Identifying Lithuanian Native Speakers Using Voice Recognition

Laurynas Dovydaitis(✉) and Vytautas Rudžionis

Kaunas Faculty, Vilnius University, Muitinės str. 8, Kaunas, Lithuania
{laurynas.dovydaitis,vytautas.rudzionis}@khf.vu.lt

Abstract. In this paper, we analyze speaker identification and present identification test results on Lithuanian native speakers' database LIEPA. Two approaches for speaker acoustic modeling are examined. We start by extracting MFCC features from audio samples, then we feed this data to create speaker acoustic model with hidden Markov models (1) and with deep neural networks (2). We compare both methods by nalyzing the subset of samples from LIEPA database. This helps to achieve more than 96% identification accuracy on sample dataset.

Keywords: Speaker identification · Deep neural networks · Hidden Markov models

1 Introduction

Continuing on previous work we presented in [1], our focus was towards implementation of speech recognition system. We proposed to use such system as gateway for security access control, or as authorization service, for phone, voice mail or voice access services. We are continuing our work on speaker recognition, with more focus on speaker identification by analyzing voice examples.

Previously we faced a challenge with our dataset size, as it was too small to have significant results. Just recently, with project LIEPA [2] completion, substantial set of speaker data became available for deep learning and analysis. This database contains approximately 100 h of samples, from more than 370 Lithuanian native speakers.

This paper shows the results of speaker recognition system for speaker identification, using acoustic modeling with Hidden Markov Models (HMM), as well as, acoustic modeling with Deep Neural Network (DNN) techniques.

1.1 Previous Work

In previous paper [1] we showed results of our experiments. We conducted proof of concept for speaker recognition system, that could be used for user authentication.

We also outlined, that the identification module performance, should be tested on larger dataset. During the experiments, we saw that best identification accuracy was achieved on voice signals without noise.

In [1] we concluded, that in order to increase accuracy, we need to split users' speech stream into smaller windows. We also considered to experiment with speech

W. Abramowicz (Ed.): BIS 2017 Workshops, LNBIP 303, pp. 79–84, 2017.
https://doi.org/10.1007/978-3-319-69023-0_8

recognizers which are based on different speech features (LPC, MFCC etc.) and different machine learning techniques.

Speaker Dataset. For this identification, project LIEPA [2] Speaker dataset was used. This data set includes 376 unique speakers and provides around 100 h of spoken sentences and words. Initial wave format .wav, sampling rate - 22 kHz, quantization - 16 bit, number of channels 1 [2].

Validation Data and Test Data. Original data subset, was split into 70% of samples for training, 30% for testing created model. Splitting was done randomly.

2 Speaker Features

Feature Extraction. Mel-frequency cepstral coefficients (MFCC) were as extracted features. This choice was made because of MFCC feature robustness for speaker recognition [3].

All samples were split using 20 ms length window function, with the help of HTK toolkit software [4]. For each windowed sample, 39 total features were extracted - 13 MFCCs, 13 delta and 13 delta-delta coefficients.

The following parameters were set in HTK configuration files

```
SOURCEKIND=WAVEFORM
SOURCEFORMAT=WAVE
TARGETKIND=MFCC_D_A_E
SAVEWITHCRC=F
SOURCERATE=454.54
TARGETRATE=100000.0
WINDOWSIZE=250000.0
USEHAMMING=T
PREEMCOEF=0.96
NUMCEPS=12
NUMCHANS=20
```

To execute feature extraction, we used HCopy executable

```
HCopy.exe -C CONFIG -S visi_wav-mfc
```

This way we created a speaker feature set, that can be processed further, to create speaker acoustic model.

3 Speaker Acoustic Model

Acoustic model for each speaker was created using two methods. By using Hidden Markov models (HMM) [5] we experimented with various number of hidden states until we got best recognition accuracy. For the second experiment, we created differing deep neural networks architectures [6, 7].

3.1 HMM Model Creation

To train HMM model, the following command was executed from HTK application

```
  HRest.exe -T 1 -S $train -i 100 -l $label -L $labeldir
$hmm
```

3.2 Neural Network

We experimented on number of different configurations in order to create best performing neural network architecture.

The network input layer was a vector of 999 × 39 dimensions, while output layer was had number of nodes, equal to number of unique speakers. Different architectures were used to choose number hidden layers. This was achieved by increasing number of nodes, as well as different depth of networks. Hidden layers consisted of recurrent neural network implementation of Long short-term memory (LSTM) cells [8, 9].

To create and train the network we Python Keras [10] module. One of the architectures that we used can be examined in code example below.

```
  model = Sequential()
  model.add(Masking(mask_value=0.,       input_shape=(999,
39)))
  model.add(LSTM(1536,
                 implementation=2))
  model.add(Dense(67,
                  activation='softmax'))

  sgd     =     SGD(lr=0.1,     decay=1e-6,     momentum=0.9,
nesterov=True)

  model.compile(optimizer='sgd',
                loss='sparse_categorical_crossentropy',
                metrics=['accuracy'])

  model.fit(X_train, y_train,
                epochs=150,
                validation_data=(X_test, y_test),
                callbacks=[csv_logger])
```

4 Test Results

We tested and compared accuracy of the speaker models in two phases. 1st phase was conducted on pilot dataset. This dataset contained 9 unique speaker examples, with total of 540 sample data. In second testing phase, we took subset of LIEPA dataset with 66 unique speakers, with total of 4691 samples.

For DNN results, shown in Table 1, input had a 25 × 39 dimensional vector (Table 2).

Table 1. Pilot dataset accuracy results for 9 speaker voice examples.

Signal to noise ratio	HMM model accuracy	DNN model accuracy
1	1	0.9958
0.9	0.9631	0.9532
0.85	0.9315	0.9295
0.8	0.9105	0.8991

Table 2. Accuracy results with HMM model for experiment running on 66 speaker subset from LIEPA database.

Test sq. number	HMM states	Accuracy
1	8	0.9562

Training for DNN model was stopped at 75 to 150 epochs, depending on loss value, which had to be below 0,05 (Table 3, Figs. 1, 2, 3, and 4).

Table 3. Accuracy results with DNN model for experiment running on 66 speaker subset from LIEPA database.

Test sq. number	DNN architecture	Accuracy
1	1 × 1000 LSTM	0.9048
2	1 × 1500 LSTM	0.8807
3	1 × 2000 LSTM	0.8970
4	1 × 3000 LSTM	0.9097
5	3 × 1000 LSTM	0.9624
6	5 × 1000 LSTM	0.9261
7	7 × 1000 LSTM	0.7990

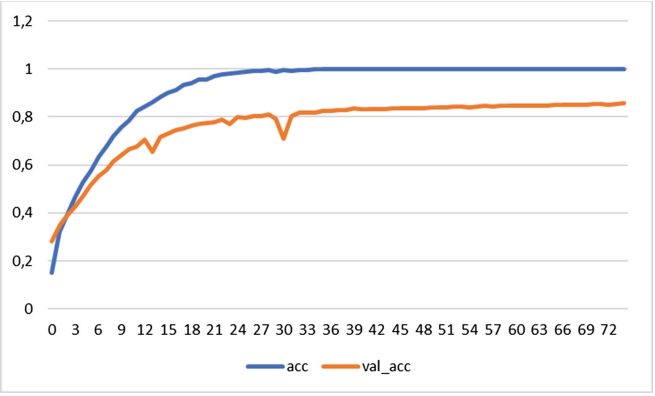

Fig. 1. Accuracy convergence through training epochs on 1st DNN Test sq.

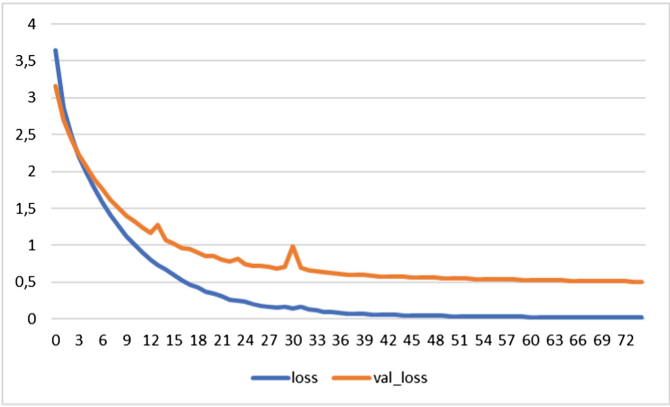

Fig. 2. Loss convergence through training epochs on 1st DNN Test sq.

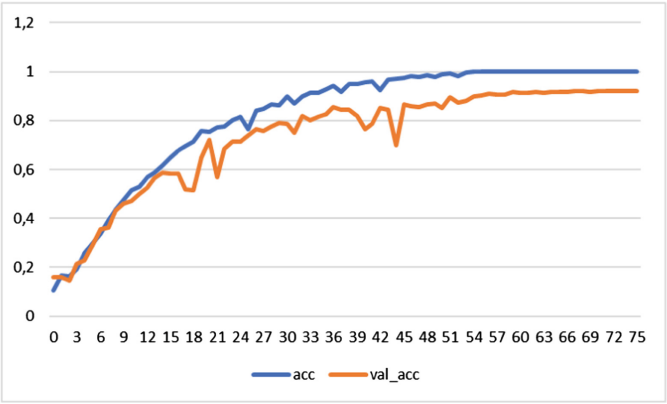

Fig. 3. Accuracy convergence through training epochs on 6th DNN Test sq.

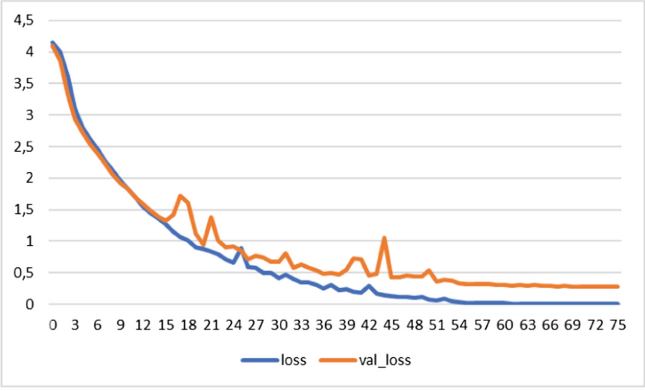

Fig. 4. Loss convergence through training epochs on 6th DNN Test sq.

In Figures above, we can observe network convergence for DNN tests, where blue line shows training set metrics, while orange line shows validation set testing.

5 Conclusions and Further Work

In this paper we shown, that with the use of deep neural networks like LSTM, it is possible to achieve high speaker identification accuracy, which in our tests reached above 96%. This is slightly higher, than speaker acoustic model created with hidden Markov models, which in our tests achieved 95% identification accuracy.

As this shows positive results, we are encouraged to further experiment and improve accuracy of this speaker identification. Also for further work, we plan to examine other LSTM network configurations, by adding additional depth and width to the network, as well as extending training time, to allow better network convergence.

References

1. Dovydaitis, L., Rasymas, T., Rudzionis, V.: Speaker Authentication System Based on Voice Biometrics and Speech Recognition, Business Information Systems Workshops, BIS International Workshops, Series Print ISSN 1865–1348 (2016)
2. LIEPA Homepage. https://www.xn-ratija-ckb.lt/liepa. Accessed 09 May 2017
3. Tiwari, V.: MFCC and its applications in speaker recognition. Int. J. Emerg. Technol. **1**(1), 19–22 (2010)
4. HTK Homepage. http://htk.eng.cam.ac.uk/. Accessed 09 May 2017
5. Abdallah, J.S., Osman, M.I., et al.: Text-independent speaker identification using hidden markov model. World Comput. Sci. Inf. Technol. J. (WCSIT) **2**(6), 203–208 (2012). ISSN: 2221–0741
6. Fandrianto A., Jin, A., Neelappa, A.: Speaker Recognition Using Deep Belief Networks [CS 229] Fall 2012:12-14-12
7. Garcia-Romero, D., Zhang, X., Alan McCree, A., Povey, D.: Improving speaker recognition performance in the domain adaptation challenge using deep neural networks. In: Spoken Language Technology Workshop (SLT), IEEE (2014)
8. Graves, A., Mohamed, A., et al.: Speech recognition with deep recurrent neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2013)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
10. Keras hompage. https://keras.io/. Accessed 09 May 2017