio 17 d. 15 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 16 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“ 1489-M mokslo literatūros knyga.

1. Introduction

1.1. Topicality of the Problem

This dissertation explores an automatic speech recognition (ASR) problem. According to Jurafsky, a system of automatic speech recognition is a system for mapping acoustic signals to a string of words. Automatic speech recognition is a first block in the voice technology system. The automatic speech recognition system can be subdivided into: feature extraction from a speech signal, formation of acoustic models and acoustic modelling, classification of unknown utterance, and its attachment to one of the acoustic models. Formation of acoustic models and acoustic modelling is one of the determinants of the accuracy of speech recognition and performance of subsequent blocks. Acoustic modelling is done for each language separately because it is closely connected to the specific speech sounds of that language.

There are various acoustic modelling investigations of Lithuanian, mostly covering phoneme-, contextual-phoneme-based acoustical modelling. At the same time, there is a lack of alternative research in syllable-, word-based acoustic modelling and research of a comparative nature.

1.2. Research Object

The systems of automatic speech recognition that are modelled in this thesis are based on a statistical method of speech recognition and use Hidden Markov Models (HMM). Training of Hidden Markov Models enables us to encode the characteristics of specific speech signals. After training each Hidden Markov Model becomes an acoustic model (AM) that represents a specific speech sound. In this thesis, we mainly focus on the selection of sub-word units (phoneme, contextual phoneme, syllable, contextual syllable, and word) for acoustic modelling, acoustic modelling itself and efficiency of acoustic models.

1.3. Aim and Task of the Work

The aim of this thesis is: to develop acoustic models for different sub-word units (words, phonemes, syllables, contextual phonemes, and contextual syllables) and to implement comparative speech recognition research, using developed acoustic models. The technologies for acoustic modelling of different sub-word units with estimates of efficiency and applicability will be proposed after investigations.

With regard to the goal of this thesis we state the following problems:

1. To construct schemes of acoustic modelling with regard to a sub-word unit and type of speech; to use them for acoustic modelling.
2. To prepare speech corpora for experimental research; technologies that lack and tools for implementation of blocks of developed schemes.
3. To investigate selection, effectiveness and adaptability of acoustic models of the sub-word units for automatic speech recognition.
4. On the ground of research, to propose technologies for acoustic modelling in development of acoustic models for automatic speech recognition systems.

1.4. Methodology of Research

The knowledge and methods from different disciplines were used for theoretical and empirical research presented here, i. e., theories of digital signal processing, hidden Markov models, mathematical statistics, Lithuanian grammar and phonetics. The results of the thesis were obtained in empirical research, for which a HTK toolkit, programs developed of the thesis author, and speech corpora designed at the Institute of Mathematics and Informatics, were used.

1.5. Scientific Novelty

The scientific novelty of this dissertation is following:

1. The results of comparative speech recognition that uses acoustic models of different sub-word units present technologies of acoustic modelling of different sub-word units.
2. A new methodology of formation of a set of acoustic models of syllables and phonemes is proposed and evaluated in experimental research.
3. A new sub-word unit – pseudo-syllable that increases accuracy of speech recognition in comparison to linguistically defined sub-word units is proposed.
4. Developed acoustic models can be used in Lithuanian automatic speech recognition systems and can increase accuracy of speech recognition.

1.6. Practical Value

The results of research of this dissertation can be applied as recommendations in the development of automatic speech recognition systems to select sub-word units and acoustic modelling aspects. The acoustic models developed can be used in the automatic speech recognition system. The speech corpora developed can be used for further speech recognition research. The results of investigations were used to pursue the program of “Lithuanian Speech in an Information Society 2000–2006”.

1.7. Defended Propositions

1. Methodology of formation of a set of acoustic models of syllables and phonemes for syllable-phoneme-based speech recognition that allows investigation of sets of acoustic models of different sub-word units.
2. Technologies for acoustic modelling of sub-word units, for processing of sub-word units and lexicon, and schemes of speech recognition that allow practical implementation of training and speech recognition.
3. Acoustic models of words, phonemes, contextual phonemes, syllables and contextual syllables that are applicable in different systems of speech recognition.
4. Two versions of continuous speech corpus LRN: LRN0 and LRN1 that allow comprehensive investigation of speech recognition.

1.8. The Scope of the Scientific Work

Dissertation is written in Lithuanian and consists of following parts: notation, acronyms, introduction, five chapters, a list of references and a list of publications. The total scope of the dissertation – 108 pages, 26 pictures, 34 tables and 3 addenda.

2. Problems in Acoustic Modelling

Statistical methods, used in automatic speech recognition, presuppose the existence of statistical models that, after the training process, become representatives of speech sounds or speech sound combinations. Speech units, according to the derivation rule, are obtained either by a linguistic criterion or by an automatic clustering technique. The objects of dissertation are sub-word units according to the linguistic criterion: phonemes, syllables and words, contextual phonemes and contextual syllables. The linguistic criterion prescribes using the sets of speech units obtained by language specialists or to extract sets of speech units according to the fixed grammar rules.

Acoustical modelling of Lithuanian remains one of important tasks. The research in dynamic time warping, which strove to solve the whole word recognition task, was gradually replaced by sub-word units, such as phonemes, recognition. Phoneme-based recognition is more universal, although it does not yield as good results as the word-based recognition. Implementation of a word-based recognizer is also simpler in comparison to sub-word-based recognizers. Modern speech recognition systems for Lithuanian employ phoneme-based recognition. These speech recognition systems are built according to the existing database resources that have a set of phonemes fixed a priori. The fixed set of phonemes is used to find optimal system parameters or to investigate additional features, such as stress, softness of consonants, and decomposition of mixed diphthongs into the basic set of phonemes. These researches established the usage of phonemes without inquiry in other sub-word units. No efforts have been made in the further more profound investigation of other sub-word units.

The analysis of literature on the subject of acoustic modelling and selection of sub-word units presented two problems: 1) researchers select a sub-word unit for acoustic modelling without the parallel research of different sub-word units; 2) research on selection of a sub-word unit for Lithuanian is scarce. Hence, we formulate the following objectives for research: investigate how different types of sub-word units and selection of units into a set, according to which acoustical models are developed, influence the speech recognition accuracy for different types of speech. We make a hypothesis that a detailed investigation of selecting sub-word units can help to increase the speech recognition accuracy. The following tasks were set: 1) to compare different types of sub-word units, 2) to investigate in detail each type, 3) if there is no technique to use a sub-word unit type in speech recognition, to propose new one, 4) investigate different types of speech (isolated words and continuous speech).

3. Description of the Structures of the Speech Recognition Systems

The HMM-based approach of speech recognition methods was chosen in for this dissertation. The task of speech recognition is to decode the sequence of words W^* from the speech signal S. We denote the speech signal S as a sequence of feature vectors $\mathbf{O} = \mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_T$, where T depends on the length of the speech signal. Then the problem of decoding becomes a problem of selecting a sequence of words W^* from all the possible sequences of words W^{**} with the highest probability:

$$W^*_{\max} = \arg\max_{W^* \in W^{**}} P\left(W^* \mid \mathbf{O}\right) \approx ... \approx \arg\max_{W^*} P\left(\mathbf{O} \mid W^*\right) P\left(W^*\right), \qquad (1)$$

here $P\left(\mathbf{O} \mid W^{*}\right)$ denotes a posterior probability of the observed feature vectors and $P\left(W^{*}\right)$ – probability of the prior sequence of words.

The Common Structure of ASR. The common structure of the ASR system is shown in Fig 1. The structure itself is applicable to all languages. Further the main elements of the structure are described.

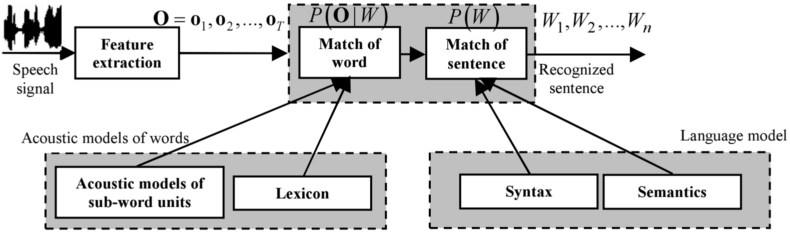

Fig 1. The structure of the ASR system

Feature extraction reduces information in a speech signal by parameterization. A speech signal is segmented into 25–30 ms overlapping frames. A vector of features is found for each frame. Mel Frequency Cepstral Coefficients (MFCC) were used in the experimental research. One feature vector consisted of 39 values, i. e., 12 MFCC, one value of energy, and first- and second-order time derivatives.

Acoustic models of words are the result of acoustic modelling – a process of building and development of acoustic models according to the sub-word units derived from a linguistic criterion and training data. Before the training starts, the structure of parameters of the acoustic model is set (the form of the acoustic model in this work is HMM). Subsequent operations of the training (the algorithm of Baum-Welch was used in this work) refine and adjust the values of parameters of the acoustic model to the training data.

The lexicon includes all the words used in the ASR system modelling and subsequent recognition task. It gives the transcription of a word in a meaningful sequence of sub-word units (each sub-word unit has an acoustic model). Each set of acoustic models has no less than one its own lexicon.

The Specification of Common Structure of ASR. Described structure was specified for empirical research of acoustic modelling of sub-word units and for speech types (isolated words, continuous speech) separately. Three schemes were developed for acoustic modelling of: 1) word-based recognition system of isolated words, 2) syllable- or phoneme-based recognition system of continuous speech, and 3) contextual syllable- or contextual phoneme-based recognition system of continuous speech (these are not given here because of the sizes of schemes). These schemes were used for experimental research. Implementation of different processes (as development and modification of lexicon, development and cloning of prototypes of acoustic models, development of questioner of clasterization) required development of new technologies and tools for completion of the tasks.

The Methodology for Syllable- and Phoneme-based Acoustic Modelling. A methodology (shown in Fig 2) is proposed for creation of a set of syllables and phonemes, later used for acoustic modelling. This methodology is distinctive as it uses new sub-word unit – pseudo-syllable. According to it, to get the basic set of syllables and phonemes and adjust the lexicon to it, you have to follow 8 steps, some of which have an alternative.

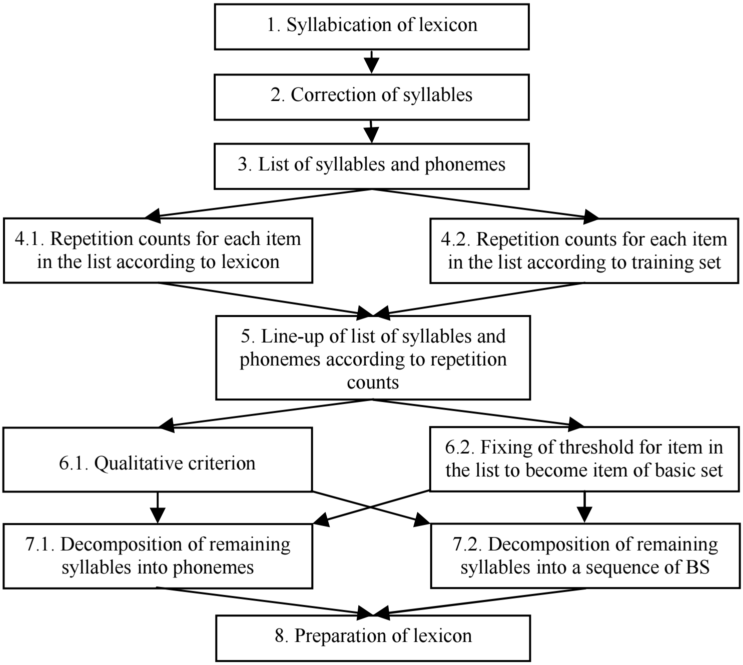
1. Syllabication of lexicon
2. Correction of syllables
3. List of syllables and phonemes
4.1. Repetition counts for each item in the list according to lexicon
4.2. Repetition counts for each item in the list according to training set
5. Line-up of list of syllables and phonemes according to repetition counts
6.1. Qualitative criterion
6.2. Fixing of threshold for item in the list to become item of basic set
7.1. Decomposition of remaining syllables into phonemes
7.2. Decomposition of remaining syllables into a sequence of BS
8. Preparation of lexicon

Fig 2. Framework for the construction of a syllable and phoneme set

The methodology was used in experimental research after implementation of different processes. It was experimentally optimised.

The Measure of Performance of Recognition Systems. The performance of recognition systems was measured by word level accuracy (WA), defined as $WA = \frac{R}{N} \times 100\ \%$ – for isolated word recognition and $WA = \frac{N - S - I - D}{N} \times 100\ \%$ – for continuous speech recognition. Here *N* is the number of words in the test or development set in total, *R* is the number of correctly recognized words, *S* is the number of word substitution errors, *I* is the number of word insertion errors, and *D* is the number of word deletion errors.

Confidence intervals were calculated following two patterns: one for the results obtained by the cross-validation principle and another – for the results obtained by testing one data set, i. e., calculation of confidence intervals by normal distribution, when variance is unknown, and calculation of confidence intervals by approximating the binomial distribution to normal distribution.

4. Development of Lithuanian Speech Corpora

An isolated word corpus and a continuous speech corpus LRN are presented in this chapter. The contributions of the author of the dissertation to development of continuous speech corpus LRN are following: setting-up of process of corpus development, selection of structure and main characteristics of corpus, selection of system of phonemes; preparation and division of tasks for members of workgroup (herself among them); verification of completed tasks; description of corpus.

The Speech Corpus LRN. The corpus LRN is increasingly expanded and at the current moment a few versions are available. The version LRN0 was used for experiments. The corpus LRN0 contains speech samples from the news broadcasts over the Lithuanian Radio in 2003–2004. The recorded speech covers political, economic, cultural, and sport areas of local and foreign affairs. There were 31 speaker: 17 females and 14 males. The distribution of speech records is not equally distributed among the speakers. Actually, speech records of 13 speakers make up 93 % of the entire speech corpus. The speech corpus is partitioned into training, development, and evaluation data sets. Speakers for all the sets are the same. The phonetic system SAMPA-LT was chosen as most suitable for speech recognition purposes. There were 18 374 distinct words in the corpus LRN0 lexicon.

The Isolated Word Corpus. The speech corpus of isolated words is of 30 min. duration. There were 31 speakers: 13 females and 18 males. Each speaker has pronounced 50 words for 20 times. Hence, each word out of 50 has 620 repetitions. Speech records of each word were placed in a separate file without silence marks at the beginning and at end of utterance. For each speech record a word level annotation file was prepared.

5. Experimental Research

5.1. Experiment 1

Acoustic models of type of sub-word unit of word are used for isolated word recognition to investigate speech recognition accuracy dependence on the size of a training set and the number of speakers. A corpus of isolated words was used. Two ASR systems were modelled: speaker dependent – IZ_PNK and speaker independent – IZ_NNK. For the speaker dependent ASR system, the dependence of size of the training set on WA and performance of additional speakers have been investigated. All AM's are listed in Table 1.

Table 1. AM of IZ_PNK and IZ_NNK systems (with division of corpus is sessions)

Name of AM Set	**Division**	**Add. speakers**	**Description**
	Training/Testing		
AM_10_40	10 / 40	4×20	AM of 50 words of one speaker formed using 10 items for training and 40 items for testing.
AM_20_30	20 / 30	4×20	AM of 50 words of one speaker formed using 20 items for training and 30 items for testing.
AM_25_25	25 / 25	4×20	AM of 50 words of one speaker formed using 25 items for training and 25 items for testing.
AM_30_20	30 /20	4×20	AM of 50 words of one speaker formed using 30 items for training and 20 items for testing.
AM_40_10	40 / 10	4×20	AM of 50 words of one speaker formed using 40 items for training and 10 items for testing.
AM_20_20	25×20	5×20	AM of 50 words of 25 speakers formed using 25×20 items for training and 20 items for testing.

The recognition results are given in Table 2. WA dependence on the training set is obvious as WA increases with an increase of the training set size from 99,80 ±0,23 % to 100 % for the system IZ_PNK, speaker P1, and from 13,89 ±8,88 % to 31,17 ±19,69 % for the same system, additional speakers. The use of additional speakers in testing AM of the system IZ_PNK showed low WA (the highest WA – 31,17 ±19,69 %) and a high dispersion of the results. This problem is circumvented in the system IZ_NNK, where 25 speakers are used for training AM. Therefore the WA of additional speakers is 99,44 ±0,48 %.

Recommendations for acoustic modelling of the unit type of word:

- To define the number of speakers whose speech will be used for recognition and to select the content of training sets of acoustic models according to a previous definition.
- To select sizes of the training set of acoustic models according to the size of a set of acoustic models.
- The sizes of training sets for poorly discriminated words have to be larger than for usual words.

Table 2. Recognition results of the systems IZ_PNK and IZ_NNK in WA

Name of AM Set	WA of Speaker P1, %	Average WA of Add. Speakers %	WA of Additional Speakers, % (20 sessions)				
			P1	P2	P5	P6	P7
AM_10_40	99,40±0,28	13,89	99,80	10,22	24,25	6,78	14,30
AM_20_30	99,87±0,15	15,71	99,90	9,32	31,56	6,67	15,30
AM_25_25	99,92±0,13	18,57	99,90	11,52	37,27	7,58	17,90
AM_30_20	99,90±0,16	23,83	99,90	16,03	46,29	9,00	24,00
AM_40_10	100	31,17	100	28,76	53,01	12,22	30,70
AM_20_20	–	99,44	99,70	99,60	99,90	99,40	98,60

5.2. Experiment 2

Acoustic models of word, syllable, and phoneme sub-word unit types are used for isolated word recognition in order to compare the speech recognition accuracy dependence on the sub-word unit type. The speaker independent speech recognition system IZ_NNK_ZSF is modelled. A cross-validation technique was applied and a set of acoustic models was developed and tested for 10 speakers out of 31. Actually, there are 3 sets of AM's for each type of a sub-word unit for each of 10 speakers.

For simplicity, the system IZ_NNK_ZSF is decomposed into 3 systems: IZ_NNK_Z – for word-based recognition, IZ_NNK_S – for syllable-based recognition, and IZ_NNK_F – for phoneme-based recognition. The average of recognition results of 10 speakers was calculated for each of the three systems. Word-based speech recognition achieved the word recognition accuracy of 97,77 ±1,40 % (IZ_NNK_Z), syllable-based – 98,04 ±1,75 % (IZ_NNK_S), and phoneme-based speech recognition – 93,91 ±3,59 % (IZ_NNK_F). These results show the efficiency of AM of syllables and words in the recognition of isolated words. Poor results of phoneme-based recognition indicate the limits of their use that depend on the type of speech.

The experiments have showed that syllables are promising units for speech recognition, but we recommend using the sub-word unit type of word for simplicity of acoustic modelling of these units.

5.3. Experiment 3

Acoustic models of the phoneme sub-word unit type are used for continuous recognition aimed to the analysis of speech recognition accuracy dependence on the formation of a phoneme set (was performed together with dr. D. Šilingas). Five phoneme sets and four triphone sets are listed in Table 3.

Table 3. The sets of phonemes and triphones

Name of Set	Description
AM_MKD	Phoneme set with marks of softness of consonants and with marks of accent
AM_KD	Phoneme set without marks of softness of consonants and with marks of accent
AM_D	Phoneme set without marks of softness of consonants and of accent
AM_MD	Phoneme set with marks of softness of consonants and without marks of accent
AM_MK	Phoneme set with marks of softness of consonants and with marks of accent, mixed dipthongs are decomposed into two parts
T_AM_MKD	Triphone set derived from set AM_MKD
T_AM_KD	Triphone set without marks of softness of consonants and with marks of accent derived from set AM_MKD
T_AM_MK	Triphone set derived from set AM_MK
T_AM_K	Triphone set with marks of accent derived from set AM_MK

Table 4. Recognition results for the sets of phonemes and triphones. The results are given in WA with 95 % confidence intervals for the development and test sets

Name of Set of Phonemes and Triphones	WA, % (development set)				WA, % (test set)
	1	2	3	4	4
AM_MKD	41,39	55,56	57,50	60,56 ±4,01	62,38 ±1,49
AM_MD	43,33	54,44	57,50	59,44 ±4,03	61,04 ±1,50
AM_KD	38,89	50,83	54,17	58,33 ±4,04	58,66 ±1,52
AM_D	37,22	48,89	52,78	55,28 ±4,08	57,44 ±1,52
AM_MK	41,94	54,17	55,56	61,94 ±3,98	61,57 ±1,50
T_AM_MKD	69,15	75,62	74,13	75,37 ±3,53	75,49 ±1,32
T_AM_KD	67,91	74,38	74,13	75,62 ±3,52	74,83 ±1,34
T_AM_MK	72,64	79,1	79,35	79,60 ±3,31	77,09 ±1,29
T_AM_K	71,14	76,37	80,35	79,35 ±3,32	77,44 ±1,29

The results of WA are given in Table 4. The best recognition result for sets of phonemes was achieved with the AM set AM_MKD – 62,38 ±1,49 %, the second result was of the AM set AM_MK – 61,57 ±1,50 %. These results suggest that the marks of accent and softness of consonants influence WA. The best recognition results for the sets of triphones were achieved with the AM set T_AM_K – 77,44 ±1,29 % and T_AM_MK – 61,57 ±1,50 %, which were derived from the phoneme set AM_MK. The best phoneme set and the triphone set are not coincident. Other than the recognition results for sets of phonemes, the best triphone set has no marks of softness, it was derived from the second best set of phonemes.

The results of this experiment lead to the following recommendations on the formation of phoneme and triphone sets for acoustic modelling:

- To use the sub-word unit type of phoneme for acoustic modelling with the view of simplicity of acoustic modelling, and to use the sub-word unit type of triphone with a view to increase the recognition accuracy.
- To use the phoneme set with (or without) marks of softness of consonants, with marks of accent, and dividing mixed diphthongs into two components.
- To use the topology of a phoneme model for modelling mixed diphthongs.

5.4. Experiment 4

Following the given schemes and methodology, given in chapter 3, we established some sets of syllables and phonemes, lexicons and carried out recognition experiments that condition the recommendations for further syllable- and phoneme-based recognition. They are listed in Table 5. During the recognition process two patterns were adopted for testing: 1) based on the cross-validation principle, 10 sets of speech records of approximately 1 hour duration have been tested and the average result calculated, 2) simple testing of 16 min test set. As the research progressed, the second pattern was used more frequently. Hence, the results of speech recognition in a few places are given for the first pattern, in others – for the second one.

The results of speech recognition accuracy for the above listed sets of syllables and phonemes are given in Table 6. The first glimpse at the results shows an advantage of set H_2, which was formed according to the repetition counts of syllables in the training data, i.e., using path 2 in the fourth block (4.2). The recognition accuracy suffers, if one chooses simplicity in building a set of syllables and phonemes according to path 1 (4.1). Taking into account that sets H_1 and H_2 are of similar size and the models of items from set H_2 had even more training data, we can conclude that the data-dependent model set meets the requirements of speech recognition better than the lexicon-dependent models.

Next we are able to examine two types of lexicons _P and _SP, designed for sets H_1 and H_2. The first one represents the case, where non-basic syllables were decomposed into phonemes (7.1); in the second case, non-basic syllables were decomposed into a sequence of basic syllables and phonemes (7.2). The recognition results show the dominance of lexicons _SP. Lexicons of this type increase the recognition accuracy. We could conclude that units of longer duration in the lexicon have a greater influence on the recognition accuracy. Meanwhile, supplementary modelling of phoneme and triphone models H_1_M, H_1_K, H_1dPm, H_1dTm did not increase the recognition accuracy and was discarded in further modelling.

Table 5. The sets of syllables and phonemes, and lexicons

Name of the Set	Description
H_1	The set of syllables and phonemes of 293 items (227 syllables, 63 phonemes and diphthongs), formed according to syllable repetition counts (> 50 repetitions) in the lexicon.
H_2	The set of syllables and phonemes of 289 items (223 syllables, 63 phonemes and diphthongs), formed according to syllable repetition counts in training data. Difference form H_1 in 45 syllables
H_1_K	The set of syllables and phonemes of 235 items (227 syllables, 5 groups of phonemes and diphthongs). Syllables are the same as H_1, phonemes and diphthongs are grouped
H_1_M	The set of syllables and phonemes of 283 items (233 syllables, 47 phonemes and diphthongs). Syllables are the same as H_1; softness of phonemes is not marked
H_1dPM	The set of syllables and phonemes where syllables are the same as H_1, models of phonemes and diphthongs are prepared in a separate experiment
H_1dTM	The set of syllables and phonemes where syllables are the same as H_1, models of phonemes and diphthongs are prepared in a separate experiment, triphone models are added
H_1P H_2P	Lexicons, formed from the set H_1 and H_2, by decomposing non-basic syllables into a sequence of phonemes
H_1SP H_2SP	Lexicons, formed from the set H_1 and H_2, by decomposing non-basic syllables into the sequence of basic phonemes and syllables
P_1	The set of phonemes with softness and stress marks, diphthongs of 227 items chosen from Experiment 3 as the best phoneme set
T_P_1	The set of triphones derived from P_1
T_H_2	The set of contextual syllables derived from H_2

The comparative results of purely phoneme-based (P_1) and syllable-phoneme-based (H_2) speech recognition show that the set H_2 is better than the set P_1. Meanwhile passage from the pure phonemes and syllables to the contextual phonemes (T_P_1) and contextual syllables-phonemes (T_H_2) demonstrated that the set H_2 is worse than the set T_P_1. It means that we could place the syllable-phoneme-based recognition between the phoneme- and triphone-based recognition.

The sixth block (6.2) of the framework in Fig 2 – fixing the threshold for an item in the list to become an item of the basic set of syllables and phonemes (BS) – has been investigated. The threshold value was increased and decreased; the change of added (withdrawn) units for each step was ~30. The results show that an increase or decrease in the value of the threshold influences the speech recognition accuracy. It is possible to achieve a threshold value, a deviation from which is insignificant (threshold values of 20 and 30 for set H_2). Inclusion of 10 % of all syllables in BS gives the best WA of 67,38 ±1,44 % in the case of the speech corpus LRN0.

Table 6. Recognition results for H_1 and H_2 sets of syllables and phonemes and two lexicons for each set

Name of the Set of Syllables and Phonemes	**Lexicon**	**WA with 95 % confidence intervals**	
		Average Results	**Test Set**
H_1	H_1P	46,96 ± 0,37	60,94 ±1,50
	H_1SP	48,69 ± 0,42	63,81 ±1,48
H_2	H_2P	53,08 ± 0,22	61,92 ±1,49
	H_2SP	56,67 ± 0,33	65,38 ±1,46
P_1	standard	51,81 ± 0,28	62,38 ±1,46
T_P_1	standard	-	75,49 ±1,32
T_H_2	H_2SP	-	71,79 ±1,38

The sixth block (6.1) of the framework in Fig 2 – a qualitative criterion – was investigated. Commonly, one decides about the accuracy of models from the performance of all models in speech recognition. The accuracy of distinct acoustic models is not calculated and estimation criteria are not known. It leaves a gap in a deeper understanding of speech recognition errors. We set a goal to test one criterion for this kind of model evaluation and by the accuracy of distinct models to modify our syllable-phoneme unit set. We predicted that such a solution could lead us to another investigation point – syllable-phoneme unit set formation according to the structure, and pattern of a syllable, i.e., according to the quality of a syllable that could be supplementary to frequency criteria. The results of investigation corroborated that the formation pattern of a set of syllable-phoneme where the threshold is used, is more effective and simpler.

For acoustic modelling of a syllable-phoneme sub-word unit we recommend:

- To use the proposed methodology for selecting a set of syllables and phonemes.
- To use sub-word unit types of a phoneme, phoneme-syllable in order to simplify acoustical modelling, and to use a sub-word unit type of a triphone in order to increase the speech recognition accuracy.

6. Results and Conclusions

The dissertation investigates word-, phoneme-, syllable-, contextual syllable- and contextual phoneme-based Lithuanian speech recognition. The following results were obtained during acoustic modelling of speech and research of speech recognition:

1. The new methodology to form a lexicon and a syllable-phoneme set that are used for the development of acoustic models has been proposed in the case of syllable- and phoneme-based continuous speech recognition. The tools were developed for implementation of methodology. Novelty of methodology:
   - The quantitative criterion is proposed for selection of syllables and phonemes in acoustic modelling.
   - The new sub-word unit – pseudo-syllable – is proposed that increase accuracy of speech recognition is proposed.
2. Three schemes of modelling of speech recognition systems were proposed according to speech and sub-word unit type:
   - The modelling scheme of word-based recognition system of isolated words.
   - The modelling scheme of syllable- or phoneme-based recognition system of continuous speech.
   - The modelling scheme of contextual syllable- or contextual phoneme-based recognition system of continuous speech.
3. The two versions of speech corpus LRN were developed. The author of the dissertation was responsible for: setting-up of process of corpus development, selection of structure and main characteristics of corpus, selection of system of phonemes; preparation and division of tasks for members of workgroup (herself among them); verification of completed tasks; description of corpus.

Results obtained and comparative investigations of speech recognition allow statement of following conclusions:

1. The review of acoustic modelling of sub-word units for Lithuanian and other languages showed that acoustic modelling of Lithuanian lacks not only analysis of more different sub-word units, but also it is impossible to compare effectiveness of acoustic modelling because of different speech corpora used for evaluation. It will be possible to propose technology of acoustic modelling of each sub-word unit with estimate of effectiveness after investigation of acoustic modelling of all sub-word units.

2. The proposed new methodology for selection of a set of syllables and phonemes for the acoustic modelling allows performing more syllable-based acoustic modelling. Optimization of separate blocks of methodology and of acoustic modelling increases accuracy of continuous speech recognition (from 61 ±1,5 % to 72 ±1,4 %).
3. The results of investigation shows that a new sub-word unit – pseudo-syllable – increases continuous speech recognition in comparison to conventionally defined syllable (from 61 ±1,5 % to 65 ±1,5 %).
4. The investigations of speech recognition of isolated words and continuous speech propose technologies for acoustic modelling of separate sub-word units using developed schemes of modelling of speech recognition systems, recommendations and tools.
5. The investigations of acoustic modelling shows following effectiveness of acoustic modelling of sub-word units:
   - The results of investigation confirmed that contextual phoneme-based continuous speech recognition affords better speech recognition accuracy (76 ±1,3 %) in comparison to other sub-word-based continuous speech recognition, but does not solve problem of inclusion of new word in lexicon.
   - The investigation showed that syllable- and phoneme-based continuous speech recognition (67 ±1,4 %) is superior to phoneme-based continuous speech recognition (62 ±1,5 %); and is recommended for use.
   - The investigations showed that word-based isolated word recognition is more proper for isolated word recognition (accuracy of recognition up to 100 %).
6. Developed speech corpus and acoustic models are universal and applicable in different systems of speech recognition. Developed acoustic models developed can increase accuracy of speech recognition.

List of Published Works on the Topic of the Dissertation
In the Reviewed Scientific Periodical Publications

1. LAURINČIUKAITĖ, S.; LIPEIKA, A. Framework for Choosing a Set of Syllables and Phonemes for Lithuanian Speech Recognition. *Informatica,* 2007, 18(3): 395–406. ISSN 0868-4952 (Thomson ISI Master Journal List).
2. LAURINČIUKAITĖ, S.; LIPEIKA, A. Syllable-Phoneme based Continuous Speech Recognition. *Electronics and Electrical Engineering*, 2006, 6(70): 91–94. ISSN 1392-1215 (Thomson Scientific (ISI)).

3. LAURINČIUKAITĖ, S.; ŠILINGAS, D.; SKRIPKAUSKAS, M.; TELKSNYS, L. Lithuanian Continuous Speech Corpus LRN 0.1: Design and Potential Applications. *Information Technology and Control*, 2006, 4: 431–440. ISSN 1392-124X (INSPEC).

In the other editions

4. ŠILINGAS, D.; LAURINČIUKAITĖ, S.; TELKSNYS, L. Towards Acoustic Modelling of Lithuanian Speech. In *Proceedings of International Conference „SPECOM 2004", held in Sankt Petersburg on 20–22 September 2004* (Tarptautinės konferencijos „SPECOM 2004", įvykusios Sant Peterburge 2004 m. rugsėjo 20–22 d., medžiaga). Sankt Petersburg: Anatolya, 2004, p. 326–333. ISBN 5-7452-0110-X.
5. ŠILINGAS, D.; LAURINČIUKAITĖ, S.; TELKSNYS, L. A Technique for Choosing Efficient Acoustic Modelling Units for Lithuanian Continuous Speech Recognition. In *Proceedings of International Conference „SPECOM 2006", held in Sankt Petersburg on 25–29 June 2006* (Tarptautinės konferencijos „SPECOM 2006", įvykusios Sant Peterburge 2006 m. birželio 25–29 d., medžiaga). Sankt Petersburg: Anatolya, 2006, p. 61–66. ISBN 5-7452-0074- X.
6. LAURINČIUKAITĖ, S. On different kinds of speech units based isolated words recognition of Lithuanian language. In *Proceedings of the Conference „Human language technologies – The Baltic Perspective", held in Riga on 21–22 April 2004* (Tarptautinės konferencijos „*Human language technologies – The Baltic Perspective*", vykusios Rygoje 2004 m. balandžio 21–22 d., medžiaga). Riga: Data Media Group, 2004, p. 139–143.

About the Author

1996–2002: studies at Vilnius Pedagogical University, Faculty of Mathematics and Informatics: Bachelor of Mathematics, Master of Informatics. 2003–2007: PhD studies at Institute of Mathematics and Informatics, Department of Process Recognition. From 2005: lecturer at Vilnius Pedagogical University, Faculty of Mathematics and Informatics

LIETUVIŲ ŠNEKOS ATPAŽINIMO AKUSTINIS MODELIAVIMAS

Mokslo problemos aktualumas. Automatinis šnekos atpažinimas yra pirmoji grandis balsinių technologijų produktuose užtikrinanti pirminės originalios informacijos gavimą iš šnekos signalo. Akustinių modelių aibės sudarymas yra susijęs su konkrečios kalbos garsų aibe ir specifika, todėl kiekvienai kalbai yra atliekamas atskirai. Lietuvių šnekai yra atlikta įvairių akustinio modeliavimo tyrimų, tačiau trūksta alternatyvių skiemenų, žodžių ir lyginamojo akustinio modeliavimo darbų.

Darbo tikslas ir uždaviniai. Darbo tikslas yra pagal skirtingiems kalbos vienetams sukurtus akustinius modelius atlikti lyginamuosius šnekos atpažinimo tyrimus, kurie leistų pasiūlyti įvairių kalbos vienetų akustinio modeliavimo technologijas ir įvertinti kalbos vienetų akustinių modelių efektyvumą ir panaudojimo galimybes. Remiantis darbo tikslu suformuluoti šie uždaviniai:

1. Sudaryti akustinio modeliavimo schemas atsižvelgiant į kalbos vieneto (žodžių, fonemų, skiemenų, kontekstinių fonemų ir kontekstinių skiemenų) ir šnekos tipą (izoliuoti žodžiai, ištisinė šneka). Šias schemas naudoti akustinio modeliavimo tyrimams.
2. Paruošti garsynus, reikalingus eksperimentiniams tyrimams, ir trūkstamas technologijas ir įrankius sukurtųjų schemų blokų realizavimui.
3. Eksperimentais ištirti akustinių modelių kūrimą lingvistinio kriterijaus būdu gaunamiems kalbos vienetams bei sukurtųjų akustinių modelių efektyvumą ir pritaikomumą įvairių šnekos tipų atpažinimui.
4. Atlikus tyrimus pateikti įvairių kalbos vienetų akustinio modeliavimo technologijas akustinių modelių kūrimui automatinio šnekos atpažinimo sistemoms.

Mokslinis darbo naujumas. Gauti moksliniai rezultatai yra šie:

1. Pagal lyginamųjų šnekos atpažinimo tyrimų rezultatus, gautus naudojant žodžių, skiemenų, fonemų, kontekstinių skiemenų ir kontekstinių fonemų akustinius modelius, pateikiamos kalbos vienetų akustinio modeliavimo technologijos.
2. Sukurta mišrios skiemenų ir fonemų akustinių modelių aibės kūrimo metodika, pagal ją atliekant eksperimentinius testavimus.
3. Pasiūlytas naujas kalbos vienetas – pseudo-skiemuo, gerinantis šnekos atpažinimo tikslumą, lyginant su lingvistiškai apibrėžiamais vienetais.
4. Sukurti akustiniai modeliai (ištekliai), naudojami konkrečioje automatinio šnekos atpažinimo sistemoje ir galintys pagerinti šnekos atpažinimo rezultatus.

Tyrimų metodika. Šiame darbe naudoti skaitmeninio signalų apdorojimo, paslėptųjų Markovo modelių teorijos, matematinės statistikos, lietuvių kalbos gramatikos ir fonetikos metodai ir sąvokos. Disertacijos rezultatai gauti naudojant įrankių paketą HTK, darbą palengvinančius autorės sukurtus automatinius įrankius ir darbo autorės sukurtus ar paruoštus izoliuotų žodžių ir ištisinės kalbos LRN0 garsynus.

Praktinė reikšmė. Atliktų tyrimų rezultatai kaip rekomendacijos gali būti taikomi kuriant konkrečias lietuvių kalbos automatinio šnekos atpažinimo sistemas kalbos vienetų parinkimo ir akustinių modelių sudarymo klausimais. Sukurti kalbos vienetų akustiniai modeliai gali būti taikomi, kaip automatinio šnekos atpažinimo sistemos ištekliai.

Ginamieji disertacijos teiginiai

1. Skiemenų ir fonemų akustinių modelių aibės sudarymo metodika, leidžianti sistemingai ištirti įvairių kalbos vienetų akustinių modelių aibes.
2. Autorės sukurtos technologijos kalbos vienetų akustiniam modeliavimui, šnekos atpažinimo schemų blokų realizavimui, kalbos vienetų ir žodynų apdorojimo automatizavimui, leidžiančios praktiškai realizuoti mokymą ir šnekos atpažinimą.
3. Sukurti žodžių, fonemų, skiemenų, kontekstinių fonemų ir kontekstinių skiemenų akustiniai modeliai, tinkantys naudojimui įvairiose šnekos atpažinimo sistemose.
4. Ištisinės lietuvių šnekos garsyno LRN versijos LRN0 ir LRN0.1, leidžiančios atlikti visapusiškus šnekos atpažinimo tyrimus.

Darbo apimtis. Disertaciją sudaro įvadas, 5 skyriai, literatūros sąrašas (102 nuorodos) ir 3 priedai. Pagrindinė darbo dalis 108 puslapiai.

Bendrosios išvados. Disertacijoje ištirtas žodžiais, fonemomis, skiemenimis, kontekstinėmis fonemomis ir kontekstiniais skiemenimis grįstas lietuvių šnekos atpažinimas. Atliekant lietuvių šnekos akustinį modeliavimą ir šnekos atpažinimo tyrimus gauti šie rezultatai:

1. Modeliuojant skiemenimis ir fonemomis grįstą ištisinės šnekos atpažinimo sistemą pasiūlyta metodika skiemenų ir fonemų akustinių modelių aibės, žodyno sudarymui. Sukurti šią metodiką realizuojantys įrankiai. Pagrindinis naujumas metodikoje:
   - Pasiūlytas akustinio modeliavimo kiekybinis kriterijus skiemenų ir fonemų atrinkimui.
   - Pasiūlytas naujas kalbos vienetas – pseudo-skiemuo, didinantis šnekos atpažinimo tikslumą.
2. Buvo pateiktos trys automatinio šnekos atpažinimo sistemų modeliavimo schemos pagal šnekos ir kalbos vienetų tipus:

- Žodžiais grįsta izoliuotų žodžių atpažinimo sistemos modeliavimo schema.
- Fonemomis ar skiemenimis grįsta ištisinės šnekos atpažinimo sistemos modeliavimo schema.
- Kontekstinėmis fonemomis ar kontekstiniais skiemenimis grįstos ištisinės šnekos atpažinimo sistemos modeliavimo schema.

3. Sukurtos ištisinės šnekos LRN garsyno LRN0 ir LRN0.1 versijos. Darbo autorė buvo atsakinga už: garsyno kūrimo proceso sudarymą; garsyno struktūros, pagrindinių garsyno charakteristikų ir fonemų sistemos parinkimą; darbo užduočių paruošimą ir paskirstymą darbo grupės nariams (tarp kurių buvo ir pati); atliktų darbų tikrinimą ir garsyno aprašymą.

Gauti rezultatai ir atlikti lyginamieji šnekos atpažinimo tyrimai leidžia daryti šias išvadas:

1. Atlikta analitinė kalbos vienetų akustinio modeliavimo apžvalga parodė, kad lietuvių šnekos akustiniam modeliavimui trūksta ne tik įvairesnių kalbos vienetų akustinio modeliavimo analizės (kai kuriems kalbos vienetams ji nėra atlikta), bet ir neįmanoma atlikti šių modeliavimų efektyvumo palyginimo dėl testavimui naudojamų skirtingų garsynų. Atlikus visų kalbos vienetų akustinio modeliavimo analizę būtų galima pasiūlyti kiekvieno kalbos vieneto akustinio modeliavimo technologiją su efektyvumo įverčiu.
2. Pasiūlyta skiemenų ir fonemų atrinkimo akustiniam modeliavimui metodika sudaro sąlygas atlikti akustinį modeliavimą, dažniau grindžiamą skiemenimis nei fonemomis. Atskirų metodikos ir akustinio modeliavimo etapų analizė didina ištisinės šnekos atpažinimo tikslumą (nuo 61 ±1,5 % iki 72 ±1,4 %).
3. Tyrimo rezultatai parodė, kad įvedus naują kalbos vienetą – pseudoskiemenį, palyginanti su standartiškai apibrėžiamais skiemenimis padidėja ištisinės šnekos atpažinimo tikslumas (nuo 61 ±1,5 % iki 65 ±1,5 %).
4. Atlikti izoliuotų žodžių ir ištisinės šnekos atpažinimo tyrimai teikia technologijas atskirų kalbos vienetų akustiniam modeliavimui, naudojant sudarytas automatinio šnekos atpažinimo sistemų modeliavimo schemas, rekomendacijas ir sukurtus įrankius.
5. Akustinio modeliavimo tyrimai pateikia šį atskirų kalbos vienetų akustinio modeliavimo efektyvumą:
   - Tyrimų rezultatai patvirtino, kad dažniausiai naudojamas kontekstinėmis fonemomis grindžiamas ištisinės šnekos atpažinimas teikia didesnį atpažinimo tikslumą (76 ±1,3 %) palyginanti su kitais

kalbos vienetais grindžiamu ištisinės šnekos atpažinimu, tačiau neišsprendžia naujo žodžio įtraukimo į žodyną problemos.

- Tyrimu buvo įrodyta, kad skiemenimis ir fonemomis grindžiamas ištisinės šnekos atpažinimas yra pranašesnis (67 ±1,4 %) už fonemomis grindžiamą ištisinės šnekos atpažinimą (62 ±1,5 %) ir todėl rekomenduojamas naudoti.
- Tyrimų rezultatai parodė, kad izoliuotų žodžių atpažinimui labiau tinka žodžiais grįstas šnekos atpažinimas (atpažinimo tikslumas iki 100 %).

6. Sudarytas garsynas, tyrimuose sukurti akustiniai modeliai yra universalūs ir gali būti taikomi įvairiose šnekos atpažinimo sistemose. Sukurtieji akustiniai modeliai gali padidinti šnekos atpažinimo tikslumą.

Trumpos žinios apie autorių

1996–2002 m. studijos Vilniaus pedagoginio universiteto matematikos ir informatikos fakultete, gaunant matematikos bakalauro ir informatikos magistro laipsnį. 2003–2008 m. doktorantūros studijos Matematikos ir informatikos instituto Atpažinimo procesų skyriuje. Nuo 2005 m. autorė dirba Vilniaus pedagoginio universiteto Matematikos ir informatikos fakultete dėstytoja.

Sigita Laurinčiukaitė

ACOUSTIC MODELLING OF LITHUANIAN SPEECH RECOGNITION
Summary of Doctoral Dissertation
Technological Sciences, Informatics Engineering (07T)

Sigita Laurinčiukaitė

LIETUVIŲ ŠNEKOS ATPAŽINIMO AKUSTINIS MODELIAVIMAS
Daktaro disertacijos santrauka
Technologijos mokslai, informatikos inžinerija (07T)

2008 05 06. 1,5 sp. l. Tiražas 100 egz.
Vilniaus Gedimino technikos universiteto
leidykla „Technika“, Saulėtekio al. 11, LT-10223 Vilnius, http://leidykla.vgtu.lt
Spausdino UAB „Baltijos kopija“,
Kareivių g. 13B, 09109 Vilnius, http://www.kopija.lt