SPEAKER IDENTIFICATION ACCURACY IMPROVEMENT USING BLSTM NEURAL NETWORK

Laurynas Dovydaitis
Kaunas faculty, Vilnius University, Muitinės str. 8, Kaunas, Lithuania
Laurynas.dovydaitis@gmail.com
http:// http://www.knf.vu.lt/

Vytautas Rudžionis
Kaunas faculty, Vilnius University, Muitinės str. 8, Kaunas, Lithuania
vytautas.rudzionis@knf.vu.lt
https://www.vu.lt

Abstract - In this work we analyze speaker identification accuracy on Lithuanian speaker dataset LIEPA. This dataset consists of 370 Lithuanian speakers reading given text samples. We preform speaker identification with HMM classification and then repeat the same test with different types of LSTM and BLSTM neural networks. On the given dataset we experimentally observe speaker identification accuracy improvement from 3% to 6% compared to best HMM implementation.

Keywords: HMM; BLSTM; speaker identification.

1. Introduction

Speaker identification is one of the more challenging tasks, because of the nature of human voice variability for each individual speaker. Not only speech signal varies between different speakers, utterances among same speaker samples differ as well. On the other hand, person recognition by voice is very attractive, because it does not require expensive equipment to collect data, compared to other biometric identification means, like iris scanners or fingerprint readers.

In the field of biometric recognition by voice, most widely used method for classifying speaker is Hidden Markov Models (HMM). In the light of the new breakthroughs in deep learning and building on the success of language recognition with deep learning, we find, that these techniques can be successfully applied to speaker recognition tasks.

In this paper we show how we can use neural networks and deep learning to improve speaker identification accuracy compared to HMM. We use grid search method to find best neural network hyper-parameter configuration and then test it for highest speaker identification accuracy.

2. Speaker identification

The enrollment for speaker identification process can be summarized in two general steps:

(1) Feature extraction;

(2) Classification model training;

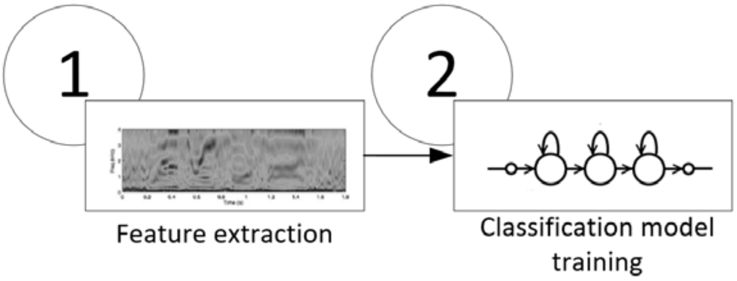

Fig. 1. Speaker enrollment process

To recognize speaker, we need to do the following:

(1) Extract features;

(2) Classify speaker against enrolled models;

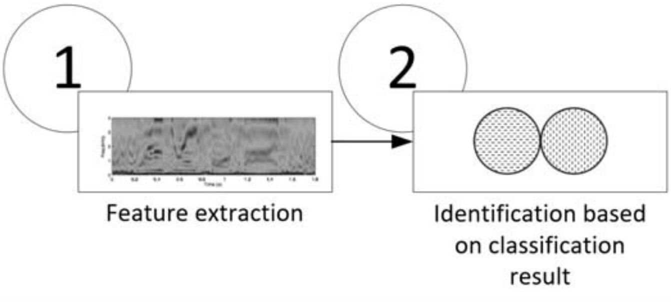

Fig. 2. Speaker recognition process

To choose the classification modeling surface we explored most popular methods using Hidden Markov Models and different configurations of recurrent neural networks. These are covered in more detail in following sections.

2.1. *Identification using HMM*

Hidden Markov model is used in variety of tasks, and performs very well we have data like voice samples. E.g. to predict the unknown outcome, in this case speaker, who is saying particular utterance. As mentioned previously, this particular field was explored deeply by number of different authors.

To highlight some of the research done in the field we can reference following authors, with identification accuracy:

- 97,4% on dataset with 5 speakers (Deshmukh S.D., Bachute M.R., 2013);
- 100% on dataset with 11 speakers (Dovydaitis, L., Rasymas, T., Rudžionis, V, 2016);
- 84,5% on dataset with 20 speakers (Mahola, U., Nelwamondo F. V., Marwala, T., 2007);
- 100% on dataset with 40 speakers (Abdallah, S. J., Osman, I. M., Mustafa, M. E., 2012);

There are several different implementations of HMMs. Below listed works use Gaussian mixture models and some Universal background model to calculate HMMs. Identification accuracy:

- 96,69% on 11 speaker datasets (Bawaskar, A. S., Kota P. N., 2015);
- 95,45% on dataset with 22 speakers (Jayanth, M., Roja, R. B., 2016);
- 92% on dataset with 33 speakers (Meglouli H., Khebli A., 2015);
- 61,9% on dataset with 42 speakers (Ganjeizadeh, F., Lei, H., Maganito, A., Pallipatta, G., 2014);
- 67,5% on dataset with 50 speakers (Zheng, R., Ulang, S., Xu, B., 2004);

All above mentioned authors use Hidden Markov Models, but recognition result depends highly on the speaker dataset and the number of speakers. Based on, training protocol, and model hyper parameters, we cannot get definitive conclusions. Hence it is impossible to summarize predicted accuracy without running additional tests.

2.2. *Identification using neural networks*

Using neural networks for speaker identification is not new. The neural network (DNN) shows better results on speech data and specifically on speaker identification. Some of these authors (Bhattacharya, G., Alam, J., Stafylakis, T., Kenny, P., 2016) analyze speaker identification in their articles achieving:

- 57% identification accuracy by using 100 to 700 neurons;
- 65% identification accuracy by using bi-directional recurrent neural network (RNN);

Another type of recurrent neural network is Long-short term memory networks (LSTM). Extensive LSTM research was done Graves, Schmidhuber, et al. Although not specific to speaker identification, results were published on phoneme recognition accuracy (Graves, A., Schmidhuber, J., 2005), where authors find validation results up to:

- 66% by using LSTM type neural network;
- 70,2% by using BLSTM neural network configuration;

Authors (Zazo, R., Lozano-Diez, A., Gonzalez Dominguez, J., Toledano, D., Gonzalez-Rodriguez, J., 2016) find 70.90% accuracy for speech recognition by using LSTM neural networks.

The result variability on the search for accuracy, shows the importance to run following tests on individual datasets. Also, the lack of research for speaker identification accuracy, using recurrent neural networks motivates to try this type of neural network performance for the given task.

3. On the search of speaker identification accuracy improvement

In order to determine the accuracy improvement, we created a search algorithm. The requirements are as follows:

- Determine the most accurate hyper-parameter configuration for HMM for given dataset;
- Determine the most accurate hyper-parameter configuration for DNN for given dataset;
- Split data set into separate tests to determine result repeatability;

The following algorithm was created (Fig. 3).

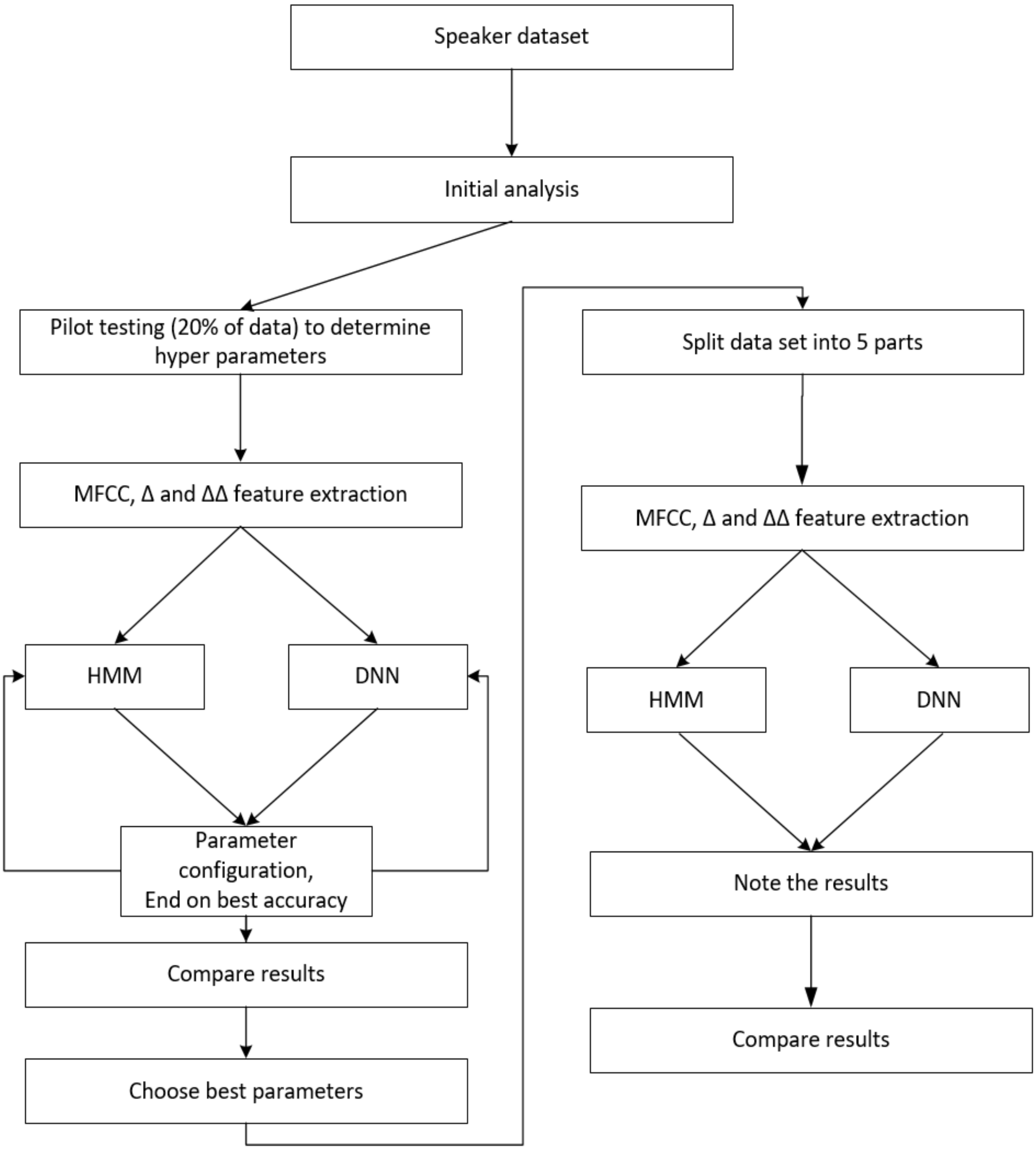
Speaker dataset
Initial analysis
Pilot testing (20% of data) to determine hyper parameters
MFCC, Δ and ΔΔ feature extraction
HMM
DNN
Parameter configuration, End on best accuracy
Compare results
Choose best parameters
Split data set into 5 parts
MFCC, Δ and ΔΔ feature extraction
HMM
DNN
Note the results
Compare results

Fig. 3. Proposed experiment workflow diagram

3.1. *Experimental dataset*

Experimental dataset consists of 370 speakers. The data set was collected as part of LIEPA project (Laurinčiukaitė, S., Telksnys, L., Kasparaitis, P., Kliukienė, R., Paukštytė, V., 2017). This data set contains Lithuanian native speakers. Each of the speaker reads varied excerpt of the text. The specific text for each speaker is not pre-determined. The dataset has the following specification:

Table 1. The planning and control components.

Parameter	Description
Volume of dataset	100 hours
Content	Words, phrases, sentences, texts
Number of speakers	376
Sampling	22kHz
Quantization	16 bits
Number of channels	1

We split this dataset into 5 equal parts. This way we can get a repeatable result within the similar samples.

Table 2. Full dataset splits.

Partial dataset	wav_set1	wav_set2	wav_set3	wav_set4	wav_set5
Speaker count	61	62	62	60	61
Sample count	4408	3523	5528	6090	4155
Training sample count	3119	2497	3905	4292	2934
Validation sample count	1289	1026	1623	1798	1221

3.2. *Classification model hyperparameter search*

In order to determine the best classification model, we had to conduct a hyper parameter grid search for each classification technique. Initial hyper-parameter grid for HMM and DNN is listed in Table 3.

Table 3. The planning and control components.

Step	Initial parameter	Adjusted parameter
HMM	Number of hidden states 5; 7; 10; 16; 22	Adjusted number of hidden states -2;-1;+1;+2;
DNN 1st step	Architecture LSTM 80;	Architecture BLSTM 80;
DNN 2nd step	Dropout: 0,2; 0,4; 0,6;	Dropout: -0,1; +0,1;
DNN 3rd step	Number of cells: +80; +160;	Number of layers: +1; +2;

3.3. *Result comparison workflow*

For the final result comparison, we identify the type of error and output it for further analysis. Possible types of identification errors:

- #1 – DNN and HMM identified speaker correctly;
- #2 – DNN and HMM miss-identified speaker and the same mistake;
- #3 – DNN and HMM miss-identified speaker and made different mistakes;
- #4 – DNN miss-identified speaker, HMM identified speaker correctly;
- #5 – DNN identified speaker correctly, HMM miss-identified speaker;

Following algorithm is used to output final results (Fig. 4).

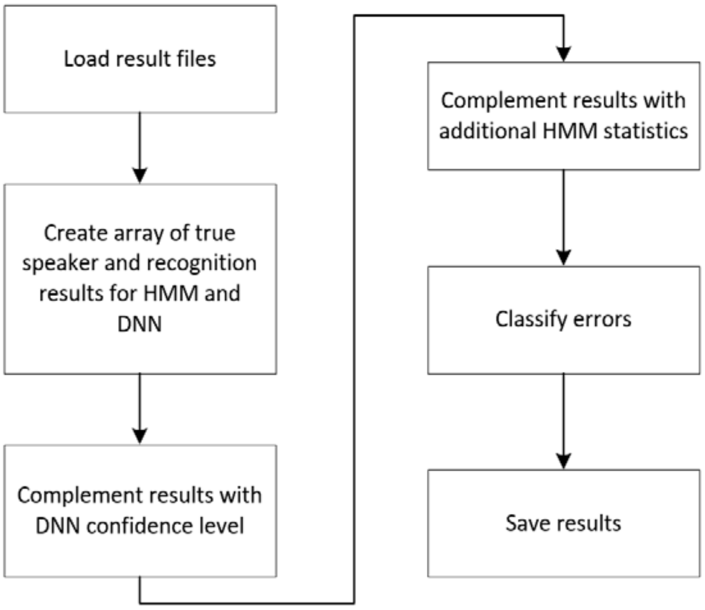
Load result files
Create array of true speaker and recognition results for HMM and DNN
Complement results with DNN confidence level
Complement results with additional HMM statistics
Classify errors
Save results

Fig. 4. Result comparison and further analysis

4. The results

Initial results for the grid search on HMM are displayed in Fig. 5 and for DNN displayed in Fig. 6. From these results we can see, that best performing HMM are those with 3, 5 and 16 hidden states. For DNN best configurations of pilot dataset are BLSTM 160 drop-out 0,3; BLSTM 240 drop-out 0,3; 2xBLSTM 160 drop-out 0,0/0,3.

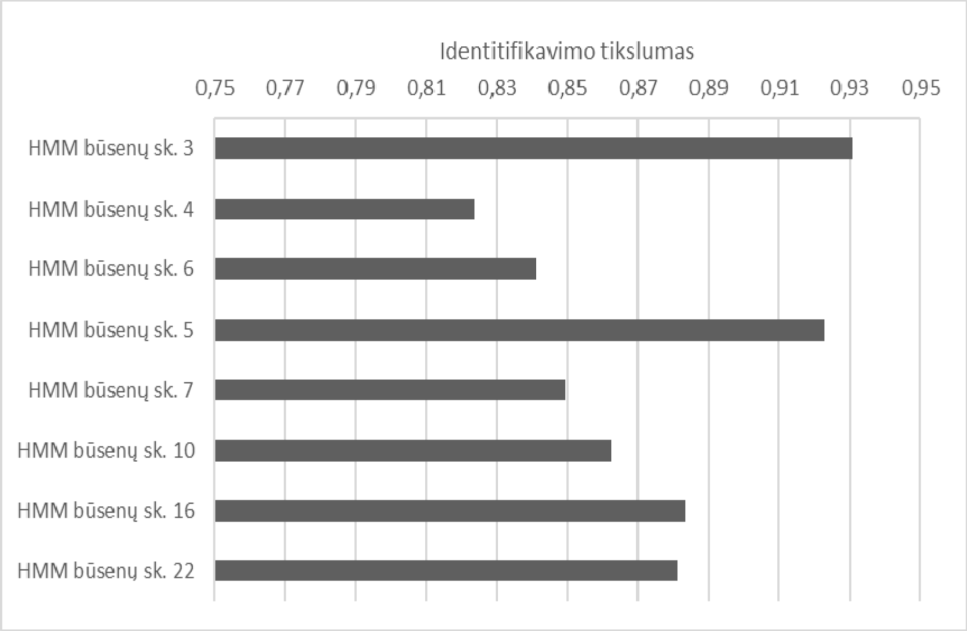

Fig. 5. HMM speaker identification accuracy results for pilot dataset

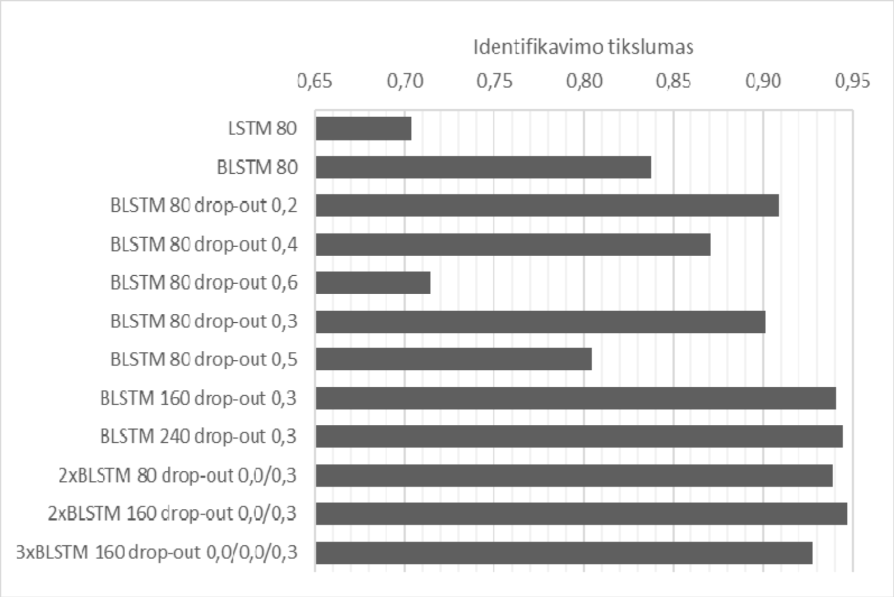

Fig. 6. DNN speaker identification accuracy results for pilot dataset

4.1. *Result comparison on full dataset*

The full accuracy result list of split datasets are listed on Table 4. From the table we can identify the best performing models for HMM and DNN configurations accordingly.

Table 4. Accuracy results for split datasets.

Model	wav_set1	wav_set2	wav_set3	wav_set4	wav_set5	average accuracy	weighted average accuracy
HMM (3 hidden states)	**0,8929**	**0,8235**	**0,8453**	**0,9310**	**0,9000**	**0,8785**	**0,8825**
HMM (5 hidden states)	0,7075	0,6062	0,6223	0,8976	0,7723	0,7212	0,7328
HMM (16 hidden states)	0,7951	0,6617	0,7553	0,9137	0,8853	0,8022	0,8123
BLSTM 160 drop-out 0,3	0,882	0,8323	0,9069	0,9721	0,9508	0,9088	0,9156
BLSTM 240 drop-out 0,3	0,9030	0,8391	0,9248	0,9766	0,9582	0,9203	0,9272
2xBLSTM 160 drop-out 0,0/0,3	**0,9208**	**0,8645**	**0,9235**	**0,9710**	**0,9639**	**0,9287**	**0,9335**

The best performing model of HMM (3 hidden states) is compared to best DNN variant (2xBLSTM 160 drop-out 0,0/0,3) in Fig. 7. We can observe improvement in speaker identification accuracy from 3% to 6% from HMM to DNN.

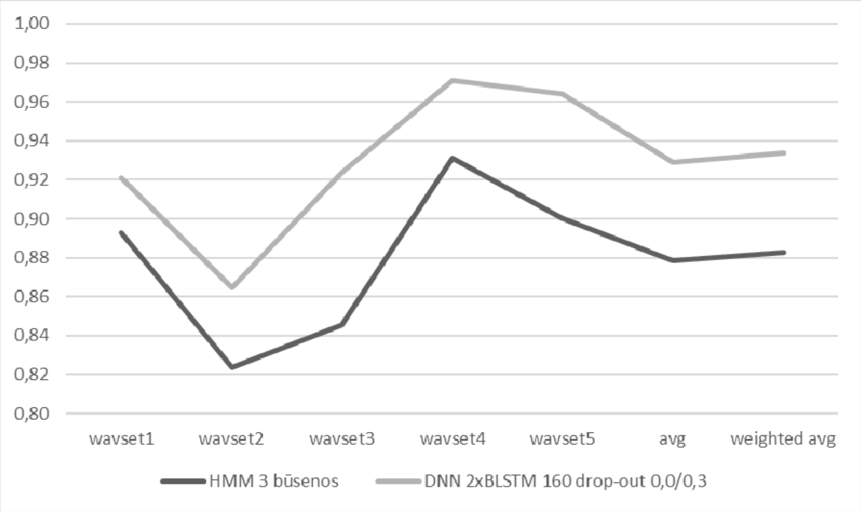

Fig. 7. HMM vs. DNN speaker identification accuracy results comparison for full 5 part dataset

5. Conclusions and further work

We conclude that, given the LIEPA dataset with Lithuanian speakers, BLSTM neural network performs better on identification accuracy, that HMM. Given the data set and particular split we achieve from 3% to 6% improvement on speaker identification accuracy. The following recommendation is to use BLSTM type neural network against HMM, when identifying speakers.

As a further work we plan to use same technique, to try it out on dataset that has lower signal to noise ratio and check whether result has the same accuracy improvement tendencies.

References

[1] Deshmukh S.D., Bachute M.R., Automatic Speech and Speaker Recognition by MFCC, HMM and Vector Quantization, International Journal of Engineering and Innovative Technology (IJEIT) Volume 3, Issue 1, July 2013

[2] Dovydaitis, L., Rasymas, T., Rudzionis, V.: Speaker Authentication System Based on Voice Biometrics and Speech Recognition, Business Information Systems Workshops, BIS 2016 International Workshops, Series Print ISSN 1865-1348

[3] Mahola, U., Nelwamondo F. V., Marwala, T., HMM Speaker Identification Using Linear and Non-linear Merging Techniques, 2007, arXiv:0705.1585

[4] Abdallah, S. J., Osman, I. M., Mustafa, M. E., Text-Independent Speaker Identification Using Hidden Markov Model, 2012, World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 2, No. 6

[5] Bawaskar, A. S., Kota P. N., Speaker Identification Based On MFCC and IMFCC Using GMM-UBM, IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 2, Ver. II (Mar. - Apr. 2015), PP 53-60, e-ISSN: 2319 – 4200, p-ISSN No. : 2319 – 4197

[6] Jayanth, M., Roja, R. B., Speaker Identification based on GFCC using GMM-UBM, International Journal of Engineering Science Invention, ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726 www.ijesi.org,Volume 5 Issue 5, May 2016, PP.62-65

[7] Meglouli H., Khebli A., Speaker recognition by gaussian mixture model (GMM), 2015

[8] Ganjeizadeh, F., Lei, H., Maganito, A., Pallipatta, G., Reducing the Computational Complexity of the GMM-UBM Speaker Recognition Approach, ISSN: 2278-0181, Vol. 3 Issue 3, March – 2014

[9] Zheng, R., Ulang, S., Xu, B., Text-independent speaker identification using gmm-ubm and frame level likelihood normalization, ISCSLP 2004

[10] Bhattacharya, G., Alam, J., Stafylakis, T., Kenny, P., Deep Neural Network based Text-Dependent Speaker Recognition: Preliminary Results, Odyssey 2016, June 21-24, 2016, Bilbao, Spain

[11] Graves, A., Schmidhuber, J., Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures, 2005 IEEE International Joint Conference on Neural Networks, 2005. IJCNN '05. Proceedings

[12] Zazo, R., Lozano-Diez, A., Gonzalez Dominguez, J., Toledano, D., Gonzalez-Rodriguez, J., Language Identification in Short Utterances Using Long Short-Term Memory (LSTM), Recurrent Neural Networks (2016), PLoS ONE 11(1): e0146917. doi:10.1371/ journal. pone.0146917

[13] Laurinčiukaitė, S., Telksnys, L., Kasparaitis, P., Kliukienė, R., Paukštytė, V., Lithuanian Speech Corpus Liepa for the Development of Lithuanian Speech Controlled Equipment, DRAFT, 2017