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 perform 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;

To recognize speaker, we need to do the following:

(1) Extract features;
(2) Classify speaker against enrolled models;
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;

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).

![Proposed experiment workflow diagram](image)
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., Klukienė, 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>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volume of dataset</td>
<td>100 hours</td>
</tr>
<tr>
<td>Content</td>
<td>Words, phrases, sentences, texts</td>
</tr>
<tr>
<td>Number of speakers</td>
<td>376</td>
</tr>
<tr>
<td>Sampling</td>
<td>22kHz</td>
</tr>
<tr>
<td>Quantization</td>
<td>16 bits</td>
</tr>
<tr>
<td>Number of channels</td>
<td>1</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Partial dataset</th>
<th>wav_set1</th>
<th>wav_set2</th>
<th>wav_set3</th>
<th>wav_set4</th>
<th>wav_set5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker count</td>
<td>61</td>
<td>62</td>
<td>62</td>
<td>60</td>
<td>61</td>
</tr>
<tr>
<td>Sample count</td>
<td>4408</td>
<td>3523</td>
<td>5528</td>
<td>6090</td>
<td>4155</td>
</tr>
<tr>
<td>Training sample count</td>
<td>3119</td>
<td>2497</td>
<td>3905</td>
<td>4292</td>
<td>2934</td>
</tr>
<tr>
<td>Validation sample count</td>
<td>1289</td>
<td>1026</td>
<td>1623</td>
<td>1798</td>
<td>1221</td>
</tr>
</tbody>
</table>

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>
<thead>
<tr>
<th>Step</th>
<th>Initial parameter</th>
<th>Adjusted parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td>Number of hidden states 5; 7; 10; 16; 22</td>
<td>Adjusted number of hidden states -2; -1; +1; +2;</td>
</tr>
<tr>
<td>DNN 1st step</td>
<td>Architecture LSTM 80;</td>
<td>Architecture BLSTM 80;</td>
</tr>
<tr>
<td>DNN 2nd step</td>
<td>Dropout: 0.2; 0.4; 0.6;</td>
<td>Dropout: -0.1; +0.1;</td>
</tr>
<tr>
<td>DNN 3rd step</td>
<td>Number of cells: +80; +160;</td>
<td>Number of layers: +1; +2;</td>
</tr>
</tbody>
</table>

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).
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.
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.

<table>
<thead>
<tr>
<th>Model</th>
<th>wav_set1</th>
<th>wav_set2</th>
<th>wav_set3</th>
<th>wav_set4</th>
<th>wav_set5</th>
<th>average accuracy</th>
<th>weighted accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM (3 hidden states)</td>
<td>0.8929</td>
<td>0.8235</td>
<td>0.8453</td>
<td>0.9310</td>
<td>0.9000</td>
<td>0.8785</td>
<td>0.8825</td>
</tr>
<tr>
<td>HMM (5 hidden states)</td>
<td>0.7075</td>
<td>0.6062</td>
<td>0.6223</td>
<td>0.8976</td>
<td>0.7723</td>
<td>0.7212</td>
<td>0.7328</td>
</tr>
<tr>
<td>HMM (16 hidden states)</td>
<td>0.7951</td>
<td>0.6617</td>
<td>0.7553</td>
<td>0.9137</td>
<td>0.8853</td>
<td>0.8022</td>
<td>0.8123</td>
</tr>
<tr>
<td>BLSTM 160 drop-out 0.3</td>
<td>0.882</td>
<td>0.8323</td>
<td>0.9069</td>
<td>0.9721</td>
<td>0.9508</td>
<td>0.9088</td>
<td>0.9156</td>
</tr>
<tr>
<td>BLSTM 240 drop-out 0.3</td>
<td>0.9030</td>
<td>0.8391</td>
<td>0.9248</td>
<td>0.9766</td>
<td>0.9582</td>
<td>0.9203</td>
<td>0.9272</td>
</tr>
<tr>
<td>2xBLSTM 160 drop-out 0.0/0.3</td>
<td>0.9208</td>
<td>0.8645</td>
<td>0.9235</td>
<td>0.9710</td>
<td>0.9639</td>
<td>0.9287</td>
<td>0.9335</td>
</tr>
</tbody>
</table>

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.
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

[13] Laurynas Dovydaitis et al. / Indian Journal of Computer Science and Engineering (IJCSE) Vol. 9 No. 2 Apr-May 2018 37

DOI : 10.21817/indjcse/2018/v9i2/180902017