Framework for Choosing a Set of Syllables and Phonemes for Lithuanian Speech Recognition

Sigita LAURINČIUKAITĖ, Antanas LIPEIKA
Recognition Processes Department, Institute of Mathematics and Informatics
Goštauto 12, LT-01108 Vilnius, Lithuania
e-mail: sigita.lau@mch.mii.lt, lipeika@ktl.mii.lt

Received: September 2005

Abstract. This paper describes a framework for making up a set of syllables and phonemes that subsequently is used in the creation of acoustic models for continuous speech recognition of Lithuanian. The target is to discover a set of syllables and phonemes that is of utmost importance in speech recognition. This framework includes operations with lexicon, and transcriptions of records. To facilitate this work, additional programs have been developed that perform word syllabification, lexicon adjustment, etc. Series of experiments were done in order to establish the framework and model syllable- and phoneme-based speech recognition. Dominance of a syllable in lexicon has improved speech recognition results and encouraged us to move away from a strict definition of syllable, i.e., a syllable becomes a simple sub-word unit derived from a syllable. Two sets of syllables and phonemes and two types of lexicons have been developed and tested. The best recognition accuracy achieved 56.67% ±0.33. The speech recognition system is based on Hidden Markov Models (HMM). The continuous speech corpus LRN0 was used for the speech recognition experiments.

Key words: speech recognition, framework for formation of set of syllables and phonemes.

1. Introduction

Acoustic modelling of Lithuanian remains one of important tasks. Research in Dynamic time warping (DTW), which strove to solve the whole word recognition task (Lipeika *et al.*, 2002; Tamulevičius and Lipeika, 2004), 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 (Laurinčiukaitė, 2004). Implementation of a word-based recognizer is also simpler in comparison to sub-word-based recognizers.

Modern speech recognition systems for Lithuanian (Raškinis and Raškinienė, 2003a, 2003b; Filipovič and Lipeika, 2004; Šilingas *et al.*, 2004) employ phoneme-based recognition. These speech recognition systems are built according to the existing database resources, which have set of phonemes fixed apriori. Fortunately, many speech recognition researchers are involved in building speech corpora and can assist in the final selection of set of phonemes. The set SAMPA-LT of phonemes (Raškinis *et al.*, 2003) nearly matches the classification of sounds for Lithuanian and is preserved in the above mentioned recognizers. The fixed set of phonemes is used to find optimal system parameters (Raškinis and Raškinienė, 2003a; Filipovič and. Lipeika, 2004) or to investigate additional features such as stress, softness of consonants, and decomposition of mixed diphthongs into the basic set of phonemes (Šilingas *et al.*, 2004).

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. Fragmental research can be found in (Šilingas *et al.*, 2006). We present here a framework and results of syllable- and phoneme-based speech recognition for Lithuanian. The research covers the areas of word syllabification, formation of set of syllables and phonemes, description of recognition system, speech recognition experiments and comparison to the results, achieved by phoneme-based recognition.

2. Related Works

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. The set of speech sounds or speech sound combinations (units) depends on the investigator's choice. Speech units, according to the derivation rule, are obtained either by a linguistic criterion or by an automatic clustering technique (Černocký, 2002). Examples of speech units according to the automatic clustering technique are fenones (Bahl *et al.*, 1993), senones (Hwang and Huang, 1992), and multones (Bahl *et al.*, 1996). Speech units according to the linguistic criterion are common to all the languages: phonemes, diphthongs, syllables. To decide – which criterion to choose – is the matter of human confidence as both of them have advantages and disadvantages with respect to each other. The linguistic criterion prescribes to use sets of speech units obtained by language specialists or to extract sets of speech units according to the fixed grammar rules. The automatic clustering technique rests on algorithms that, supposedly, are able to extract sets of speech sounds from speech signals. The problem of the linguistic criterion is to map specific speech sounds to segments of a speech signal. At the same time, the problem of the automatic clustering technique-based criterion is mapping of speech signal segments to the common speech sounds.

In the past decade a dominant linguistic criterion used a phoneme or triphone as a unit. The reason is that all the phoneme-based languages (Lithuanian, German, French, English, etc.)[1] have finite sets of phonemes. Adjustment by adding some features enables us to use them for speech recognition. The lack of contextual information is circumvented by adding of triphones, the main advantage of which is to integrate coarticulation phenomena. In this case, a new problem is introduced by a large number of triphone models with poorly trained parameters due to the small amount of training data. The subsequent model clusterization reduces the amount of memory for model storage, but does not solve the main problem of model accuracy.

[1] Phoneme-based languages are the languages that have a fixed number of phonemes. Syllable-based languages (East Asian) such as Chinese have a fixed number of phonemes and syllables ($\sim$ 1200), but the writing system takes a syllable as a basic unit of representation.

It is essential to stress the role of a syllable as a sub-word unit in speech recognition. The motivation for using a syllable as a sub-word unit comes from human speech perception and generation studies. Appealing properties of a syllable are stability and duration, which makes a more precise reflection of contextual information possible. Stability and robustness for pronunciation variations of a syllable have been observed from the estimation of deletion rates of phonemes (12%) and syllables (1%) in the recognition process (Greenberg, 1998). On the other hand, major languages have large syllable sets, finite, but not enumerated. The number of syllable units to be used in recognition heavily depends on the lexicon size and it is not fixed. Definition of a set of syllables requires additional investigation. The syllable- and phoneme-based recognition was investigated for large vocabulary continuous speech recognition (LVCSR) (Ganapathiraju *et al.*, 1997), spontaneous speech recognition (Sethy *et al.*, 2003) and it shows recognition improvements against the triphone-based recognition. The syllable- and phoneme-based recognition has found its place in the recognition of names (Sethy *et al.*, 2002; Wang, 2004), and recognition of syllabic languages.

The syllable-based recognition has not been investigated for Lithuanian so far and we could find first attempts only in the works of students. It is important to compare the performance of different sub-word units in speech recognition and to introduce an alternative to the phoneme-based recognition.

3. Statement of the Problem

In this research work we sought to build a syllable-based continuous speech recognition system using the HTK toolkit (Young, 2002). Our aim was to start a syllable-based recognition research for Lithuanian. We investigated acoustic modelling without language models.

With the progress of research, we replaced the aim of building a purely syllable-based recognition system by building a syllable- and phoneme-based recognition system due to the large amount of syllables with an improper size of training material. Subsequently we diverged from a strict definition of a syllable and used sub-word units, derived from syllables. We named a new sub-word unit a pseudo-syllable. The use of pseudo-syllable in the lexicon transcription was corroborated by the improved speech recognition results. Our framework is referred to the construction of the basic set of pseudo-syllables and phonemes[2]. The framework proposed is shown in Fig. 1. 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.

The first block in the framework is syllabication of lexicon. Accomplishing of this task needs a syllabication tool. The tool was prepared by the algorithm, given in (Kasparaitis, 2005). The second block is correction of the ending of syllables and the inner part of syllables according to adjacent syllables. Automatically we corrected such artefacts as

[2] The basic set of syllables and phonemes is derived from the list of all syllables and phonemes, extracted in the third step. Finally it was prepared in the seventh step.

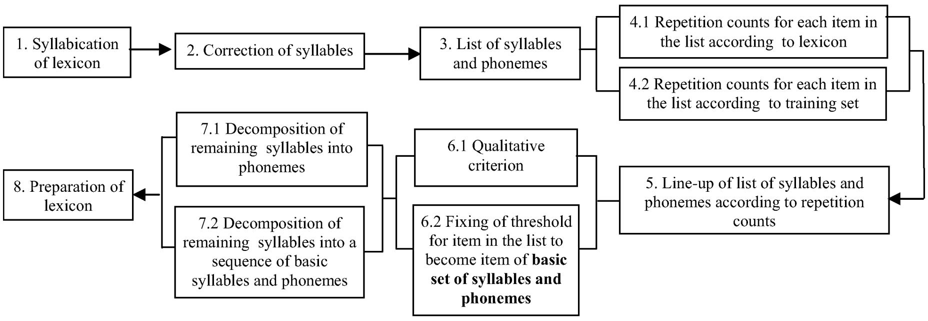
1. Syllabication of lexicon
2. Correction of syllables
3. List of syllables and phonemes
4.1 Repetition counts for each item in the list according to lexicon
4.2 Repetition counts for each item in the list according to training set
5. Line-up of list of syllables and phonemes according to repetition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

Fig. 1. Framework for syllable and phoneme set construction.

softness and voiceness / unvoiceness of unvoiced / voiced consonants. As a third step, we have to extract all the syllables and phonemes from the lexicon. The extracted set of phonemes is not complete. It is finally filled in seventh step. If the lexicon is large, usually one will be confronted with a long list of syllables. To check whether all of them are used in acoustic modelling of speech, we need to extract repetition counts of each syllable in the list according to the training set (as usual) or lexicon (proposed way). In the fifth step we line-up syllables and phonemes according to the repetition counts and in the sixth step we need to choose a criterion for selecting syllables for the basic set of syllables and phonemes, used further for acoustic modelling. At this point, two alternative ways are possible, i.e., the criterion link to the repetition count (the way we have chosen) or the criterion link to the structure of syllable (the way that was not very successful). The choice which repetition count should be fixed as a threshold is questionable. At this point we have fixed the repetition count and are going to examine it in the future. In the seventh step, the list of remaining syllables needs to be revised with a view to "destroy", i.e., decompose them. Since we have chosen linguistically defined units for modelling the ASR system, we will have to decompose the remaining syllables into phonemes. In addition, we tried another way – to decompose the remaining syllables into phonemes and syllables from the basic set of syllables and phonemes, giving preference to syllables. The eighth step adjusts the lexicon to the basic set of syllables and phonemes.

Following the given framework we established some sets of syllables and phonemes, lexicons and carried out recognition experiments, conditioning recommendations for further syllable- and phoneme-based recognition.

4. Description of the Speech Recognition System

The speech recognition system is based on HMM methods and has been built using the HTK toolkit and additional programs. The description includes formation of sets of syllables and phonemes, processing of speech records, construction of models, and set-up of modelling modes. First of all we specify training and testing procedures and introduce speech corpora, used for research.

4.1. *Training and Testing Modes*

System modelling consists of two modes: training and testing. The training mode accomplishes the model preparation task. The testing mode estimates models and the training process in terms of speech recognition accuracy. The performance of the recognition system was evaluated by word level accuracy defined as:

$$WA = \frac{N - S - I - D}{N} \times 100\%,$$

where N is the total number of words in the test set, 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.

The testing mode was based on the cross-validation principle, i.e., $n - 1$ subsets of data are used for training and one remaining subset of data is used for recognition. In our case, $n = 10$, because the speech corpus was divided into 10 nearly equal subsets, each of them approximately of one hour duration. One experiment consists of 10 training and testing cycles, i.e., 9 subsets ($\sim 90\%$) were used for training and one for testing ($\sim 10\%$). The average result of one experiment was calculated from 10 recognition results.

4.2. *Speech Corpus*

Experiments were based on the Lithuanian radio news speech corpus LRN0 (Lithuanian Radio News version 0). The corpus contains over 10 hours of speech records. Speech samples were recorded directly from the Lithuanian Radio broadcasting, first program (LR1). The content of records covers the most important political, economic and sport events of local and foreign areas. Each speech waveform file is provided with sentence level annotations. Specific marks, such as pauses, silence, breathing, and mispronunciations were included as well. The main characteristics of records are: the sampling rate – 11 kHz, resolution – 16 bits. The speech corpus is accompanied by words-to-phones transcription lexicon. It contains $\sim$ 18000 entries. Phonetic transcriptions and stress marks were created manually referring to (Vaitkevičiūtė, 2001; Keinys 2000). Semi-automatic lexicon transformations were carried out in the process of word syllabification. The corpus contains records of 23 male and female speakers, all Lithuanian Radio newsreaders speaking with a correct and clear pronunciation. The records of 10 speakers (4 females and 6 males) make 89% of the speech corpus.

4.3. *Formation of Syllable and Phoneme Sets, and Lexicons*

The set of syllables and phonemes is compiled following the framework given in Fig. 1.

The first step was **syllabification of lexicon.** Syllabification of words was implemented according to the algorithm description, given in (Kasparaitis, 2005). In short, we mention only the main principles of word syllabification and for deeper understanding please refer to the source, mentioned above. In Lithuanian, a syllable has a defined structure – a formula, expressed in a sequence of letters STRARTSK. Each of these letters represents a particular set of the alphabet:

S = {s, š, z, ž},

T = {b, d, g, k, p, t, c, č, dz, dž, ch, h, f},

R = {j, l, m, n, r, v},

A – any vowel or vowel combination,

K = {k, t}.

Syllabification starts from the end of a word by establishing to which set of the alphabet, represented by a letter in the formula, the letter belongs (note that we check each letter in turn from the end of the formula). No matter a letter belongs or not to the alphabet, we move on to the next one to the end of the formula. When we reach the end, the letters observed in the word compose a syllable. We proceed with syllabification by applying the formula to the remaining part of the word. The algorithm does not yield a completely correct syllabification at the points where the prefix meets the root or at the juncture of two words. These mistakes would have required additional, manual correction, but emphasizing the weight of automated work we have skipped these grammar rules. These assumptions resulted in a slight deviation of syllable repetition counts from the real ones, and examples of syllables that do not exist in the spoken language. For example, word *suimt* was syllabised as *suim-ti*, whereas the correct decomposition into syllables is *su-im-ti*. We can pay no attention to the latter problem due to the fact that these uncommon patterns are rare and do not enter the basic set of syllables and phonemes.

The step of **correction of syllable** involved a verifying effect of inter- and intra-syllable coarticulation of syllable sounds. The phenomenon of voiceness / unvoiceness of unvoiced / voiced consonants was verified for intra- and inter-syllable coarticulation. For the inter-syllable coarticulation, the softness of consonants was checked in addition. This block was introduced with a view to reduce the amount of syllables covering the same acoustic information.

The list of syllables and phonemes has been obtained in the third step.

The fourth step of the framework proposes two alternatives: to get **the repetition count for each item in the list using the lexicon or training data** (according to sentence level annotations). This step required to examine the amount of training material of each syllable and to judge its sufficiency. The first repetition result shows the usage level of a syllable in the syllabification process. Following it formation of the set of syllables and phonemes becomes simpler and leads to the model set dependency on the lexicon. The second result shows a real availability of training material for the models of syllable. The set of syllables and phonemes would result in the construction of a more accurate and data-dependent model set. With a view to rate the syllable-phoneme formation approach, both ways were applied and two sets of syllables and phonemes were formed. The basic set of syllables and phonemes in the fourth block, formed by the first way, was named H_1 (lexicon-dependent) and by the second way – H_2 (data-dependent). Since the research cross-validation principle was chosen for evaluation, there exist 10 training sets. Following the first way, we need to calculate the repetition counts only once and it will be valid for all the 10 training sets. Hence, the basic set of syllables and phonemes will be only one. Following the second way, we need to calculate the repetition counts for each of the 10 training sets. Thus, we will have 10 basic sets of syllables and phonemes. After forming 10 basic sets of syllables and phonemes, we compared all the pairs of 10 sets and noticed that all of them differ from one another within 4 syllables, i.e., all of them were quite similar.

The lists of syllables and phonemes were **lined up** according to the repetition counts in the fifth step. The sixth step was aimed at determining a criterion, according to which syllables would be selected for the basic set of syllables and phonemes. At this point, we have chosen to **fix a threshold for an item in the list to become the item of the basic set of syllables and phonemes**. For the set H_1 of syllables and phonemes, the threshold was 50 (Šilingas *et al.*, 2004). The final basic set H_1 consisted of 290 items (227 syllables, 63 phonemes and diphthongs). The sizes of all the 10 basic sets H_2 are approximately the same, which allows us to compare the recognition results of model sets H_1 and H_2. It is possible to change the size of the basic set by changing the threshold. We are sure it could improve the results.

Apart from the fixed number of syllables and phonemes that were moved to the basic set of syllables and phonemes, we have to consider the remaining part of syllables that did not exceed the threshold. The seventh **decomposition** step allows for two possibilities again: to decompose syllables in a sequence of phonemes or in a sequence of basic phonemes and syllables. These two approaches differ in the role which syllable units play in the subsequent lexicon design. Hence, this block applies to the lexicon design. Lexicons designed according to first way, are denoted by _P, and according to second way, by _SP. Following this way, a syllable is called a pseudo-syllable, because pseudo-syllables are not derived by the grammar rules.

The eighth step is a continuation of the seventh one. We have to modify the lexicon, obtained in the first step to match the basic sets of syllables and phonemes. For sets H_1 and H_2 we have designed the following lexicons: H_1P, H_1SP, H_2P, and H_2SP. Actually, the latter two lexicons represent only their formation patterns since each of the 10 sets H_2 has its own two lexicons, i.e., we had 10 lexicons of H_2P and 10 lexicons of H_2SP.

The derived sets of syllables and phonemes, and lexicons are summarized in Table 1.

4.4. *Model Parameters*

The next issue to be considered is definition of HMM structure parameters. The standard left-to-right model topology with no skips was used. The structural element that we focus here on is the number of states in a model[3]. Traditionally duration of a phoneme is expressed in 3 emitting and 2 non-emitting states. In line with this, all the models of phonemes and diphthongs in our experiments consisted of 5 states. The number of states in each syllable model was set after calculating the phonemes in each syllable, multiplying the obtained number of phonemes by 3, and adding 2 non-emitting states.

[3] Other elements such as distribution of observations in the states and state transition matrices resulting from the training process are not of interest to us.

Table 1

The sets of syllables and phonemes, and lexicons

Name of the set of syllables and phonemes / lexicons	Description
H_1	The set of syllables and phonemes of 290 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 $\sim$ 290 items, formed according to syllable repetition counts in training data.
H_1P	Lexicon, formed from the set H_1, non-basic syllables decomposing into sequence of phonemes.
H_1SP	Lexicon, formed from the set H_1, non-basic syllables decomposing into the sequence of basic phonemes and syllables.
H_2P	Lexicons, formed from the set H_2, non-basic syllables decomposing into the sequence of phonemes.
H_2SP	Lexicons, formed from the set H_2, non-basis syllables decomposing into the sequence of basic phonemes and syllables.

4.5. *Training and Testing*

The models were trained using 13th-order feature vectors of Mel Frequency Cepstral Coefficients (MFCC) and their delta and delta-delta values (these were 39-dimensional feature vectors), extracted from raw speech waveforms. The training process involved augmentation of the number of mixtures in each model and each state to 4, after performing training of each model. Testing was carried out after the fourth augmentation of mixtures. As a result we consider the recognition results achieved by acoustic models with 4 mixtures per state. A more thorough description how the work is organized by HTK can be found in (Šilingas and Telksnys, 2004).

5. Results and Discussion

Next we list the results of experiments, achieved while testing two earlier defined sets H_1 and H_2, and lexicons H_1P, H_1SP, H_2P, H_2SP derived for each of them. Recognition results, given in Table 2, are average result of 10 recognition results.

Recognition results, given in Table 2, suggest twofold discussion of the fourth (pattern of syllable repetition counts) and seventh blocks (decomposition pattern of non-basic syllable) in the framework of syllable and phoneme set formation, presented in Fig. 1. Having these results we are able to compare the calculation pattern of syllable repetition counts and decomposition pattern of a non-basic syllable.

First glimpse at the results shows advantage of set H_2, which was formed according to the repetition counts of syllables in training data, i.e., using path 2 in the fourth block (4.2). The recognition accuracy suffers approximately by 4–10%, if one chooses simplicity in building the set of syllables and phonemes according to path 1 (4.1). Taking into

Table 2

Recognition results for lexicon-dependent H_1, and data-dependent H_2 sets of syllables and phonemes and two lexicons for each set. The results are average of ten word accuracy (WA) results. Each WA was outcome of speech recognition of the acoustic model set with 4 mixtures per state

Name of the set of syllables and phonemes	Lexicon	Average results of WA with 95% confidence intervals
H_1	H_1P	46.96 ±0.37
	H_1SP	48.69 ±0.42
H_2	H_2P	53.08 ±0.22
	H_2SP	56.67 ±0.33

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 suits the requirements of speech recognition better than 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 dominance of lexicons _SP. Lexicons of this type increase the recognition accuracy by 2–4%. We could conclude that units of longer duration in the lexicon have a greater influence on the recognition accuracy.

6. Comparison of Syllable- and Phoneme-Based Recognition with Phoneme Based Recognition

In this section, we compare the recognition results, achieved for the same speech corpus, with different speech units, i.e., phonemes and syllables. Phoneme-based recognition has been conducted to discover the importance of additional features (softness, stress marks) to the phoneme set (Šilingas *et al.*, 2004). Five phoneme sets have been tested and 2 with the best recognition results were selected for further research. We took the phoneme set with the best recognition results from the experiment mentioned above. Using it we modelled a speech recognition system in the same training and testing mode as in the case of syllable- and phoneme-based recognition. The comparison results are given in Table 3. The results show that purely phoneme-based speech recognition surrenders to the syllable- and phoneme-based speech recognition by $\sim 5\%$.

All the results, obtained on this research set, are presented in Fig. 2.

7. Conclusions

The target of our research was to develop a framework for formation of the set of syllables and phonemes, and lexicon adjustment. The result output of the proposed framework is

Table 3

Recognition results for the set H_2 of syllables and phonemes, lexicon H_2SP and the phoneme set. The results are given as the average values of 10 word accuracy (WA) results. Acoustic models with 4 mixtures per state were tested

Syllable and phoneme / phoneme set	Size in models	Average results of WA with 95% confidence intervals
H_2	∼ 290	56.67 ±0.33
P_1		
Phoneme set with softness and stress marks, diphthongs	227	51.81 ±0.28

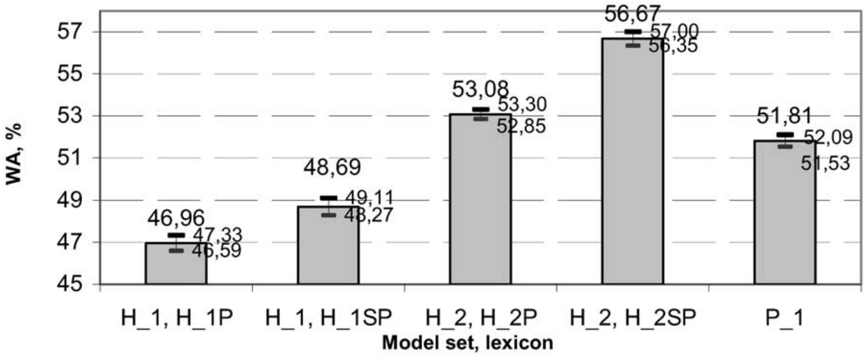

Fig. 2. Recognition results for the sets H_1, H_2 of syllables and phonemes, lexicons H_1SP, H_2SP and phoneme set P_1. The results are given in as the average values of 10 word accuracy (WA) results and confidence intervals at 95% level of confidence. Acoustic models with 4 mixtures per state were tested.

the set of syllables and phonemes, and lexicon used subsequently for Lithuanian speech recognition. The framework proposed has alternative at some places that are partly investigated in experiments. We analyzed the fourth and seventh blocks of the framework, i.e., the calculation pattern of syllable repetition counts and decomposition pattern of non-basic syllables. At this point we have identified that training data (average WA 53.08% for H_2, H_2P and 56.67% for H_2, H_2SP) are superior to the lexicon (average WA 46.96% for H_1, H_1P and 48.69% for H_1, H_1SP) when calculating the syllable repetition counts. At the same time, decomposition of non-basic syllables into sequences of basic syllables and phonemes (average WA 48.69% for H_1, H_1SP and 56.67% for H_2, H_2SP) are superior to that of non-basic syllables into phonemes (average WA 46.96% for H_1, H_1P and 53.08% for H_2, H_2P).

With a view to compare the syllable- and phoneme-based recognition and purely phoneme-based recognition, phoneme-based speech recognition has been modelled. The obtained results, i.e., 51.81% for purely phoneme-based recognition, show that syllable- and phoneme-based recognition is superior to the phoneme-based recognition.

In the future we are going to explore more blocks from the framework proposed with a view to enhance the accuracy of speech recognition.

References

Bahl, L.R., P.F. Brown, P.V. de Souza, R.L. Mercer, M.A. Picheny, (1993). A method for the construction of acoustic Markov models for words. In *Speech and Audio Processing, IEEE Transactions*. Vol. 1(4). pp. 443–452.

Bahl, L.R., J.R. Bellegarda, P.V. deSouza, P.S. Gopalakrishnan, D. Nahamoo, M.A. Picheny (1996). A new class of fenonic Markov word models for large vocabulary continuous speech recognition. In *Proceedings of ICSLP 96*. Vol. 2. pp. 1077–1080.

Černocký, J. (2002). Units for automatic language independent speech processing. In *Proceedings of LREC 2002 – Workshop on Portability Issues in human Language Technologies*. Las Palmas, ES, ELRA. pp. 7–13.

Filipovič, M., A. Lipeika (2004). Development of HMM/neural network-based medium-vocabulary isolated-word Lithuanian speech recognition system. *Informatica*, **15**(4), 465–474.

Ganapathiraju, A., V Goel, J. Picone, A. Corrada, G. Doddington, K. Kirchhoff, M. Ordowski, B. Wheatley (1997). Syllable – a promising recognition unit for LVCSR. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*. pp. 207–214.

Greenberg, S. (1998). Speaking in shorthand – a syllable-centric perspective for uderstanding pronuncation variation. In *Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*. pp. 47–56.

Hwang, M.Y., X. Huang (1992). Subphonetic modeling with Markov states-Senone. In *IEEE International Conference of Acoustics, Speech, and Signal Processing*. ICASSP-92. Vol. 1. pp. 33–36.

Young, S., G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland (2002). *The HTK Book* (for HTK Version 3.2).

Kasparaitis P. (2005). Syllabification and hyphenation of word (in Lithuanian). *Lecture notes.* Reachable by Internet: `http://www.mif.vu.lt/~pijus/CL/cl.htm`

Keinys, S. (2000). *Lexicon of Modern Lithuanian*. 4th edition, Vilnius (in Lithuanian).

Laurinčiukaitė, S. (2004). On different kinds of speech units based isolated words recognition of Lithuanian language. In *Proceedings of the First Baltic Conference on Human Language Technologies: The Baltic Perspective*. Riga. pp. 139–143.

Lipeika, A., J. Lipeikienė, L. Telksnys (2002). Development of isolated word speech recognition system. *Informatica*, **13**(1), 37–46.

Raškinis, A., G. Raškinis, A. Kazlauskienė (2003). Speech assessment methods phonetic alphabet (SAMPA) for encoding transcriptions of Lithuanian speech corpora. In *Proceedings of Information Technology and Control*. Technology, Kaunas. **4**(29). pp. 52–55.

Raškinis, G. and D. Raškinienė (2003a). Parameter investigation and optimization for the Lithuanian HMM_based speech recognition system. In *Proceedings of Information Technologies*. Technology, Kaunas. pp. IX–41–48 (in Lithuanian).

Raškinis, G. and D. Raškinienė (2003b). Building medium-vocabulary isolated-word Lithuanina HMM speech recognition system. *Informatica*, **14**(1), 75–84.

Sethy, A., Sh. Narayanan, S. Parthasarthy. (2002). A syllable based approach for improved recognition of spoken names. In *Proceedings of the ISCA Pronuncation Modeling Workshop*. pp. 30–35.

Sethy, A., B. Ramabhadran, S. Narayanan (2003). Improvements in English ASR for the MALACH project using syllable-centric models. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '03)*. pp. 129–134.

Šilingas, D., L. Telksnys (2004). Specifics of hidden Markov model modifications for large vocabulary continuous speech recognition. *Informatica*, **15**(1), 93–110.

Šilingas, D., S. Laurinčiukaitė, L. Telksnys (2004). Towards acoustic modeling of Lithuanian speech. In *Proceedings of SPECOM'2004*. St Petersburg. pp. 326–333.

Šilingas, D., S. Laurinčiukaitė, L. Telksnys (2006). Technique for choosing efficient acoustic modeling units for Lithuanian continuous speech recognition. In *Proceedings of SPECOM 2006*. St. Petersburg. pp. 61–66.

Tamulevičius, G., A. Lipeika (2004). Dynamic time warping based speech recognition system. In *Proceedings of the First Baltic Conference on Human Language Technologies: The Baltic Perspective*. Riga. pp. 156–161.

Vaitkevičiūtė, V. (2001). *Fundamentals of Pronunciations for Lithuanian*. Lexicon, Vilnius, Pradai (in Lithuanian).

Wang, N.J.C., Tsai Ching-Ho, P. Huang, J.L. Shen (2004). Chinese large-vocabulary name recognition system using character description and syllable spelling recognition. In *Proceedings of International Symposium on Chinese Spoken Language Processing*. pp. 17–20.

S. Laurinčiukaitė received her master of science degree in informatics from the Vilnius Pedagogical University in 2002. She is a PhD student in Recognition Processes Department at the Institute of Mathematics and Informatics since 2003. Her research field is HMM based methods for Lithuanian speech recognition.

A. Lipeika is a doctor of technical sciences, a senior researcher of the Recognition Processes Department at the Institute of Mathematics and Informatics and a professor at the Mathematics and Informatics Department of Vilnius Pedagogical University. His scientific interests include processing and recognition of random processes, detection of changes in the properties of random processes, signal processing, speech processing, speech and speaker recognition.

Skiemenų ir fonemų aibės sudarymo struktūra lietuvių šnekos atpažinimui

Sigita LAURINČIUKAITĖ, Antanas LIPEIKA

Skiemenimis grįstas lietuvių šnekos atpažinimas yra nagrinėtas fragmentiškai, prioritetą skiriant fonemos kalbos vienetui, nėra susiklosčiusios skiemenų naudojimo tradicijos. Straipsnyje pristatoma skiemenų ir fonemų aibės sudarymo struktūra lietuvių šnekos atpažinimui. Pateiktoje struktūroje galimos kelios elgesio alternatyvos, kurių dalis yra tiriama eksperimentiškai. Aprašomas kelių skiemenų ir fonemų aibių, žodynų konstravimas. Išryškinamas skiemens vaidmuo žodyno skiemenavimo procese. Sukonstruotos ir testuotos dvi skiemenų ir fonemų aibės, du žodynai. Nenaudojant kalbos modelio pasiekiamas 56.67% ±0.33 atpažinimo tikslumas. Šnekos atpažinimo modeliavimas grindžiamas paslėptais Markovo modeliais (PMM). Tyrimams naudojamas Lietuvos radijo naujienų ištisinės šnekos garsynas LRN0.