COMPUTATIONAL SCIENCE AND TECHNIQUES
Volume 6, Number 1, 2018, 574-580

Publisher: Klaipėda University
http://journals.ku.lt/index.php/CST
Online ISSN: 2029-9966
DOI: 10.15181/csat.v6i1.1579

BUILDING BI-DIRECTIONAL LSTM NEURAL NETWORK BASED SPEAKER IDENTIFICATION SYSTEM

Laurynas Dovydaitis, Vytautas Rudžionis
Kaunas faculty, Vilnius University

Abstract. In this paper we analyze speaker identification accuracy by using two different classification methods. One of these methods – is well known hidden Markov models that are commonly used for language recognition and speaker identification tasks, while another method is to use classifier, based on neural networks, which is less popular in the speaker identification field. By using grid search, we find the best performing models for both mentioned methods. We compare those methods to find highest speaker identification accuracy on the subset of LIEPA dataset.

Keywords: Lithuanian speaker identification, neural networks, hidden Markov models, bi-directional LSTM.

Introduction

When we want to adapt any content to specific person, first we need to know his or her identity. One of the main advantages of identifying a person from one's voice is that it can be done remotely. From technical standpoint, you don't need to have specialized equipment, which is required for the other identification means e.g. fingerprint reader or iris scanner. All you need for the speaker identification is a microphone connected to a computing device.

There is a very little research for Lithuanian speaker identification, so if one would want to identify person from given demographic, one would need to make thorough investigation beforehand. One of the main reasons for the lack of such research was the availability of Lithuanian speaker dataset. Upon completion of project LIEPA[1], substantial set of speaker data became available. This database contains approximately 100 hours of samples and a total of 370 Lithuanian native speakers, most of which can be used for various language recognition and speaker identification tasks.

In this paper we show preliminary testing results of speaker identification for Lithuanian speakers. We compare identification accuracy results, based on classification with hidden Markov models (HMM) and deep neural networks (DNN), more specifically bi-directional long short-term memory (BLSTM) networks.

1 Previous work on speaker identification

Hidden Markov models are used in speaker identification field and are already explored by number of different authors (Fakotakis et al., 1997; Mahola et al., 2007; Abdallah et al., 2012; Deshmukh et al., 2013). For example, Deshmukh et al. (2013) and Mahola et al. (2007) in their research achieve 97.4% and 84.5% of speaker identification accuracy

[1] https://www.xn--ratija-ckb.lt/liepa

respectively. These tests are done with different size of speaker datasets. In the first referenced case, result achieved on 5-speaker dataset, while the second conducted with 20-speaker dataset.

As we can see, identification accuracy results vary as they are dependent on the size of the dataset and the number of speakers, as well as test parameters. We can observe, that identification accuracy downgrades, as dataset gets larger.

Using neural networks for speaker identification is not new, yet the research is pretty sparse. In research by Bhattacharya et al. (2016) authors achieve identification accuracy of 57% on test set by using recurrent neural network type. By running additional tests, using bi-directional recurrent neural networks, observed accuracy increases up to 65%. These results are achieved on "vanilla" recurrent neural network configurations and tests are done using proprietary dataset with 98 speakers.

Graves et al. (2005) did another research, for phoneme identification. Though, not specific to speaker identification, this research is useful to reference, because it provides identification accuracy results and its dependency, based on neural network type. In the mentioned research, best results are shown on LSTM and BLTSM type neural networks, where the tests show 66% and 70.2% of phoneme identification accuracy accordingly.

2 Speaker dataset and experimental settings

Speaker dataset LIEPA includes 376 unique speakers and provides around 100 hours of spoken sentences and words. Initial data format is .wav, with sampling rate of 22 kHz, quantization of 16 bit and mono channel recording (Laurinčiukaitė et al., 2017).

For experimental tests, conducted in this paper, we take only part of this dataset. This allows us to run initial accuracy tests more efficiently and in more timely manner. Hence, we limit our experiments to 66 unique speakers, which equal to 66 individual classification sets. Selected speaker subset gives us total of 4691 unique audio samples. The exact number of samples per speaker is not uniform, but we can highlight, that each speaker has at least 28 unique samples.

Selected dataset is split into 70% of samples, used for training the models, and 30% which are used for accuracy testing. 70%/30% splits are done at the speaker level, this gives us at least 8 unique samples per speaker, that are excluded from model training.

To prepare raw data for classification, we extract features from individual .wav files. Feature extraction is done with Mel-frequency cepstral coefficients (MFCC). Two common reasons to choose MFCCs are their robustness and common use for speaker recognition tasks (Tiwari, 2010). All audio samples are split with 20ms window function. For each windowed sample, 39 total features are calculated - 13 MFCCs, 13 delta and 13 delta (Ringelienė et al., 2011).

2.1 Experimental settings

Based on research outlined in 2nd paragraph, it is worth to note, that in order to find the most accurate identification results, we need to fine tune hyper-parameters for each model. We use grid search method to find best parameter set.

HMM classification. In the HMM hyperparameter configuration, one of the most significant configuration change can be achieved, by changing number of hidden states in the model (Fig. 1).

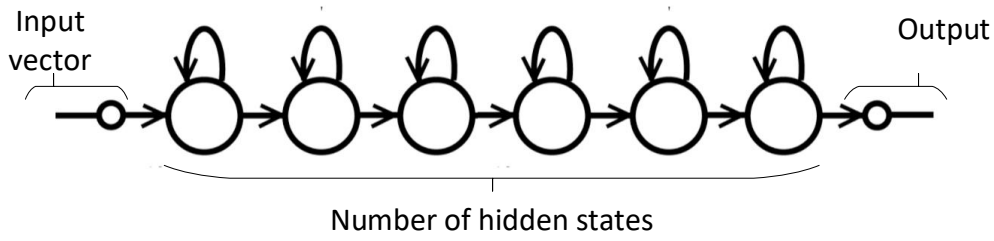

Fig. 1. **Hidden Markov model chain**

For the initial search we take 5; 7; 10; 16; 22 number of hidden states. Afterwards, we choose the best performance and search for a smaller step in parameter range.

Neural network classification. To create a neural network with highest accuracy, we take the same grid search approach, as with HMM parameters.

Principal neural network architecture is shown in Fig. 2. Here input layer is a MFCC feature vector and hidden layer contains long Short-term memory cells. Output layer is a softmax classifier, which predicts speaker identification with confidence value.

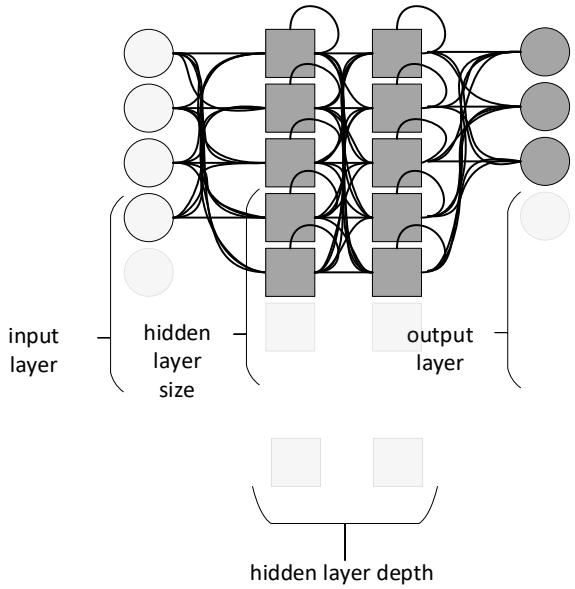

Fig. 2. **Neural network generic architecture**

2.2 Software configuration

For feature extraction we use The Hidden Markov Model Toolkit (HTK)[2]. Using this toolkit, we will also build hidden Markov models (HMM) for each enrolled speaker.

[2] HTK software, http://htk.eng.cam.ac.uk/

The following HTK configuration used for feature extraction:

```
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
```

For neural network we use Pyhton3 with Keras4 package. As deep learning backend we use Theano[5] package. All neural networks built with the default software settings.

3 Test results

Parameters for the HMM grid search are shown in Table 1, respectively you can see identification accuracy results for the test data in the second column. The test data was excluded from model training.

Table 1. HMM parameter search

Number of HMM states	Identification accuracy for test dataset (1=100%)
5	0.9228
7	0.8493
10	0.8624
16	0.8835
22	0.8813

After initial results we refined hyper-parameters on best performing model to improve accuracy even more. Additional results are shown in Table 2. Here we can see, that the best accuracy is achieved with HMM containing 3 hidden states.

Table 2. HMM parameter finetuning

Number of HMM states	Identification accuracy for test dataset (1=100%)
3	0.9308
4	0.8238
6	0.8413

[3] Python software, https://www.python.org/
[4] Keras software, https://keras.io/
[5] Theano software, http://deeplearning.net/software/theano/

The initial DNN parameter setup is listed in Table 3. Compared to HMM, neural networks have more hyper-parameters, that's why we split this process in to more steps. As before, we refine our results by tuning the model. In the second step, we are adding additional dropout layer to avoid overfitting. The initial results with added dropout layer shown in Table 4 and dropout layer fine-tuned adjustments shown in Table 5. DNN architecture naming convention outlined in Fig. 3.

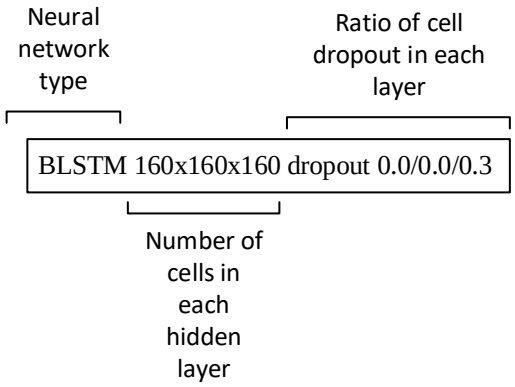

Fig. 3. **DNN architecture naming convention**

Table 3. Neural network parameter search

DNN architecture	Identification accuracy for test dataset (1=100%)
LSTM 80	0.7038
BLSTM 80	0.8377

Table 4. Neural network accuracy with additional dropout layer

DNN architecture	Identification accuracy for test dataset (1=100%)
BLSTM 80 dropout 0.2	0.9090
BLSTM 80 dropout 0.4	0.8704
BLSTM 80 dropout 0.6	0.7147

Table 5. Neural network accuracy with additional dropout layer finetuning

DNN architecture	Identification accuracy for test dataset (1=100%)
BLSTM 80 dropout 0.3	0.9010
BLSTM 80 dropout 0.5	0.8042

We finalize our neural network hyper-parameter search by increasing width and depth of hidden LSTM cells in the network. These results are shown in the following tables: Table 6 – depth (size) increase; Table 7 – width (number of hidden layers) increase.

Table 6. Neural network depth (size) increase

DNN architecture	Identification accuracy for test dataset (1=100%)
BLSTM 160 dropout 0.3	0.9403
BLSTM 240 dropout 0.3	0.9446
BLSTM 256 dropout 0.3	0.9337

Table 7. Neural network width increase

DNN architecture	Identification accuracy for test dataset (1=100%)
BLSTM 80x80 dropout 0.0/0.3	0.9388
BLSTM 160x160 dropout 0.0/0.3	0.9468
BLSTM 160x160x160 dropout 0.0/0.0/0.3	0.9272

As we can see from results, accuracy starts to decrease, when we increase network width size above two layers.

Statistical significance. To calculate the statistical significance of these results, we use McNemar's chi-squared test. This is done to check, whether the best performing HMM model, with 3 hidden states and accuracy of 93.08%, has a statistically significant difference, compared to the best performing neural network - BLSTM 160x160 dropout 0.0/0.3, with accuracy of 94.68%. The results of this test are listed in Table 8.

Table 8. Statistical significance of the best performing models

McNemar's Chi-squared statistic	p-value
3.872	0.0490980

Given that both models perform very well, we can see, that the difference on this dataset, between the best performing HMM and DNN models, is statistically significant. By using BLSTM neural network we increased speaker identification accuracy by 1.6%.

Conclusions and further work

In this paper we have shown, that identification accuracy of Lithuanian speaker is 93% or more by using techniques like hidden Markov models and BLSTM neural networks for datasets with 66 speakers.

We made comparison, which shows, that neural networks can perform better accuracy wise, than well-known methods like hidden Markov models on Lithuanian speaker identification tasks. We achieved increase of 1.6%. in the identification accuracy.

We tested accuracy results for a statistical significance, which for this dataset and the best performing models, shown difference as statistically significant (p value of 0.0490980).

For further work, we plan to use full LIEPA dataset to find a potential accuracy improvement between different methods. Also, we plan to examine denser BLSTM type neural networks, by adding additional depth and width to the network.

References

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

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

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

Fakotakis, N., Georgila, K., Tsopanoglou, A., A Countinous HMM text-independent sepeaker recognition systems based on viwel spotting, 1997, EUROSPEECH'97, 5th European Conference on Speech Communication and Technology

Graves A., Schmidhuber J., Framewise phoneme classification with bidirectional LSTM and other neural network architectures, 2005, Neural Networks, Volume 18, Issues 5–6, July–August 2005, Pages 602-610

Laurinčiukaitė, S., Telksnys, L., Kasparaitis, P., Kliukienė, R., Paukštytė, V., Lithuanian Speech Corpus Liepa for the Development of Lithuanian Speech Controlled Equipment, Submitted to journal INFORMATICA, Vilnius University, 2017

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

Ringelienė, Ž., Filipovič, M., Žodžių atpažinimo, grįsto paslėptaisiais Markovo modeliais, vizualizavimo ir analizės programinė įranga, INFORMACIJOS MOKSLAI. 2011 56, ISSN 1392-0561

The Hidden Markov Model Toolkit (HTK), prieiga per internetą http://htk.eng.cam.ac.uk/, paskutinį kartą kreiptasi 2017/05/02

Tiwari, V.: MFCC and its applications in speaker recognition. International Journal on Emerging Technologies 1(1), 19-22(2010)

L. Dovydaitis is PhD student in informatics engineering at Kaunas faculty, Vilnius University. His research interests include machine learning, neural networks, biometrics on speaker identification.

V. Rudžionis is associate professor at the Institute of Applied Informatics, Kaunas faculty, Vilnius University. He received Ph. D. degree (informatics engineering) in 1998 from Kaunas university of technology. His research interests include signal processing, speech processing, machine learning. He has been participant in several international and national research projects, he published more than 50 research papers in international and national journals.

ASMENS ATPAŽINIMAS PAGAL BALSĄ NAUDOJANT DVIKRYPTĮ ILGĄ TRUMPALAIKĖS ATMINTIES NEURONINĮ TINKLĄ

Laurynas Dovydaitis, Vytautas Rudžionis

Santrauka

Šiame straipsnyje yra analizuojamas asmens identifikavimo uždavinys pagal balsą, naudojant du skirtingus klasifikavimo metodus. Vienas iš minėtų metodų naudoja paslėptuosius Markovo modelius, kitas yra pagrįstas neuroniniais tinklais ir yra rečiau sutinkamas asmens identifikavimo srityje. Atliekant tyrimą yra surandami tiksliausiai diktorius identifikuojantys klasifikatorių parametrai. Gauti rezultatai yra palyginami ant dalies LIEPA garsyno rinkinio.

Pagrindiniai žodžiai: lietuviškai kalbančio asmens identifikavimas, neuroniniai tinklai, paslėptieji Markovo modeliai, dvikrypčiai ilgi trumpalaikės atminties neuroniniai tinklai.