LITHUANIAN SPEECH DATABASE LTDIGITS

Algimantas Rudzionis*, Vytautas Rudzionis**
Kaunas university of technology*, Vilnius university**
* Studentu 65-108, Kaunas, Lithuania
** Muitines 8, Kaunas, Lithuania
alrud@mmlab.ktu.lt

The Lithuanian speech database LTDIGITS was developed. Some details of this database could be of more general interest. These features are related with collected set of nasal consonant realizations in different vowel contexts. First, LTDIGITS contains nasal – vowel syllables where nasal is before open, middle and closed vowels. Second, the database includes special continuous phrase with above mentioned nasal – vowel syllables. These nasal-vowel pairs are in the stressed positions in the front of short 2 – 3 syllable words. Third, both the utterances to words and word to phones marking and labeling procedures were applied and are presented here.

I. INTRODUCTION

It is well known fact that speech corpora are of crucial importance for speech technology development. Work to develop methods and approaches for speech corpora has been started nearly two decades ago (Fisher, 1986). One of the first systematic speech signal databases was TIMIT. It became standard tool for speech technology research worldwide. Other popular speech signal databases used in various countries are YOHO, OGI Spelled and Spoken Words (Cole, 1992), Switchboard, etc.

Another well-known factor influencing progress in speech technology is incorporation of specific properties of particular language. Speech corpora LTDIGITS have been developed. It contains patterns of Lithuanian speech pronounced by native Lithuanian speakers. These data are Lithuanian digits (0-9) and words that could be used to control computer or some other devices (e.g. home appliances). But some details of LTDIGITS could be of more general interest. First, this corpora contains nasal – vowel syllables before open, middle and close vowels. Second, the database includes special continuous phrase with nasal – vowel pairs that are located in the stressed positions in the front of short 2 – 3 syllable words. Third, both the utterances to words and words to phones marking and labeling procedures were developed and are presented at conference.

Importance of acoustic realizations of nasal consonants in various vowel environments lies in the fact that better discrimination of phonemes should lead to significant progress in speech recognition. Nasal consonants form one of the most difficult groups of phonetic units to discriminate. Also it is well known that discrimination peculiarities are strongly influenced by the vowel in the consonant context: it is easier to recognize consonant in open vowel context than in the closed vowel.

We believe that CV clusters will provide capabilities to carry out deeper analysis of phoneme discrimination problems including multilingual peculiarities. E.g., it is meaningful to compare results from OGI telephone quality database where methods with strong discrimination capabilities weren't applied (Loizou & Spanias, 1996; Chengalvarayan & Deng, 1998) with results achieved using discriminant methods applied on other data (Ayer, 1993). Also methods used in our earlier works are worth to explore better (A.Rudzionis & V.Rudzionis, 1999) where significant phoneme discrimination error rate reduction using carefully selected features and discriminating methods has been observed. Analysis of E – set phonemes could be promising using similar CV and VC cluster data.

II. LTDIGITS CHARACTERISTICS

LTDIGITS has been created as the Lithuanian speech database and as a tool that could be used for analysis of discrimination of nasals in various vowel environments. LTDIGITS contains:

- the continuous Lithuanian digits strings – 5 utterances for each speaker in random order;
- the isolated Lithuanian control words – one utterance for each speaker;
- the syllables CV, where consonant C denote two nasals M, N before the open vowel A, middle U and the closed vowel I – 1 utterance for every speaker
- the continuous Lithuanian phrase MI_ MA_ NI_ MU_ NA_ NU_ speaker where the mark _ shows the final part of the word – 1 utterance for each speaker.

The recordings were collected from 225 female and 125 male speakers in the laboratory (acoustic camera) conditions. Listening control of each pronunciation has been used immediately after recording session. Speaker has been asked to repeat those phrases that were recorded improperly. The isolated Lithuanian control words are similar as in the various speech databases. These are such words as *start*, *stop*, *pause*, *resume*, etc. Each speaker pronounced 5 Lithuanian digit strings containing 5 digits. Digits in the string were selected randomly. Speech data were recorded using 16 kHz sampling rate.

Also LTDIGITS contains realizations of nasal consonants M and N in different acoustic contexts. These consonants has been recorded in continuous sentence and pronounced in isolated syllables that contain vowel and nasal. These CV

syllables contains nasals M, N before the contrastive open, middle and closed vowels A, U, I. The pairs of English words "might – night", "moon – noon", "meat – neat" could be examples where two nasals M, N precede vowels A, U, I.

The Lithuanian continuous phrase *MIikas MAto NIshoje MUsu NAmo NUmeri* also contains all above mentioned nasal – vowel pairs in stressed initial positions in the short 2 – 3 syllable words. All words in this sentence are meaningful.

In some sense this database supplements telephone quality OGI corpora where the similar inverse clusters VC (vowel – consonant) were collected. So the analysis extensions of phoneme cluster nasal – vowel are possible. Such data could be convenient tool to look for alternative ways to discriminate phonetic units. The regularized discriminant analysis (A. Rudzionis & V. Rudzionis, 1999; Rudzionis, 1998) could be one of the alternatives for deeper understanding of the phoneme discrimination problems.

Both the phrases to words and word to phone marking and labeling procedures were developed. The visual inspection of the signal, energy, sonogram were implemented together with the audio control. The labeling results are saved in TIMIT style (Lamel,1987) and in the modified formats where only word/phone boundaries are kept.

Most important options of speech database processing tools are: go to next stage option (the action is completed right), go back (it is necessary to repeat the previous action) or error (too many problems, the procedure should be interrupted without fixing any of the marking/labeling data and the new marking/labeling trial has to be completed). Next chapter describes in detail tools for LTDIGITS corpora processing.

III. PHRASE/WORD MARKING AND LABELING PROCEDURES

3.1. GENERAL PRINCIPLES

Possibilities to use speech corpora increases significantly if speech data is processed: there are determined and marked bounds of words in the phrases (sentences) and bounds of phones in the words. Below are presented our procedures for phrase to words and word to phones labeling that were used to process LTDIGITS (at least in part).

3.2. PHRASE TO WORD SEGMENTATION

This stage assumes location of word boundaries in the phrase. If phrase contains spontaneously pronounced sequences of words sometimes occur situations when proper marking of such boundaries isn't plain action. Actions performed during phrase to word segmentation could be summarized as follows:

- Positioning of marker before the first word
- Labeling of first word
- Alternative continue – repeat procedure (while marking has been done correctly);
- Marker positioning recursion (before the word start)
- Recursion of word boundaries labeling.

Start of phrase to word labeling procedure

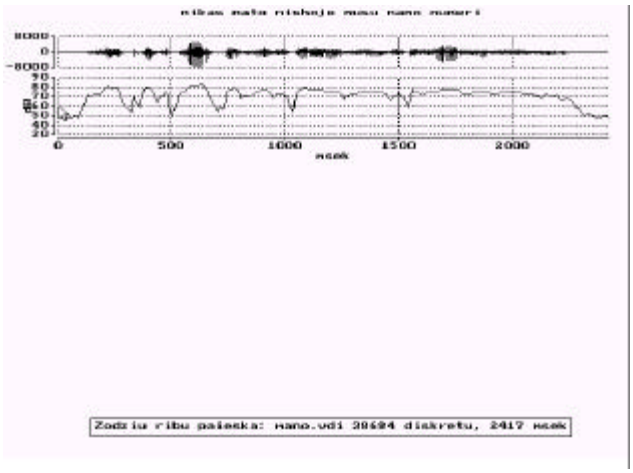

Figure 1: The phrase to word segmentation (1st step). There are two pictures for the whole phrase. From top: a) speech oscilogram, b) energy and mouse mark that should be placed before the first word in the phrase. There are displayed the phrase orthography (at top) and the speech file length (at bottom).

- In the upper part of Figure 1 we see orthographic description of phrase "Mikas mato niðoje mûsø namo numerá": specific Lithuanian symbols aren't used but in general orthographic not phonetic characteristics are expressed;
- signal oscilogram below;
- third graph shows energy level of phase in dB and time marks are presented in msec. At the left we see cursor which should be placed at the start of the considered word;
- in the bottom we see name of file, duration of phrase in samples and msecs.

Performed actions: cursor seen in the left of the picture must be placed just before the start of considered word. In this case cursor must be placed before the beginning of first word.

Labeling of first word boundaries

- In the upper part of Figure 2 we see text of phrase, signal oscilogram and energy curve (similarly as in first step). But in energy curve we see vertical line, which shows approximately start of first word;
- in the next graph we see oscilogram of signal which length is one sec (from the point selected in previous phase). It allows better orientation when performing following steps;

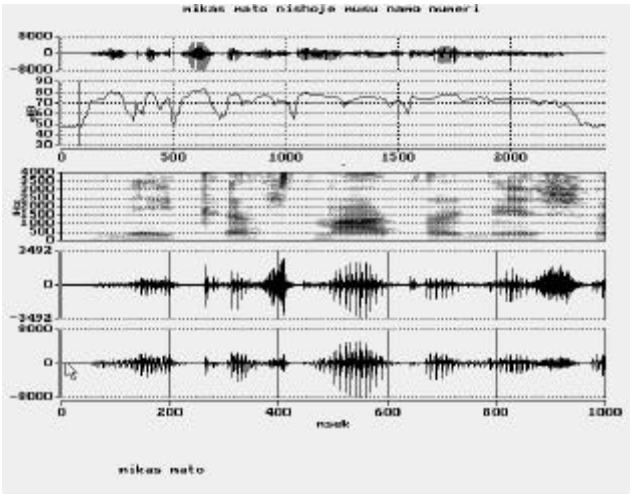

Figure 2: The phrase to word segmentation (2nd step). The two top pictures are the same as before: a) speech oscilogram, b) energy and marked by mouse line before the first word in the phrase. The three bottom pictures belong to word region: c) sonogram, d) preemphasized speech oscilogram, e) speech oscilogram. Orthography of the current (first) and the following (second) words is displayed in the bottom of this picture.

- in the next two graphs we see two oscilograms of the same part of signal. In the first oscilogram higher frequencies are emphasized (e.g., it is easier to determine end of phone ‘s’ in the first word “mikas”);
- in the left side of latter oscilogram cursor could be seen. It should be placed accurately at the start of word and then at the end of word;
- textual note “mikas” could be seen at the bottom of picture. It means that we are trying to determine boundaries or word “mikas”. After that next word “mato” should follow.

Performed actions: first accurately mark the beginning of word with cursor, then mark the end of word.

Labeling of consequent words in the phrase

- in picture (Figure 3) we see same view as in previous picture except bottom oscilogram;
- in the bottom oscilogram we see interchanged colors of background and signal curve. This shows that boundaries of these words has been determined;
- computer plays continuously selected part of signal. **This is very efficient procedure which allows almost precisely determine if word boundaries has been located correctly;**
- in text row below could be seen caption “if okay – q, else – repeat, if error – d“ which indicates three possible actions depending on the results of current bound labeling;
- in the bottom of this picture we see record “mikas mato“ as in the previous phase as well as digital representations of word boundaries.

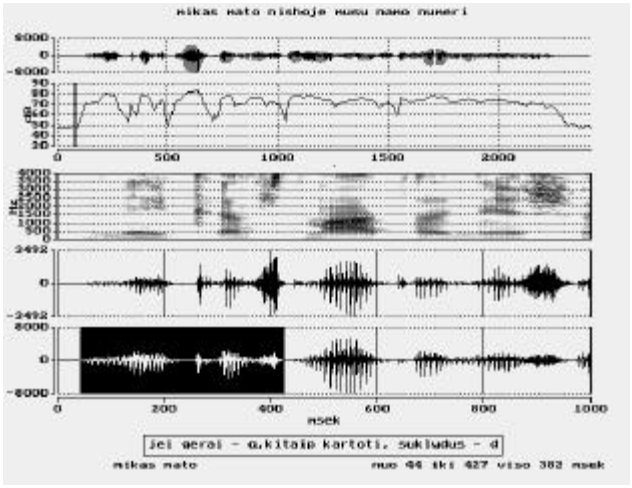

Figure 3: The phrase to word segmentation (3rd step). The pictures are the same as in Figure 2. The marked first word is highlighted and continuously played. Additionally there are displayed options at the bottom: q- continue, d- escape program or repeat previous step (any other key).

Performed actions:

- if by listening we decide that word boundaries has been located properly press key q (or Q) and move to the next phase (location of next word bounds);
- if by some reason we decide that serious mistake has been made then press key d (or D). The program stops, results of current labeling aren’t saved and marking should be performed once more; if determined bounds are suspicious then press any key and return back to the previous phase (it means to repeat labeling of the same word once more).

Marker positioning recursion (before the first word)

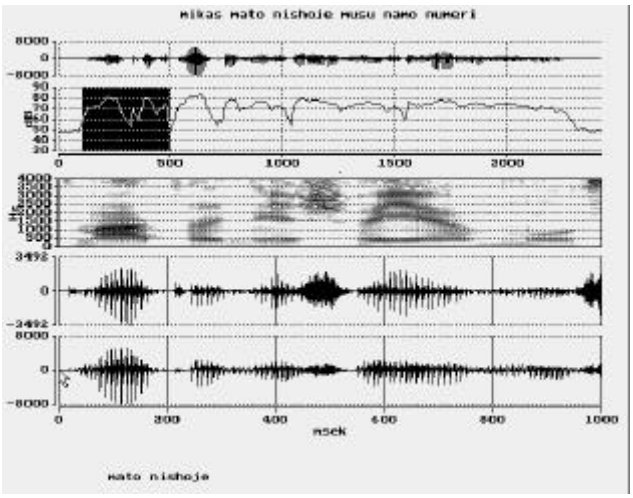

Figure 4: The phrase to word segmentation (4th step). The pictures are the same as in Figure 2. The marked first word is highlighted in the energy picture. There is displayed at bottom the orthography of the current (second) and the following (third) words.

- this step (shown in Figure 4) is very similar to the second step of procedures except that here we are trying to locate boundaries of any other word of the phrase (not the first one). So colors in the energy curve graph of marked words were changed and bottom oscilogram is prepared for accurate location of word boundaries. The start of oscilograms in lower part of picture is the end of previously located word.

Performed actions: actions at this stage are the same as in second phase: we must to place cursor accurately first at the start of the considered word and then at the end of the same word.

Word boundaries labeling recursion

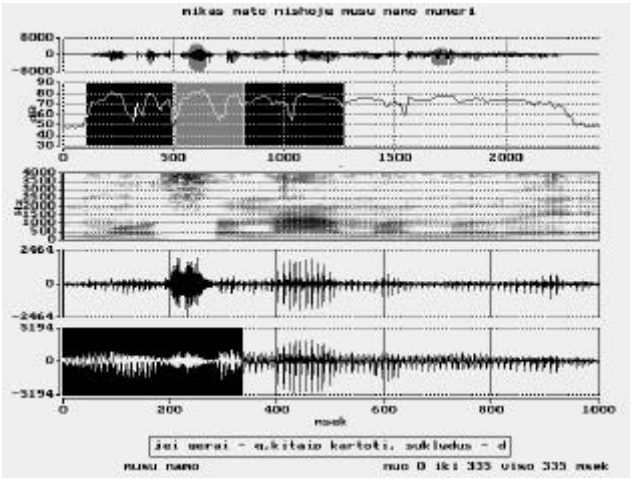

Figure 5: The phrase to word segmentation (5th step). The pictures are same as in Figure 2. The marked three initial words are highlighted in the energy picture. The task is to label fourth word. Orthography of the current (fourth) and the following (fifth) words could be seen in the bottom of this picture.

- Similarly as described above it is almost third phase of these procedures except that here we are trying to find boundaries not of the first word in the phrase but any other word. So earlier found boundaries of the first three words “mikas mato nishoje“ are marked using different colors in the energy oscilogram;
- it is necessary to locate bounds of fourth word “mato“ (next word is “namo“) what could be seen in the bottom graph.

Performed actions: the same as in the third phase.

3.3. WORD TO PHONE SEGMENTATION

Word to phone labeling means location of phone bounds within single word. In this stage information about word boundaries from previous stage is used. Basic ideas of word to phone segmentation are:

- Alternative labeling – extension to right
- Phone end labeling
- Phone end labeling recursion

Alternative labeling – extension to right

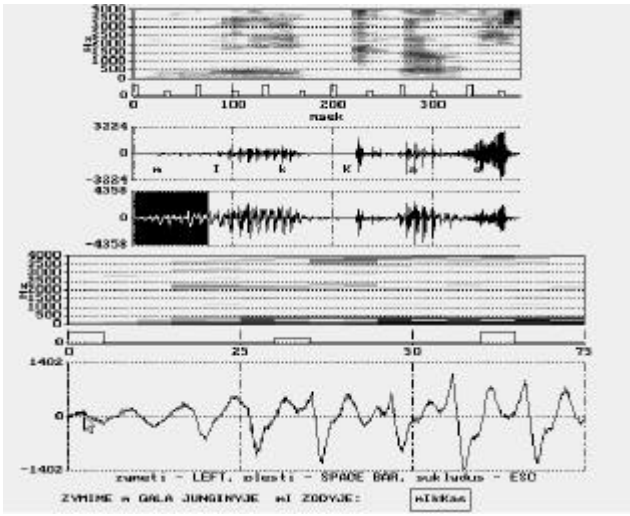

Figure 6: The word to phone segmentation (1 step). There are four pictures for the whole word. From top: a) sonogram, b) preliminary phone marks, c) preemphasized speech oscilogram and word transcription, d) speech oscilogram. The following pictures are in the phone region: e) sonogram, f) preliminary phone marks, g) signal oscilogram. There are the options at the bottom: LEFT- to label SPACEBAR- extension to right, ESC - escape program. The current and following phones are displayed too. Since in this case we can’t see all phone must press SPACEBAR.

- In the top of the Figure 6 we see sonogram of the word (in this case word “mikas”) and below we see preliminary bounds of phones in this word as well as two oscilograms of the same word (oscilograms are slightly narrower). In the last word oscilogram colors in the possible phones boundaries are changed;
- further there are presented part of possible phone, sonogram, bounds of phone and oscilogram;
- suggestions of possible actions are shown at the bottom “to mark – LEFT, to expand – SPACE BAR, to correct – ESC “ and “MARK END OF m IN SYLABLE mi IN THE WORD: mikKas“.

Performed actions: in the left side of the graph in the bottom shown cursor should be placed at the end of considered phone and then left mouse button must be pressed. This means transition to the location of next phone; if the pictures in the bottom can not cover end of phone we must press SPACEBAR and expand shown part of signal to the right; if serious error has been made by labeler then key ESC must be pressed what causes exit from program. The word to phone labeling procedure must be repeated.

Phone end labeling

- in Figure 7 we se same data as in the first phase except that in the last oscilogram phone boundaries were expanded;

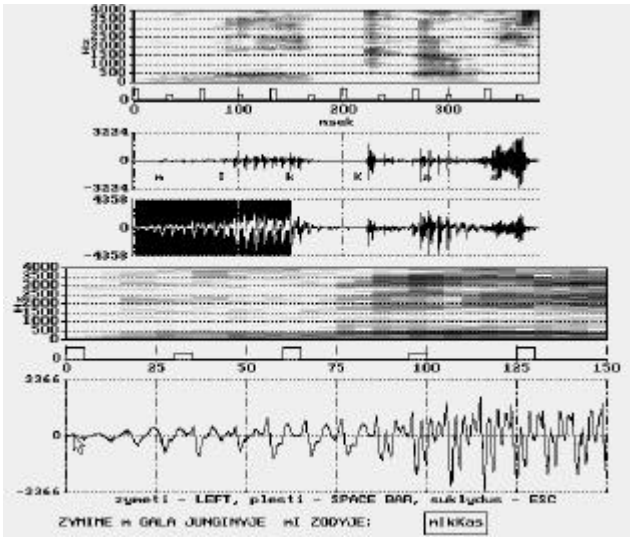

Figure 7: The word to phone segmentation (2nd step). The pictures are the same as in Figure 6. In this case we perform phone end labeling (with LEFT) since we see all m and part of I.

Performed actions: same as in the first phase.

Phone end labeling recursion

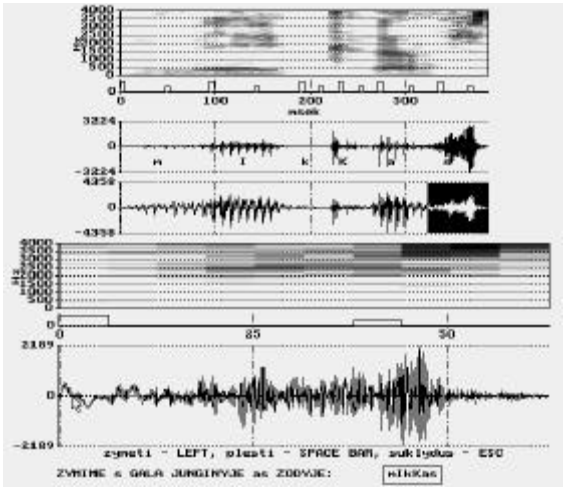

Figure 8: The word to phone segmentation (3rd step). From top: a) sonogram, b) preliminary phone marks, c) preemphasized speech oscilogram, d) speech oscilogram, e) sonogram in the phone region, f) signal oscilogram in the phone region.

- We see the same data as in Figure 6 (first phase) but here we are trying to locate bounds of the last phone (this case phone *s*) in the word;

Performed actions: same as in the first stage.

Figure 9 summarizes phrase to word segmentation and word to phone labeling results. Here we can see whole phrase ("mIkas mAto nisoje mUsu nAmo nUmeri") and each of the words from this sentence. Each word in the phrase oscilogram (graph in the top of picture) is shown in different color background (blue or yellow). Below you could see six graphs showing oscilograms of consequent words in the phrase. Similarly each phone is shown in different color background here (green or dark red). Such tool could be used for visual inspection and illustration purposes.

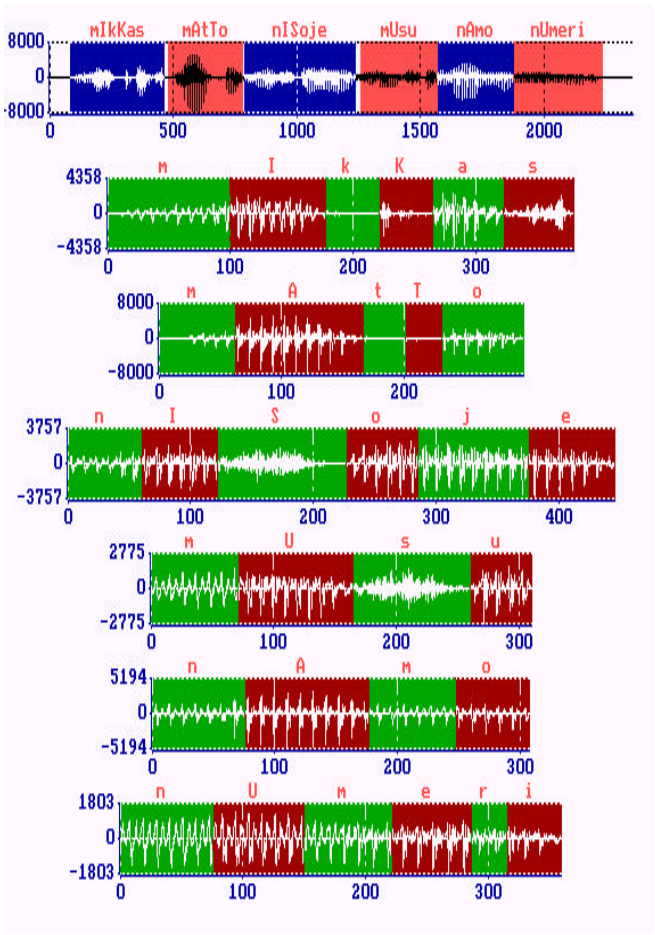

Figure 9: Example of the final result. There is phrase and labeled words in the top raw followed by all six words and labeled phones.

IV. CONCLUSIONS

Lithuanian speech corpora LTDIGITS have been prepared. It contains Lithuanian continuous digit sequences similarly as in TIDIGITS (Leonard, 1984) and Lithuanian computer control words. Important feature of LTDIGITS is collected set of CV (consonant - vowel) clusters containing nasals and various vowels (MA, NA etc.) and same syllables in continuous phrase. Recordings from 115 male and 225 female speakers were gathered.

Also this paper describes developed and tested procedures for phrase to word and word to phrase expert level segmentation and labeling. Segmentation procedures provides TIMIT like (Lamel, 1987) label data. These tools allows to speed up substantially speech database processing.

We believe that collected CV clusters will allow to carry out deeper analysis of phoneme discrimination problems including multilingual peculiarities. E.g., it is meaningful to compare results obtained using OGI telephone quality speech where methods with strong discrimination capabilities wasn't

applied with results achieved using discriminant methods applied on other data. Also it is important to verify methods used in our earlier works with strong discrimination capabilities on LTDIGITS and other data. Previous results showed significant phoneme discrimination error rate reduction using carefully selected features and discrimination methods.

Analysis of similar CV and VC cluster data of E – set data could be promising also.

REFERENCES

Ayer C.M., Hunt M., Brookes D. (1993). A Discriminatively Derived Linear Transform for Improved Speech Recognition, Proc. of EUROSPEECH, Berlin

Chengalvarayan, R. & Deng, L. (1998). "Speech Trajectory Discrimination Using the Minimum Classification error Learning". IEEE Trans. on Speech and Audio Processing, vol. 6, no 6

Cole, R., Fanty, M., Roginski, K. A (1992). Telephone Speech Database of Spelled and Spoken Names. Proc. Int. Conf. Spoken Language Processing , pp. 891 - 893.

Fisher, W.M., Doddington, G.R., Goudie-Marshall, K. (1986). The DARPA Speech Recognition Research

Database: Specification and status. Proc. of DARPA Workshop on Speech Recognition, pp. 93 - 99.

Lamel, L.F., Kasel, R.H., Seneff, S. (1987). Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus. In Proc. of the DARPA Speech Recognition Workshop, 1987, pp. 26-32

Leonard, R. G. (1984), A Database for Speaker Independent Digit Recognition, Proceedings of ICASSP, San Diego, Vol. 3, p. 42.11

Loizou, P. & Spanias, A. (1996) "High Performance Alphabet Recognition," IEEE Trans. on Speech and Audio Processing, vol.4, no 6, pp. 430-445.

Rudzionis, A. & Rudzionis, V. (1999). Phoneme recognition in fixed context using regularized discriminant analysis. Proc. 6th European Conference on Speech Communication and Technology, Eurospeech'99, Budapest, Hungary, pp. 2745 - 2748.

Rudzionis, V. (1998). Speech Recognition by Phonetic Units, Ph.D. Thesis, Kaunas University of Technology