ISSN 1392 – 124X INFORMATION TECHNOLOGY AND CONTROL, 2007, Vol. 36, No. 1 A

AVERAGED TEMPLATES CALCULATION AND PHONEME CLASSIFICATION

Kęstutis Driaunys, Vytautas Rudžionis, Pranas Žvinys

Department of Informatics, Vilnius university Kaunas faculty of humanities

Abstract. Modeling of acoustic processes has not been considered sufficiently in experimental research of phoneme classification. This paper presents phoneme template calculation method which is motivated by modeling of transitional and stationary phonetic processes and enables to represent each of the distinctive process by a single feature vector. Experimental study confirmed advantages of the proposed method. The best achieved phoneme recognition accuracy was 65.2% in our study. This result is by 26.4% better than the recognition accuracy achieved with phoneme representation composed from all frames composing phoneme. It was observed that the best vowel, semivowel and fricative consonant representation were got from 40-50 ms while best representations of plosive consonants were got from 30 ms.

Introduction

In recent years, hidden Markov models (HMM) are the dominant technology in speech recognition systems and almost all of commercial automatic recognition systems simulate acoustic-phonetic phenomena's using HMM [0]. Also, the search for alternative speech recognition methods is under way since further progress in recognition systems based on HMM is rather moderate. Phonemic recognition may be one of such supplemental approaches.

We call phonemic recognition method, which tries to locate explicitly boundaries of phonemes in speech signal, to find an appropriate description method and assign it to one of the phoneme classes using statistical classification methods. These recognition organization methods have several advantages [0]:

- Phonetically segmented utterance provides opportunity to use different classification methods for different phonemes, in other words, to use acoustic and phonetic knowledge gained from prior statistical analysis and other experiments;
- Phonemic recognition approach allows contextual and speaker related variations modeling and use them as a powerful tool for classification;
- Different types of features could be used to recognize different phonemes.

Despite of the advantages mentioned above the phonemic recognition approach is implemented rather seldom. One of the most important difficulties is the necessity for an efficient phonetic segmentation algorithm. The errors made in this stage of operation are very hard to neutralize later. So the number of studies oriented towards phonemic speech recognition is not big.

It could be seen that research devoted to phoneme recognition problems that most often attention has been paid to selection of classification algorithm [0], [0] or search for the features best suited to identify phoneme [0]. But in these studies little attention has been paid to analysis of acoustic structure of phoneme. Most often phoneme is modeled as part of fixed size speech signal [0] or all frames are used covering phoneme from the left boundary to right boundary. Traditional three emitting state left to right HMM is applied most frequently. Such a HMM configuration is based on the fact that phoneme consists of three distinctive parts: steady-state (middle) part and left as well as right contextual processes describing parts. This means that each state simulates approximately one third of a phoneme. But durations of different states of phonemes are not equal and it is unclear how good these durations are captured by HMM training. This is clearly seen in Viterbi training of HMM when a phoneme is decomposed into three equal parts where the first part represents dependency from previous phoneme or represents left context, the middle part is a stationary part which is not affected by context and the third part represents dependency from next phoneme or represents right context [0]. Phonetic events may be modeled sufficiently well using rule based methods but, applying them for phoneme classification, the number of rules and acoustic features grows essentially and causes a significant increase in error rate. Also, implementation of acoustic knowledge about phonetic units requires well-established theoretic

background, which is in incipiency for Lithuanian sounds.

Several studies have been performed aiming to model phoneme templates, which try to evaluate acoustic processes occurring inside phoneme. Wu with colleagues [0] in their experimental research have shown that maximum likelihood (ML) based frame selection, which selects reliable frames from a high resolution along the time axis, helps to improve the discrimination between phonemes. In further papers they present research on single frame selection for a phoneme classification task. This method selects one frame for one state in an (HMM). This technique takes likelihoods of frames and their positions in a phoneme segment into account at the same time, and selects very few frames to represent the spectral evolution of the phoneme. Experiments show that phoneme model trained by selected frames is more discriminative than a model using all frames [0].

Hosom and Cole [0] carried out experiments trying to model diphones according to their duration. If duration is less than 120 ms, then phoneme is subdivided into two parts. If duration is longer, then phoneme is subdivided into three parts taking 60 ms from the left boundary of the phoneme and 60 ms from right boundary of the phoneme while other part represents stationary part. These parts were excluded when modeling diphones. Such a phoneme decomposition approach allowed increasing phoneme recognition accuracy from 44 percent to 48.4 percent. The study showed that more precise modeling of phoneme parts is perspective seeking to increase the accuracy of phonemic recognition.

In this paper we will present a method to derive phoneme template, which is motivated by the modeling of acoustic stationary, and transitional processes observed in the phoneme. The method uses a single feature vector to describe each of the processes.

1. Theoretical framework

It could be observed analyzing oscilograms and spectrograms of speech signal that each phoneme has three distinctive parts. Figure 1 shows the oscilogram, spectrogram and phoneme boundaries of the word *nulis*. It could be seen relatively stationary parts of phonemes (spectral characteristics vary slowly) and more dynamic contextual parts which emerge on both sides of phoneme at the boundaries between two different phonemes.

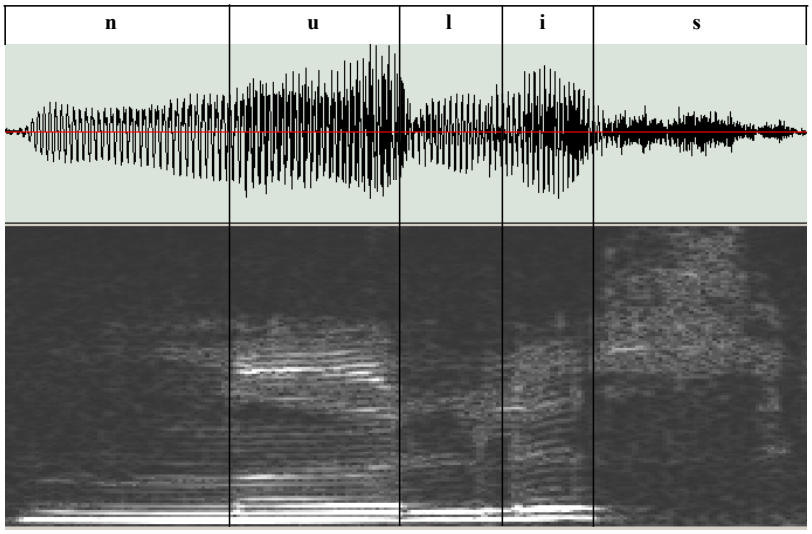


Figure 1. Oscilogram and spectrogram of the word "*nulis*"

Trying to describe and exploit those three parts we need to prepare three templates for each phoneme. The stationary part is described by locating the center of the phoneme and taking a part of the signal (in example, 20 ms to both sides from the center). Let us assume that most characteristic properties of phoneme concentrate in this central part. We can find more accurate intervals experimentally maximizing recognition accuracy.

We describe parts of phoneme influenced by context by taking fixed size part of the signal from both the left and right boundary. Let us assume that most valuable information about context of phoneme concentrates in the transition.

The algorithm presented below shows template calculation for the *i-th* feature. Here:

phoneme templates: *LC* – left contextual part of the template, *SP* – stationary part of the template, *RC*– right contextual part of the phoneme;

ELn – number of elements in feature vector.

Let's determine the variable *TFn*, which shows the size of part of the phoneme or the number of frames used to get the template.

Let's determine the variable *PhFn* which shows the size of the phoneme or how many frames cover the whole phoneme

If $PhFn <= TFn$ **then**

$$\begin{cases} LC_i = \dfrac{1}{PhFn}\sum_{k=1}^{PhFn} x_{i;k} \\ SP_i = \dfrac{1}{PhFn}\sum_{k=1}^{PhFn} x_{i;k} \ \ where\ i = 1,...,ELn \\ RC_i = \dfrac{1}{PhFn}\sum_{k=1}^{PhFn} x_{i;k} \end{cases} \quad (1)$$

else

$$\begin{cases} LC_i = \dfrac{1}{TFn}\sum_{k=1}^{TFn} x_{i;k} \\ SP_i = \dfrac{1}{TFn}\sum_{k=\frac{PhFn}{2}-\frac{TFn}{2}}^{\frac{PhFn}{2}+\frac{TFn}{2}} x_{i;k} \ \ where\ i = 1,...,ELn \\ RC_i = \dfrac{1}{TFn}\sum_{k=PhFn-TFn}^{PhFn} x_{i;k} \end{cases} \quad (2)$$

end

Algorithm 1: *Phoneme template calculation.*

The above described algorithm is illustrated in example shown in Figure 2. This example shows how templates are formed for phonemes of different durations. If the length of phoneme is shorter than defined length of phoneme *TFn* (phonemes *I* and *L* in the figure), then all templates are derived from the same frames of speech signal (both stationary and contextual parts). If phoneme length is bigger than defined size of phoneme (*N* and *S* in Figure 2), then the part of speech signal is ignored making assumption that this part has not enough important information for recognition.

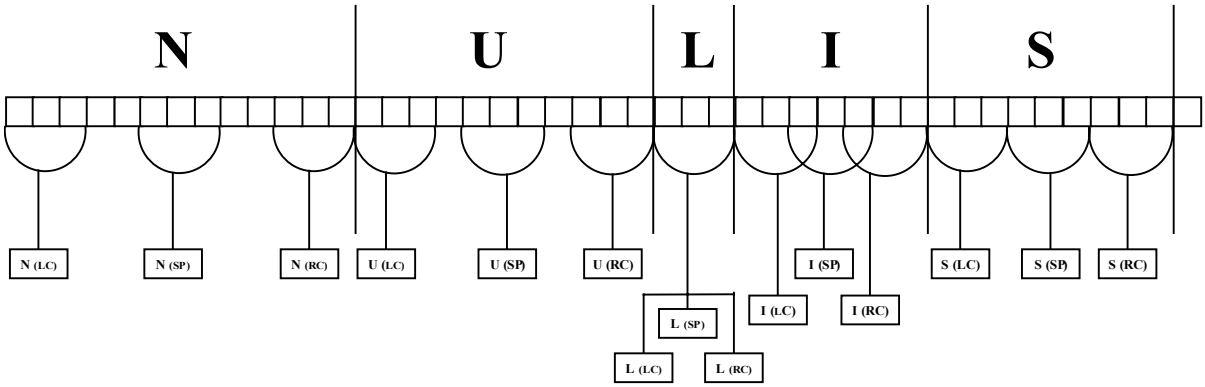


Figure 2. Example of phoneme template formation

This phoneme-modeling algorithm enables to reduce the amount of processed information using for classification only stationary part (acoustic models of monophones) and, at the same time, maintain important contextual information. If phoneme is classified to some class with low probability level, then it is possible additionally to introduce classification of contextual parts with the hope to increase recognition accuracy.

This method allows us to expect that best results of vowels classification will be achieved using templates formed from the stationary part. Otherwise by the analyzing structure of plosive consonants we could conclude that the most important acoustic events concentrate near the right boundary of phoneme where a burst occurs. So for recognition of plosive consonants it may be advisable to use a template formed from right context. It was expected to check these assumptions experimentally.

2. The experiment

Speech signal has been described using mel frequency cepstrum coefficients (MFCC) which are the most popular technique in speech recognition recently. The mel frequency cepstrum coefficients were obtained using HTK 3.0 package. While calculating MFCC coefficients there were used the following parameters: the length of the window – 16 ms, the step – 6.25 ms, 20 filters from the bank of Mel filters and 12 cepstrum coefficients. The feature vector contains 12 MFCC coefficients, energy and their first and second derivatives. Totally 39 elements were derived for each signal frame.

Phoneme is decomposed into three parts during template formation. Taking middle or stationary part and calculating average value of each feature vector element per all frames of this part we obtain the template of phoneme stationary part. Later we took a part of the predefined size (number of frames) speech signal from the left boundary of the phoneme and calculated the average of each feature vector element in this part. The template of left context of the phoneme was formed in this way. Similarly the right context template was formed also. A series of 8 experiments were carried out to define the number of frames that provides the best phoneme recognition accuracy.

The classifier was trained deriving averages (μ) and covariance matrix (Σ) for each feature of template class observations:

$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$, where x_i is parameters in the feature vector, N is number of frames (3)

$$\sum_{et} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)(x_i - \mu)^T . \quad (4)$$

During classification the Mahalanobis distance between tested phoneme feature vector and each template feature vector was calculated:

$$g_{et} = (\mu - \mu_{et})' \Sigma_{et}^{-1} (\mu - \mu_{et}),\ et = 1,...,ET , \quad (5)$$

where ET is the number of templates or phonemes used for recognition.

During decision making, the phoneme is assigned to the class with the closest Mahalanobis distance:

$$\hat{Ph} = \arg\max_{et} g_{et} . \quad (6)$$

Experimental study of phoneme recognition was performed using LTDIGITS [0] speech corpora. Utterances of 100 speakers (50 male and 50 female) were used during experiments. Training and testing sets were formed according to the data presented in Table 1, trying to obtain speaker independency.

Table 1. Sizes of testing and training samples

Step	Test set	Training set
1	M001-M010; F001-F010	M011-M050; F011-F050
2	M011-M020; F011-F020	M001-M010; M021-M050; F001-F010; F021-F050
3	M021-M030; F021-F030	M001-M020; M031-M050; F001-F020; F031-F050
4	M031-M040; F031-F040	M001-M030; M041-M050; F001-F030; F041-F050
5	M041-M050; F041-F050	M001-M040; F001-F040

At the first stage phoneme testing was comprised from the first ten LTDIGITS male (M001-M010) and first ten female (F001-F010) speakers utterances while other 80 speakers data were used for training. In following experiments data were rearranged according to the rules presented in Table 1. The covariance matrices of features were derived using together male and female utterances data.

3. Results

A total of eight experiments were carried out in order to evaluate the influence of stationary part and contextual (left and right) part lengths and to find best representation of a particular phoneme. In seven experiments phoneme parts were formed from fragments of different length. In the last case, the template values were calculated using all the windows constituting a phoneme. All the classification experiments were performed using the same set of 25,553 phoneme utterances. Phoneme parts consisting of 2, 3, 4, 5, 6, 7 and 8 analysis frames were tested. These cases could be expressed in milliseconds - 22.25 ms, 28.5 ms, 34.75 ms, 41 ms, 47.25 ms, 53.5 ms and 59.75 ms, respectively. In order to substantiate the expedience of splitting into three parts an additional experiment was carried out where phoneme template was derived from all frames covering phoneme. Detailed confusion matrices of each experiment were presented in [0]. Summarized results, reflecting phoneme recognition accuracy in percentages, are provided in Table 2. Here LC (left context) denotes the phoneme correct classification accuracy when only the template values of the left part of phonemes were used for training and testing; SP (stationary part) and RC (right context), respectively, means that in experiments when only the fragments of stationary part or right context of phonemes were used and Wph means that experiment was performed using all frames covering phoneme.

Phoneme recognition accuracy was evaluated using an algorithm, which compares phoneme transcriptions with outputs of classifier and provides classification accuracy. Phoneme Recognition Accuracy (*PhRA*) was calculated using the following formula:

$$PhRA = \frac{C}{N} * 100\% , \quad (7)$$

where C is the number of correctly recognized phonemes, N – number of phonemes used for testing.

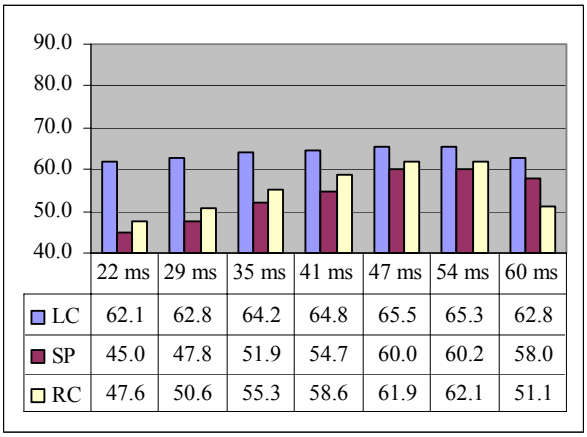
Table 2. Dependency of PhRA on the size of part phoneme used to derive the template (in percent)

	22 ms	29 ms	35 ms	41 ms	47 ms	54	60 ms	Wph
LC	60.2	62.0	63.8	64.6	65.2	63.1	61.8	38.8
SP	48.5	50.6	53.9	55.6	58.5	59.2	56.4	
RC	55.7	58.0	60.8	61.7	62.0	60.3	48.9	

As can be seen, the best results were achieved when templates of phonemes were formed from 5-6-7 frames. This means that each phoneme part was modeled by fragments of 40-50 ms. One can draw a conclusion that splitting a phoneme into three parts is superior comparing with the recognition when templates formed from all the frames of a phoneme were used. The best results of phoneme recognition were attained when phoneme templates were formed from 6 frames of left context parts of phoneme (65.2%).

Analyzing phoneme recognition accuracy dependency on the part of phoneme used for training and testing, we see that best results were obtained by performing phoneme classification using templates of the left part of the phoneme. This shows that important information for recognition concentrates in stationary part of phoneme and at the beginning of phoneme.

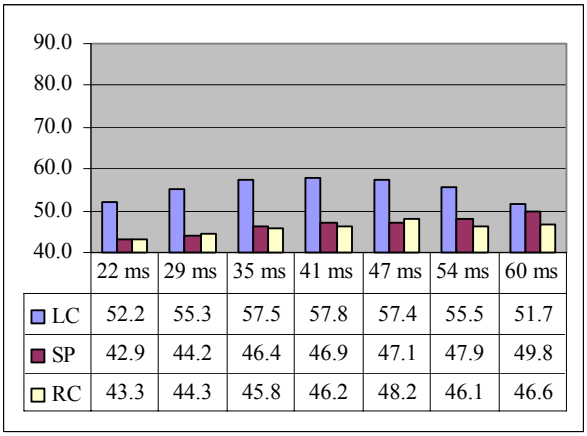
Further we will discuss *PhRA* for different groups of phonemes as a dependency from the size of part of phoneme used to derive the template.



	22 ms	29 ms	35 ms	41 ms	47 ms	54 ms	60 ms
LC	62.1	62.8	64.2	64.8	65.5	65.3	62.8
SP	45.0	47.8	51.9	54.7	60.0	60.2	58.0
RC	47.6	50.6	55.3	58.6	61.9	62.1	51.1

Figure 3. PhRA of vowels as dependency on the number of frames used to derive template

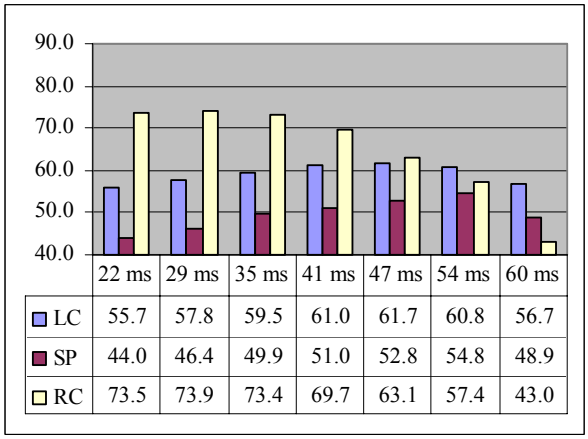
Looking at the *PhRA* results of vowel recognition (Figure 3) we could see that the highest accuracy (65.5%) has been obtained using template formed from 6 frames of left context. The template formed from 7 frames of stationary part of phoneme provides 60.2% accuracy. The template formed from 7 frames of right context provides the best accuracy in this case (62.1%). This results show that vowels have a dynamic structure and left context modeling is particularly important while significant information is concentrated in right context as well.



	22 ms	29 ms	35 ms	41 ms	47 ms	54 ms	60 ms
LC	52.2	55.3	57.5	57.8	57.4	55.5	51.7
SP	42.9	44.2	46.4	46.9	47.1	47.9	49.8
RC	43.3	44.3	45.8	46.2	48.2	46.1	46.6

Figure 4. PhRA of semivowels as dependency on the number of frames used to derive template

Similar situation is observed looking at the *PhRA* results of semivowels classification (Figure 4). The best results in this class of phonemes were achieved using templates formed from 4-6 frames of left context (average recognition accuracy over 57.4 %) while, using templates composed from only stationary part or right context, the average semivowels recognition accuracy is below 50%. Also, it should be noted that in all experiments the best semivowel recognition accuracy was achieved using templates formed from left context. Semivowels classification accuracy falls significantly when the length of left context becomes larger while lengthening of right context does not give so rapid decrease in accuracy. Lengthening stationary part we see a stable increase in recognition accuracy up to 49.8%. This fact is mainly caused by significant increase in recognition accuracy of phoneme *n*. Since duration of phoneme *n* is longest among semivowels [0] lengthening the part of signal used to form the template leads to higher recognition accuracy while recognition of other semivowels at the same time deteriorates. It looks that it is sensible to separate nasals m and n from other semivowels. Semivowels recognition accuracy was lowest comparing with other groups of consonants in these experiments.



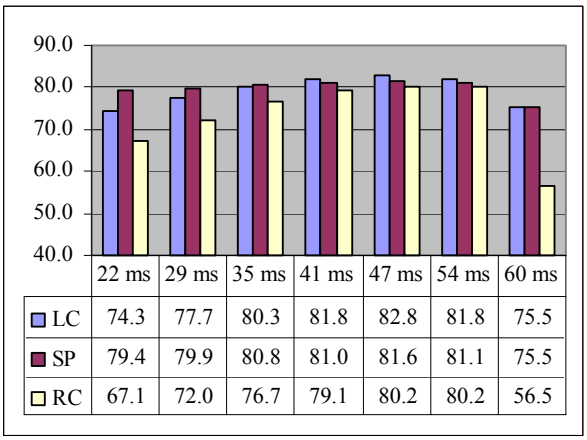
	22 ms	29 ms	35 ms	41 ms	47 ms	54 ms	60 ms
LC	55.7	57.8	59.5	61.0	61.7	60.8	56.7
SP	44.0	46.4	49.9	51.0	52.8	54.8	48.9
RC	73.5	73.9	73.4	69.7	63.1	57.4	43.0

Figure 5. PhRA of plosive consonants as dependency on the number of frames used to derive template

The assumption that most of the important information to recognize plosives concentrates at the right context was confirmed by our study. Figure 5 shows those best plosive consonants classification results were obtained using template formed from 3 frames of right context – 73.9%. Experimental study shows that *PhRA* increases when reducing the number of frames used to derive template. This agrees with the fact that acoustic event of plosive – burst – is very short and lengthening part of phoneme used to derive template introduces additional unhelpful artifacts.

We also could observe similar situation as with vowels and semivowels comparing recognition accuracy using templates derived from the stationary part and left context. Here contextual templates have superiority over stationary templates. The best results with templates derived from the left context were obtained using 6 frames (61.7%) while the best results achieved using templates derived from 7 frames of stationary

part were (54.8 %). Summarizing we can say that recognizing plosive consonants particular attention should be paid to the right context.



	22 ms	29 ms	35 ms	41 ms	47 ms	54 ms	60 ms
LC	74.3	77.7	80.3	81.8	82.8	81.8	75.5
SP	79.4	79.9	80.8	81.0	81.6	81.1	75.5
RC	67.1	72.0	76.7	79.1	80.2	80.2	56.5

Figure 6. Fricative consonants PhRA as dependency on the number of frames used to derive template

Summarized recognition accuracy results of fricative consonants *PhRA* are presented in Figure 6. As could be seen, the stationary part is more important for fricatives. But absolutely best classification results (82.8 %) were obtained using templates derived from 6 frames of left context. It should also be noted that templates derived from the frames provided best results using stationary part, left and right context for fricatives.

We can see that fricative consonant recognition accuracy was highest in our experiments comparing with recognition accuracy in other phoneme groups.

4. Conclusions

The performed experiments confirmed the assumption that phoneme modeling using left context, stationary part and right context separately has advantages over phoneme modeling as a single unit. The best recognition accuracy of speaker independent recognition of phonemes from LTDIGITS corpora was 65.2%. *PhRA* research showed that the most appropriate way to form phoneme description template is using 5–6–7 frames or approximately 40–50 ms speech signal. The best recognition accuracy for vowels, semivowels and fricative consonants was obtained using templates derived from the left context data. The best recognition results for plosive consonants were achieved using phoneme description derived from the right context which duration was 20-30 ms.

References

[1] **A.M. Abdelatty Ali, J. Van der Spiegel, P. Mueller, G. Haentjens, H. Berman H.** An Acoustic-Phonetic Feature-based System for Automatic Phoneme Recognition in Continuous Speech. *Proceedings of IEEE International Symposium on Circuits and Systems* (ISCAS-99), 1999, III-118 - III-121.

[2] **M. Antal.** Speaker Independent Phoneme Classification in Continuous Speech. *Studia Univ. Babes – Bolyai. Informatica. Vol.*XLIX, *No*.2. 2004, 55 – 64.

[3] **O. Dekel, J. Keshet, Y. Singer.** An Online Algorithm for Hierarchical Phoneme Classification. *Workshop on Multimodal Interaction and Related Machine Learning Algorithms* (MLMI 2004).

[4] **K. Driaunys.** Analysis of Lithuanian Spoken Language Labeling and Phonemic Recognition using LTDIGITS corpora. *PhD thesis, Vilnius University, Vilnius*, 2006, (*in Lithuanian*).

[5] **J.P. Hosom, R.A. Cole.** A Diphone-Based Digit Recognition System using Neural Networks. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (ICASSP 97). *Vol*.4, 1997, 3369-3372.

[6] **J.R. Ye, J. Povinelli, M.T. Johnson.** Phoneme classification using naive bayes classifier in reconstructed phase space. *Proceedings of IEEE Signal Processing Society* 10*th Digital Signal Processing Workshop*, 2002, 37-40.

[7] **A. Rudžionis, V. Rudžionis, P. Žvinys.** Lithuanian Speech Corpora. *Information sciences*, ISSN 1392-0561, *No*.17, 2001, 77–84, (*in Lithuanian*).

[8] **V. Rudžionis.** Consonant discrimination in consonant-vowel diphones. *Information sciences*, ISSN 1392-0561, *No*.9, 1998, 47–54, (*in Lithuanian*).

[9] **T.Y. Wu, D.Van Compernolle, J. Duchateau, H.Van Hamme.** Maximum likelihood based temporal frame selection. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Toulouse, France, May* 2006, 349-352.

[10] **T.Y. Wu, D.Van Compernolle, J. Duchateau, H.Van Hamme.** Single frame selection for Phoneme Classification. *Proceedings of. International Conference on Spoken Language Processing, Pittsburgh, USA, September* 2006, 641-644.

Received March 2007.

DOI: 10.5755/j01.itc.36.1.11817