Advanced Recognition of Lithuanian Digit Names Using Hybrid Approach

Kastytis Ratkevicius[1], Gintare Paskauskaite[1], Gintare Bartisiute[2]
[1]*Department of Automation, Kaunas University of Technology, Studentu St. 48, LT-51368 Kaunas, Lithuania*
[2]*Department of Multimedia Engineering, Kaunas University of Technology, Studentu St. 50–401, LT-51367 Kaunas, Lithuania*
gintare.paskauskaite@ktu.lt

***Abstract*—The paper deals with the recognition of digits and with the hybrid recognition technology. By the hybrid approach, we assume the combination of two or more different recognizers have to achieve higher recognition accuracy. Two Lithuanian recognizers using the word based and phoneme-based hidden Markov models (HMM) together with the Spanish language recognizer 8.0 (Spanish-US) and Microsoft Speech Server Spanish language recognizer 9.0 (Spanish-US) were investigated. Using data mining package Weka, classification research was carried out with five different recognizer combining scenarios. The results of connecting two or three recognizers showed that the suggested method of using machine learning method to connect different recognizers greatly improved the recognition accuracy of digits speech corpus in all five cases. Manual annotation of the part of speech corpus enables to increase the recognition accuracy of Lithuanian digits names about 40 % using sub-word-based recognizer. SAMPA_LT set of phonemes is redundant for the digits recognition.**

***Index Terms*—Classification algorithms; Hybrid intelligent systems; Machine learning; Speech recognition.**

I. INTRODUCTION

Automatic Speech Recognition (ASR) could be useful in many areas of life. One of such areas is the recognition of PIN codes, telephone numbers or credit card numbers, but very high requirements to the accuracy of digits recognition are raised in such case – the accuracy of digits sequence is calculated by raising the accuracy of digits recognition to the power of digits sequence length.

Speech recognition system design includes a number of different methods, models, and algorithms. The Hidden Markov Model (HMM) is the most common and successful approach for isolated voice commands recognition today. Looking for the large vocabulary continuous speech recognition methods, the Deep Neural Networks (DNN) gives us quite precise results of speech recognition [1], despite the fact that it requires high data resources, which is complicated for low recourse languages. Such languages typically have a low presence on the internet, with limited textual resources in electronic form and little available knowledge about the language [2]. The Lithuanian language is among some other low-resource languages because there is no necessary annotated and transcribed acoustic training data for the Lithuanian language: the collection, transcription, and annotation of speech data are typically expensive and time-consuming tasks.

The aim of the research is to find the solution to reach as high as possible recognition accuracy of Lithuanian digits names considering that the Lithuanian language is a low resource language.

The one possible solution to achieve this goal is to develop a hybrid speech recognition system – the combination of two or more different recognizers. The hybrid approach provides the possibility that in the case when the answers of recognizers differ, the correct answer could be found using machine learning methods. Another important factor is that different recognizers give different features of recognition. Nevertheless, it has to be noted that in many of the current state-of-the-art speech recognizers, hybrid recognition principles are implemented in many different methods [3].

The inclusion of adopted foreign language speech recognizer is aimed at exploiting the elements of well-developed acoustic models of foreign languages, moreover, for example, Microsoft speech recognizer 8.0 for Windows (Spanish - Spain) enables good voice command end – pointing. It was used in our experiments.

Two Lithuanian recognizers using the word based and phoneme-based hidden Markov models (HMM) were prepared for the implementation of the hybrid recognizer. Taking into account that the annotation of examined speech corpora requires a lot of time and human outlay, the studies on the recognition of speech corpora were limited by word based and phoneme-based HMMs. However, the influence of the manual annotation of the part of speech corpus to the recognition accuracy of Lithuanian digits names was investigated.

Microsoft Speech Recognizer 9.0 for Microsoft Office Communications Server Speech Server (MSS'2007) (Spanish-US) was selected for telephony applications.

The similar experiments of Lithuanian digits recognition were done in 2015 [4]. The best result of 97.51 % was acquired when some foreign language recognizers and Naive Bayes classifier were used. Such recognition accuracy is not suitable for the digits sequence recognition.

This publication presents the new results of experiments

Manuscript received 3 October, 2017; accepted 28 February, 2018.

of Lithuanian digits names recognition using a hybrid approach. The methodology of two recognizers combining, selection of the classifier and description of features, used for two recognizers combining was presented in [5].

II. SPEECH CORPUS

Speech corpus DIGITS30 was selected for our experiments. 30 speakers pronounced 10 digit names 20 times using the high quality Sennheizer IE8i headphones with microphone. 7 men and 23 women took part in the speech corpus preparation, the parameters of speech corpus: 16 kHz sampling rate and 16 bits resolution.

The demarcation of phonetic units – whether segments or others – can proceed in two ways: automatically or manually. A number of automatic instruments have been developed, most frequently based on HMMs (Hidden Markov Models) [6]. Unfortunately, these methods are at present not accurate enough for phonetic research and they need manual correction [7]. The part of speech corpus (the utterances of 6 speaker-women) was manually annotated using SFSWin program providing HTK-based labels [8] of segments. The annotation was done using a minimal set of phonemes and the SAMPA_LT [9] set of phonemes expecting to find if the SAMPA_LT set of phonemes is not redundant for the digits recognition task.

III. FOREIGN LANGUAGE RECOGNITION SYSTEM

The operation of the Non-Lithuanian recognizer is based on multilingual recognition principles with expectations that phonetic features of the recognizable language to a large extent reflect in acoustic-phonetic models of the basic recognition language.

The purpose of training is to find out which acoustic models of basic recognition language describe the best properties of phonetic units of Lithuanian speech. The selection of appropriate sequences (transcriptions) of phonetic units is a central adaptation ("mapping") task.

A. Speech Server Recognition System

As was mentioned above, MSS'2007 [10] was chosen for preparing of telephony services. It provides tools for developing applications that run over a telephone. The following language packs are available in MSS'2007: English, German, French and Spanish. All they were used in our experiments.

Speech server needs telephony format of speech input, so speech corpora were adopted by down-sampling the speech corpus from the original 16 kHz to 8 kHz sampling rate.

The experiments of Lithuanian digit names recognition by MSS'2007 using four recognizers were performed. All possible UPS (Universal Phone Set)-based transcriptions were generated using the lexicon design tool of MSS'2007 for all used languages. After the testing experiments only two or three transcriptions of digits names were left and the recognition accuracy of speech corpus DIGITS30 was measured for each recognizer. The results showed, that Spanish recognizer of MSS'2007 greatly outperformed other recognizers [11].

B. Spanish Speech Recognition System

Spanish language recognizer implemented in Windows'8 operation system (Microsoft Speech Recognizer 8.0 (Spanish-US)) (REC_SP) was chosen for adaptation process to the Lithuanian language recognition.

For the adaptation of the Spanish language recognizer, it is necessary to find the best phonetic sequences (transcriptions) for the Lithuanian voice commands: the transcriptions selected using the lexicon design tool of Microsoft Speech Server'2007 were used for Lithuanian digits recognition by REC_SP recognizer.

IV. HTK BASED LITHUANIAN SPEECH RECOGNIZER

Practically HMM-based speech recognizers are either word-based or sub-word – based. In the word-based case, the whole word in the system's vocabulary is modeled by a single model, which is trained on examples of each word spoken in isolation.

In our experiments, the Lithuanian speech recognizer is HTK - based recognizer. Continuous density hidden Markov models (CD-HMM) were used for the creation of a Lithuanian recognizer by using an open code software toolkit HTK v.3.2 (Hidden Markov Toolkit) [8]. In the first place, sound materials were transformed into the feature vectors. After that, speech recordings were sampled at 16 kHz frequency and broken down into 20 ms duration frames using 10 ms displacement to each other (overlapping analysis windows). During the coding procedure 39 coefficients for each signal frame were computed consisting of 12 Mel-frequency cepstrum coefficients (MFCC) and the energy plus the delta coefficients and the acceleration coefficients [5].

A. Recognition with Word-Based HMMs

In the case of isolated word recognition, word-based HMM recognition should be investigated before sub-words – based HMM recognition, because co-articulation and prosody phenomenon are more likely to be reflected in units of longer duration.

In our previous experiments the important HMM parameters – the number of states and the number of Gaussians per state – were selected during the recognition experiments of speech corpus DIGITS30. The best recognition accuracy of this corpus was achieved using 2 additional states and 6 Gaussians: the number of states is equal to the number of phonemes of digit name plus two states [5].

B. Recognition with Sub-Word-Based HMMs

Several experiments were conducted for the preparation of sub-word-based recognition models. 24 different sets of phoneme experiments were established and carried out using DIGITS30 corpora. 24 speaker's entries were used for the learning process and remaining 6, for testing. The first set Dig1, made of phonemes without palatalization and accentuation, is the primary set. It consists of 10 consonants, 5 vowels, 2 diphthongs and 2 silence phonemes *sp* and *sil*, as it is practiced in HTK. The set Dig2 contains 28 phonemes selected using SAMPA_LT [9] requirements: stressed *i*, *u*, long *i*, palatalized *t*, *d*, *sh*, *s*, *n* and velar *n* were added to phoneme set Digit2. Further phoneme selection and set expansion were made as the research progressed: *a* new allophone *ud* (phoneme *u* after consonant *d*) was

incorporated for command DU (two in English), which increased the accuracy of DU recognition, two allophones *ish* (phoneme *i* after consonant *sh*) and *esh* (phoneme *i* before consonant *sh*) were added to increase the accuracy of command SESI (six in English) recognition. During the research process, four additional allophones (two allophones of phoneme *e* and two allophones of phoneme *i*) were added to the set Dig16. The expanded phoneme set Dig16, containing 35 phonemes had the highest recognition of all 24 phoneme sets recognition experiments, so it was used for phoneme-based Lithuanian recognizer REC_LTp.

Another phoneme-based Lithuanian recognizer REC_LTp_a was prepared using the annotated part of speech corpus (the utterances of 6 speaker-women) for its training.

V. RESULTS OF LITHUANIAN DIGITS CORPUS RECOGNITION

The results of Lithuanian digits corpus DIGITS30 recognition by five different recognizers are presented in Table I.

In order to test all speakers' recordings, a cross-validation based experiments were carried out with REC_LTw and REC_LTp (Dig16 set of phonemes) recognizers. 5 times cross-validation (5TCV) method for DIGITS30 speech corpus was chosen: speech corpus was divided into 5 folds, 1 fold used for training, 4 folds – for testing. It is repeated 5 times changing training and testing folds and averaging the results of testing.

The recognition experiments with REC_SS and REC_SP recognizers were carried out testing all utterances of 30 speakers without training phase. The REC_LTp_a recognizer was trained using the annotated part of speech corpus (6 speakers) and was tested using all speech corpus (30 speakers) or the utterances of 24 speakers. The results of REC_LTp (Dig1) and REC_LTp (Dig2) are given using only one fold of speech corpus for testing and are useful showing the influence of annotation on recognition accuracy.

TABLE I. THE RECOGNITION ACCURACY OF DIFFERENT RECOGNIZERS.

SS	**SP**	**LTw**	**LTp**			**LTp_a**	
30 sp.	**30 sp.**	**5TCV**	**Dig1 6 sp.**	**Dig2 6 sp.**	**Dig16 5TCV**	**Dig1 30 sp.**	**Dig2 24 sp.**
99,12	92,05	99,19	59,25	80,92	97,10	99,28	98,00

VI. COMBINING OF RECOGNIZERS

The hybrid approach would have the sense only if their performance is uncorrelated (both recognizers have high enough recognition accuracy, but the errors are largely different) or at least their performance could help to make a final decision.

Using data mining package Weka, classification research was carried out with five different recognizer combining scenarios:

1. REC_LTw / REC_SP;
2. REC_LTw / REC_LTp;
3. REC_LTw / REC_LTp / REC_SP;
4. REC_LTw / REC_SS;
5. REC_LTw / REC_LTp_a.

For the connection of recognizers, two different methods were applied. The Ordinary 10 times cross-validation (10TCV) with graphical WEKA interface was involved: one file with attributes of all speakers is prepared, then WEKA by default randomly distributes data: 90 % for the training, 10 % for testing. It performs the classification 10 times changing the set of test objects and then calculates the average of the obtained results. This classification method allows predicting the accuracy of the classification (at the same time, the accuracy of the hybrid recognizer) for the "known speaker" (one of the announcers of the speech corpus). Another method – already mentioned 5TCV method. The results of such classification allow the prediction of the classification accuracy for an "unknown speaker".

Classification experiments were conducted by using 10 different classifiers: RIPPER, C4.5, Multinomial Logistic Regression (MLR), Multilayer Perceptron (MP), ZeroR, AdaBoost, K-Nearest Neighbour (kNN), Random Forest (RF), Support Vector Machines (SVM), and Naive Bayes (NB). The task of classification was to separate two classes: TF (both recognizers produce a different hypothesis, the first recognizer produces correct decision) and FT (both recognizers produce a different hypothesis, the second recognizer produces correct decision).

Each object in the training set has been described using features. Among those features are such parameters as the confidence measure of the result provided by REC_SP or REC_SS recognizers, the average logarithmic probability of the REC_LTw or REC_LTp recognizers, the proportions of all sounds presented in the hypothesis produced by both recognizers and some other parameters [5].

The highest classification accuracy from 10 most popular data mining algorithms for digit names was achieved by Random Forest (RF) classifier when the number of trees is equal to 100.

The recognition accuracy of the hybrid recognizer is calculated by summing the number of subset TT (both recognizers produces correct decisions) with the number of subsets TF and FT, multiplied by the accuracy of the best classifier and dividing the result by the number of all used utterances. The classification and recognition accuracies of three best-combining scenarios are presented in Table II.

The investigation of the impact of different features to the classification accuracy showed that the best classification accuracy was reached using classification data with all features.

TABLE II. THE CLASSIFICATION AND RECOGNITION ACCURACY OF DIFFERENT COMBINING SCENARIOS.

	LTw/SP		**LTw/SS**		**LTw/LTp_a**	
	10TCV	5TCV	10TCV	5TCV	10TCV	5TCV
Classif.	98,15	98,26	100	93,33	99.80	98,37
Hybr. Rec.	99,78	99,79	100	99,89	99,93	99,91

The recognition errors of different recognizers used in the combining scenarios are shown in the Fig. 1.

The results of two combining scenarios REC_LTw/REC_LTp and REC_LTw/REC_LTp/REC_SP, which are worse compared with other three scenarios, presented only in Fig. 2 and Fig. 3.

The results of investigations in the form of the recognition

error are presented in the Fig. 2 (the results received using 5TCV classification method) and in the Fig. 3 (the results received using 10TCV classification method).

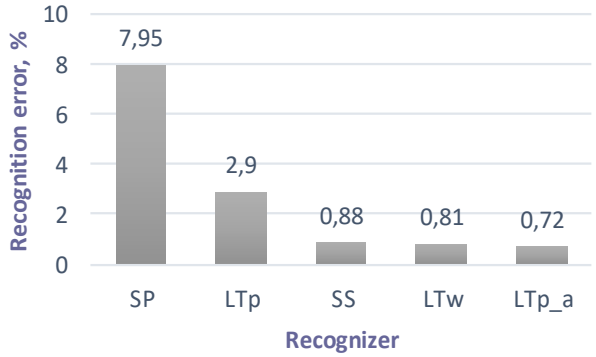

Fig. 1. The recognition errors of different recognizers, used in the combining scenarios.

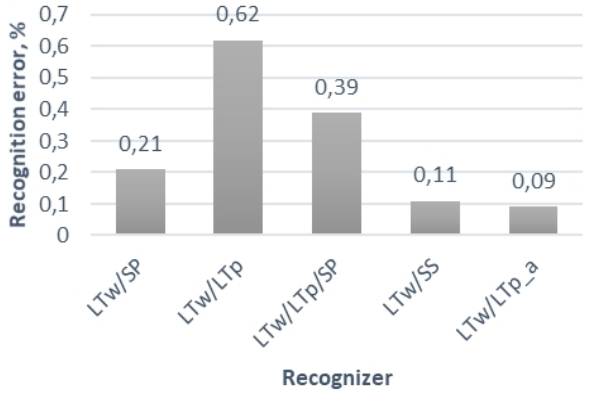

Fig. 2. The recognition errors of hybrid recognizers received using 5TCV classification method.

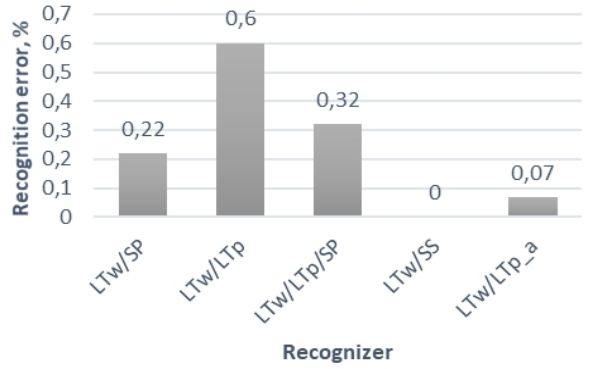

Fig. 3. The recognition errors of hybrid recognizers received using 10TCV classification method.

Results represented in figures, show that the combining of recognizers could significantly improve recognition results in all investigated cases. They outperform all known results of similar experiments done in Lithuania with digits corpora.

VII. CONCLUSIONS

The results of connecting two or three recognizers showed that the suggested method of using machine learning method to combine different recognizers greatly improved the recognition accuracy of digits speech corpus in all five cases.

Manual annotation of the part of speech corpus enables to increase the recognition accuracy of Lithuanian digits names by 40 % using sub-word-based recognizer.

The results presented in Table I shows that the SAMPA_LT set of phonemes is redundant for the digits recognition: the recognition accuracy of REC_LTp_a recognizer using 19 phonemes (set Dig1) is better compared with the accuracy obtained using 28 phonemes (set Dig2).

REFERENCES

[1] L. Deng, G. Hinton, B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: an overview", *IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Vancouver, BC, pp. 8599–8603, 2013. DOI: 10.1109/ICASSP.2013.6639344.

[2] T. Fraga-Silva, A. Laurent, J. L. Gauvain, L. Lamel, V. B. Le, A. Messaoudi, "Improving data selection for low-resource STT and KWS", *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2015)*, Scottsdale, AZ, 2015, pp. 153–159. DOI: 10.1109/ASRU.2015.7404788.

[3] G. Saon, J. T. Chien, "Large-vocabulary continuous speech recognition systems: a look at some recent advances", *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 18–33, 2012. DOI: 10.1109/MSP.2012.2197156.

[4] T. Rasymas, V. Rudzionis, "Lithuanian digits recognition by using hybrid approach by combining Lithuanian Google recognizer and some foreign language recognizers", *Information and Software Technologies. Communications in Computer and Information Science*, vol. 538, pp. 449–459, 2015. DOI: 10.1007/978-3-319-24770-0_38.

[5] K. Ratkevicius, G. Paskauskaite, G. Bartisiute, "Recognition of ICD-10 codes by combining two recognizers", in *Proc. 7th Int. Conf. Baltic (HLT 2016)*, Riga, Latvia, 2016, vol. 289, pp. 51–58. DOI: 10.3233/978-1-61499-701-6-51.

[6] F. Brugnara, D. Falavigna, M. Omologo, "Automatic segmentation and labeling of speech based on Hidden Markov Models", *Speech Communication*, vol. 12, no. 4, pp. 357–370, 1993. [Online]. Available: https://doi.org/10.1016/0167-6393(93)90083-W

[7] R. Skarnitzl, P. Machac, *Principles of Phonetic Segmentation.* Phonetica, 2011, pp. 198–199. DOI: 10.1159/000331902.

[8] S. Young, G. Evermann, M. Gales, T. Hain, Dan Kershaw, X. Liu, G. Moore, J. Odell, D. Ollson, D. Povey, V. Valtchey, P. Woodland. *The HTK Book*. Cambridge University Engineering Department, England, 2002.

[9] A. Raskinis, G. Raskinis, A. Kazlauskiene, "SAMPA (Speech Assessment Methods Phonetic Alphabet) for Encoding Transcriptions of Lithuanian Speech Corpora", *Information Technology and Control*, vol. 29, no. 4, pp. 50–56, 2003.

[10] M. Dunn, *Pro Microsoft Speech Server 2007: Developing Speech Enabled Applications with .NET.* Apress, Berkely, CA, USA, 2007.

[11] G. Bartisiute, K. Ratkevicius, "Speech server based lithuanian voice commands recognition", *Elektronika ir Elektrotechnika*, vol. 18, no. 10, pp. 53–56, 2012. DOI: 10.5755/j01.eee.18.10.3061.