Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)

doi:10.3233/978-1-61499-442-8-33

Combining Multiple Foreign Language Speech Recognizers by using Neural Networks

Tomas RASYMAS [a,1] and Vytautas RUDŽIONIS [a]
[a] *Vilnius university, Kaunas faculty, Lithuania*

Abstract. This paper proposes a new method for combining multiple foreign language speech recognizers which are adapted to recognize Lithuanian voice commands. The recognizers are combined by using neural network. The type or structure of speech recognizer is not important for method but at least it must return recognized command and recognition hypothesis. These two parameters are used to train neural network and to make the final decision about recognized command. The proposed method showed that recognition accuracy was increased by 4.94 % as compared to the best single recognizer.

Keywords. Speech recognition, hybrid recognizer, Lithuanian language recognition, adaptation of recognizer

Introduction

The development of large vocabulary speech recognition systems requires enormous resources: both material and human resources. It is difficult to find such resources in a countries were relatively not widely spoken languages are used as a primary mean of communication. Companies such as Microsoft, Apple, Google, Nuance are not interested in developing Lithuanian speech recognition system, because Lithuanian language is not so widely used as some others and don't have significant market potential. At the same time it has been shown that proper adaptation of existing foreign language acoustic models could speed up the development of recognizer and lead to the acceptable recognition level in that language [1], [2], [3]. One of the solutions for this problem might be to try to create our own speech recognition engine, or to adapt the ones created for other foreign languages. Some previous studies have shown that speech recognition systems of languages such as English or Spanish can be quite well adapted for Lithuanian speech recognition [1], [3]. However, the results are not always good and depend on many factors. So, it is logical to try to create hybrid systems, which are based on combinations of different foreign language speech recognition systems and try to achieve better recognition accuracy. The main point of hybrid recognition is a parallel use of several different recognizers expecting that at least one of the recognizers will give the right result. If we want to use hybrid speech recognition method then at least one recognizer should produce correct result and there exists

[1] Corresponding Author: Tomas Rasymas, Vilnius University, Kaunas faculty, Lithuanian; E-mail: tomas.rasymas@khf.stud.vu.lt

parameters enabling to find the correct result. There are and other types of combinations for achieving better accuracy, for example combining different feature extraction methods like MFCC and LPC, or combining different pattern classification methods like DTW/GHMM [4], [5]. In this paper we are analyzing possibility to combine full recognition systems. Currently hybrid speech recognition systems most often are using several methods to combine the results: if-then rules or maximum likelihood selection as well as discriminant analysis [2], [6]. In this paper, we are proposing method based on neural network technology for combining multiple foreign language speech recognition engines.

1. Proposed method description

Proposed method consists mainly from two parts: adapted foreign language speech recognizers and neural network. First part may consist of any number of recognizers, but all of those recognizers should return recognized command text and hypothesis. Speech signal is sent to the multiple engines simultaneously. The speech utterance is then recognized by these all foreign recognizers. Each engine then returns its own best recognition hypothesis and recognized command. After that recognizers output is passed to neural network as input and final decision of result is made by neural network. Block diagram of system which is working using proposed method is illustrated in Figure 1. Of course in order to use any foreign language recognition system firstly we have to transcribe Lithuanian words using foreign language phonemes. In this paper we are not focused on discovering and analyzing the best transcription rules so the phonemes were transcribed as they sound. Of course this can lead to greater recognition errors, but objective of the work is not to increase recognition accuracy by transcribing words. Main objective was to evaluate possibilities to increase recognition accuracy by combining multiple foreign languages acoustic models.

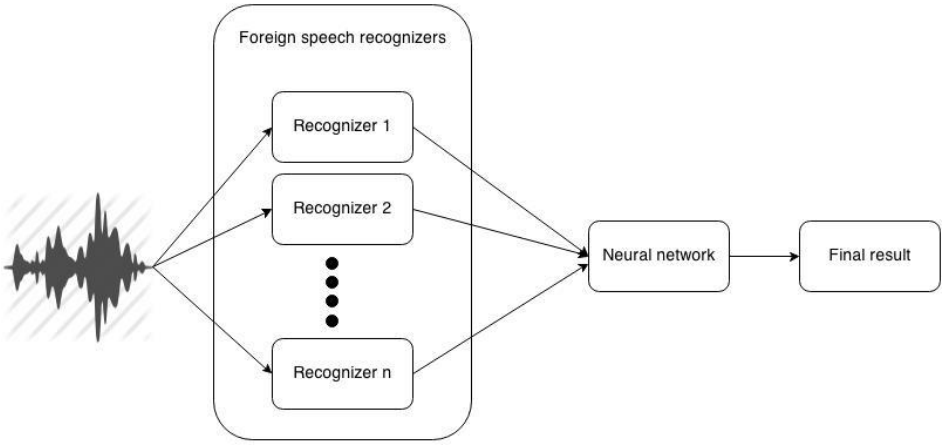

Figure 1. Block diagram of proposed hybrid speech recognition system.

1.1. Foreign speech recognizers

CMU Sphinx 4 was used as a speech recognition and simulation tool. Sphinx 4 is hidden Markov model based speech recognition framework which provides simple way for creating custom speech recognition systems [7]. There are few open source acoustic models which are suitable for Sphinx 4: English, Russian, German and Dutch. These four models will be used for testing proposed method. All acoustic models are trained with 16 kHz recordings. All recognizers were using the same pipeline for speech feature extraction it is displayed in Figure 2.

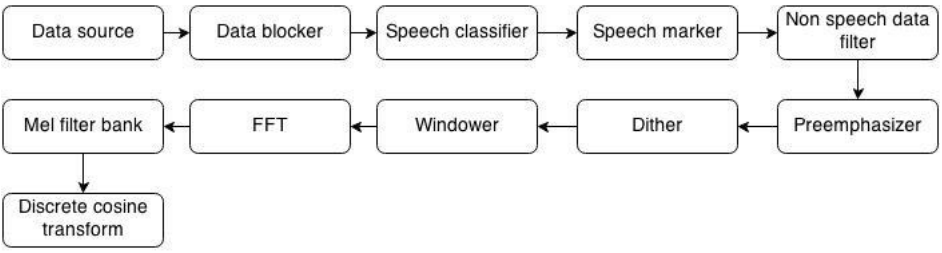

Figure 2. Pipeline of speech feature extraction.

This pipeline is quite standard and is used in majority of modern speech recognition systems [8]. Pipeline components are described below:

- Data source – audio file or audio stream.
- Data blocker, speech classifier, speech marker, non speech data filter – all these components acts as voice activity detection. Signal leaving those components is filtered and only speech is past to further processing.
- Preemphasizer – high-pass filter that compensates for attenuation in the audio data.
- Dither – small amount of random noise is added to the signal to avoid floating point errors and prevent the energy from being zero.
- Windower – in order to minimize the signal discontinuities at the boundaries of each frame, we multiply each frame with a raised cosine windowing function.
- FFT – computes Discrete Fourier Transform.
- Mel filter bank – filters an input power spectrum through a bank of number (40) of mel-filters.
- Discrete cosine transform – applies Discrete Cosine Transform.

Each speech recognition engine as a result returns two parameters: recognized command name and hypothesis. These two parameters from each foreign language recognizer were used for neural network training and classification.

1.2. Neural network

For neural network performance modeling open source software Neuroph was used. After experimenting with different neural network learning rules and types using existing data best results were acquired when multi layer perceptron and resilient propagation learning rule with sigmoid transfer function were used. Lowest mean

square error was acquired with neural network with two hidden layers: first – 27 neurons, second – 63 neurons. Resilient propagation performs a direct adaptation of the weight step based on local gradient information. The main difference to the ordinary backpropagation is that the effort of adaptation is not blurred by gradient behavior whatsoever, it only depends on the sign of the derivative not its value and therefore it will converge from ten to one hundred times faster than the simple backpropagation algorithms [9]. As mentioned before neural network input is command unique identifier and recognition hypothesis. Output of neural network was formed depending on whether the particular recognizer recognized command correctly or not, if recognition result is correct then output is 1 otherwise 0. Neural network architecture is illustrated in Figure 3.

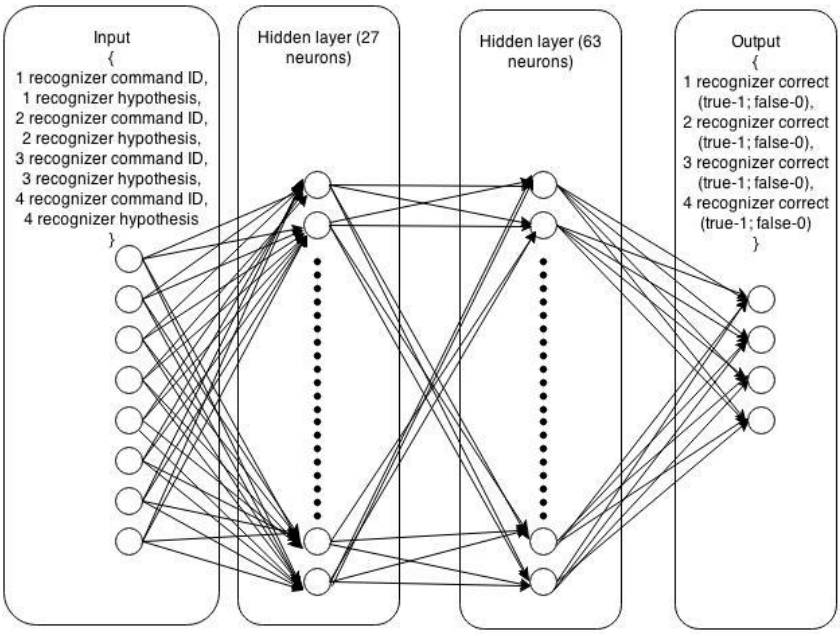

Figure 3. Neural network architecture.

Recognizer with the highest neural network output value is selected and final method result is equal to that recognizer result.

2. Experimental evaluation

The accuracy of the proposed method was tested using Windows 7 based laptop computer (Core i5 CPU, 4 GB of RAM). Main speech corpora containing 25 drug names were used. Speech corpus used in the experiments was gathered by recording speech of 12 people (5 female and 7 male). Each of these speakers pronounced each drug name 20 times in a single session. So every drug name was pronounced for 240 times. Performance of the recognizer has been evaluated in a speaker-independent mode using leaving-one-out methodology. Vocabulary of all drug names used in this experiment is listed in Table 1.

Table 1. A vocabulary of drug names.

No.	Drug name
1	ANALGINAS
2	BIFOVALIS
3	CYKLODOLIS
4	ENARENALIS
5	FERVEKSAS
6	GASTROVALIS
7	HEKSORALIS
8	HEMATOGENAS
9	KETANOVAS
10	KETONALIS
11	KREONAS
12	METFORALIS
13	MIKARDIS
14	NEBIKARDAS
15	PANANGINAS
16	PREDUKTALIS
17	PROPODEZAS
18	RADIREKSAS
19	RANIGASTAS
20	TRACHISANAS
21	TRAVATANAS
22	TRENTALIS
23	TRILEPTALIS
24	VALOKORDIN LAŠAI
25	VERDINAS

First of all single recognizers were tested using obtained recordings and recognition results are shown in Table 2.

Table 2. Single recognizer average error.

Speech recognizer	Average error, %	
English	40.33	
Russian	32.27	
German	53.60	
Dutch	43.87	

Several different configuration neural networks were trained with different combinations of foreign language speech recognizers. 180 recordings were used for neural network training and 60 recording were used for testing. After training neural networks accuracy was evaluated. The obtained results are presented in the Table 3.

Table 3. Combined foreign recognizers average error.

Combined foreign recognizers	Average error, %
English + Russian + German + Dutch	29.68
English + Russian + German	30.60
English + Russian + Dutch	28.93
English + Russian	33.87

Obtained results showed that proposed method allowed reduce the average error by absolute 3.34 % compared with the best individual recognizer result. This result was achieved by using combination of Russian, English and Dutch recognizers.

The difference between the best system and the best single system is very small, which could be attributable to chance. So we asked three more people to take part in experiment. They repeated every drug name for 20 times and obtained recordings were added to neural network testing data set. After that experiments with single and

combined recognizers where repeated. Results of repeated experiments are presented in Table 4 and Table 5.

Table 4. Single recognizer average error after repeated experiment.

Speech recognizer	Average error, %
English	37.62
Russian	31.37
German	55.14
Dutch	41.79

Table 5. Combined foreign recognizers average error after repeated experiment.

Combined foreign recognizers	Average error, %
English + Russian + German + Dutch	28.92
English + Russian + German	31.83
English + Russian + Dutch	26.45
English + Russian	32.86

After comparing both experiments results we can see the difference between the best hybrid system and the best single system increased from 3.34 % to 4.92 %.

3. Conclusions

The results of our experiments showed that it is quite reasonable to use neural networks for combining multiple speech recognizers. Comparing best single recognizer and best combined foreign recognizer average error was decreased by 4.92 %. This experiment also demonstrated that it is possible to adapt foreign language acoustic models for Lithuanian language recognition using just transcriptions.

In the end, even these results still does not “prove” much, as the experiment was done with quite a small corpus, recorded in a controlled environment. However we hope to repeat this evaluation as soon as we’ll get our hands on more speech data.

4. Future work

Best results were achieved using Russian, English and Dutch recognizers so we plan to continue experimenting with those recognizers. Now we are planning to increase recognition accuracy by finding better transcriptions to recognize Lithuanian commands using Russian, English and Dutch language speech engines. Also it is necessary to increase the vocabulary used in the experiments. Especially important is to increase the variety of the phonetic elements used in the adaptation process. With the recent advent of deep belief networks it is important to evaluate the efficiency of these types of networks allowing to use more complicated structures and to capture more subtle characteristics of recognizer properties.

In the future we are planning to repeat experiments with large corpus also compare proposed method results to other most popular speech recognition combination methods.

References

[1] R. Maskeliūnas, A. Rudžionis, K. Ratkevičius, V. Rudžionis, Investigation of Foreign Languages Models for Lithuanian Speech, *Electronics and Electrical Engineering (2009)*, 15 – 20.
[2] V. Rudžionis, K. Ratkevičius, A. Rudžionis, G. Raškinis, R. Maskeliūnas, Recognition of Voice Commands Using Hybrid Approach, *Communications in Computer and Information Science (2013)*, 249 – 260.
[3] R. Maskeliūnas, A. Esposito, Multilingual Italian – Lithuanian Small Vocabulary Speech Recognition via Selection of Phonetic Transcriptions, *Electronics and Electrical Engineering (2012)*, 85 – 88.
[4] E. H. Bourourba, M. Bedda, R. Djemili, Isolated Words Recognition System Based on Hybrid Approach DTW/GHMM, *Informatica (2006)*, 373 – 382.
[5] K. R. Aida-Zade, C. Ardil, S. S. Rustamov, Investigation of Combined use of MFCC and LPC Features in Speech Recognition Systems, International Journal of Electronical, Robotics, Electronics and Communications Engineering (2008), 72 – 78.
[6] W. W. Cohen, Fast Effective Rule Induction, *Proceedings of the Twelfth International Conference on Machine Learning (1995)*, 115 – 123.
[7] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, J. Woelfel, *Sphinx-4: A Flexible Open Source Framework for Speech Recognition*, Sun Microsystems, California, 2004.
[8] N. Dave, Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition, *International Journal for Advance Reseach In Engineering and Technology (2013)*.
[9] M. Riedmiller, H. Braun, A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm, *IEEE International Conference On Neural Network (2011)*.