VILNIUS UNIVERSITY

LAURYNAS DOVYDAITIS

RESEARCH ON THE ACCURACY OF LITHUANIAN SPEAKER'S IDENTIFICATION USING RECURRENT NEURAL NETWORKS

Summary of Doctoral Thesis

Technological Sciences, Informatics Engineering (07 T)

Vilnius, 2018

The doctoral dissertation was written in 2013–2017 at Vilnius University.

Scientific Supervisor

Assoc. Prof. Dr. Vytautas Rudžionis (Vilnius University, Technological Sciences, Informatics Engineering – 07 T).

Dissertation to be defended at the Dissertation Defense Council:

Chairman

Prof. Dr. Julius Žilinskas (Vilnius University, Technological Sciences, Informatics Engineering – 07 T).

Members:

Prof. Dr. Stefano Bonnini (University of Ferrara, Italy, Physical Sciences, Informatics – 09 P),

Prof. Dr. Audrius Lopata (Vilnius University, Technological Sciences, Informatics Engineering – 07 T).

Dr. Virginijus Marcinkevičius (Vilnius University, Technological Sciences, Informatics Engineering – 07 T).

Prof. Dr. Habil. Rimvydas Simutis (Kaunas University of Technology, Technological Sciences, Informatics Engineering – 07 T).

The dissertation is to be defended during a public meeting of the Council at the Institute of Data Science and Digital Technologies of Vilnius University, auditorium 203, on September 27, 2018 at 12:00 PM.

Address: Akademijos str. 4, LT-04812 Vilnius, Lithuania.

A summary of the doctoral dissertation was provided for review on August 27, 2018.

The dissertation is available at the library of Vilnius University.

VILNIAUS UNIVERSITETAS

LAURYNAS DOVYDAITIS

LIETUVIŠKAI KALBANČIO DIKTORIAUS IDENTIFIKAVIMO NAUDOJANT GRĮŽTAMOJO RYŠIO NEURONINIUS TINKLUS TIKSLUMO TYRIMAS

Daktaro disertacija

Technologijos mokslai, informatikos inžinerija (07 T)

Vilnius, 2018

Disertacija rengta 2013–2017 metais Vilniaus universitete.

Mokslinis vadovas

doc. dr. Vytautas Rudžionis (Vilniaus universitetas, technologijos mokslai, informatikos inžinerija – 07 T).

Disertacija ginama viešame disertacijos Gynimo tarybos posėdyje:

Pirmininkas

prof. dr. Julius Žilinskas (Vilniaus universitetas, technologijos mokslai, informatikos inžinerija – 07 T).

Nariai:

prof. dr. Stefano Bonnini (Feraros universitetas, Italija, fiziniai mokslai, informatika – 09 P),

prof. dr. Audrius Lopata (Vilniaus universitetas, technologijos mokslai, informatikos inžinerija – 07 T),

dr. Virginijus Marcinkevičius (Vilniaus universitetas, technologijos mokslai, informatikos inžinerija – 07 T),

prof. habil. dr. Rimvydas Simutis (Kauno technologijos universitetas, technologijos mokslai, informatikos inžinerija – 07 T).

Disertacija bus ginama viešame disertacijos Gynimo tarybos posėdyje 2018 m. rugsėjo 27 d. 12 val. Vilniaus universiteto Duomenų mokslo ir skaitmeninių technologijų instituto 203 auditorijoje.

Adresas: Akademijos g. 4, LT-04812 Vilnius, Lietuva.

Disertacijos santrauka išsiuntinėta 2018 m. rugpjūčio 27 d.

Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje ir VU interneto svetainėje adresu: www.vu.lt/lt/naujienos/ivykiu-kalendorius.

INTRODUCTION

1. Scope of the research

The present research focuses on the problem of speaker identification accuracy. This problem is part of the informatics engineering research field dealing with signal processing. The scope of the present research involves multiple tasks: processing of a voice signal, finding of features enabling speaker identification, and the task of feature classification. This thesis focuses on the problem of classification of said features. In order to perform accurate and successful speaker identification it is necessary that the identification process undergoes different stages, namely, collection of a voice sample, extraction of features, classification of those features and their assignment to a specific speaker. While each stage requires using different methods, each of these methods add additional layer of complexity and significant challenges.

2. Relevance of the problem

Speaker identification by voice is one of the most convenient ways of personal biometric identification. There is no speaker identification algorithm that could be used to identify a speaker with 100% accuracy, thus there is always room for error. The main reason why such errors might occur is particularly high signal variance. The fact that each pronunciation may vary for every utterance, a speaker's voice may be affected by an illness, or there could be attempts by malicious actors to impersonate another speaker are just a few examples where voice variance plays a significant role. This is why despite significant efforts made in this area the problem of speaker identification still remains a relevant challenge in the field of scientific research.

The accuracy of speaker identification becomes especially relevant when there is a need to recognize a person by one's voice. In the absence of 100% accurate systems that identify a speaker, their application is somewhat limited. For instance, speaker identification may be applied in the field of forensic investigation with the purpose to identify an individual uttering a

phrase; however, its application has certain limitations due to the lack of sufficiently high level of accuracy reliability. On the other hand, this method is highly useful in the areas where certain levels of accuracy error are allowed, thus enabling identification of a speaker with the purpose of providing more personalized information or additional services that require information about a person. According to Bimbot et al (2004), speaker identification is highly relevant in order to protect any type of interfaces where person's authentication is required. Hence, any increase in the accuracy of the speaker identification would expand the number of potential applications.

The task of speaker identification is also challenging because of a number of methods one needs to use. For successful speaker identification it is necessary to collect voice samples, find unique features and only then to perform classification.

Part of a successful process of speaker identification requires choosing the most suitable classifier. Based on a well-known "no free lunch" theorem, each classifier may only be sufficiently accurate to classify the specific data set (Goodfellow et al, 2016). The speaker identification problem is even more relevant for the experiments carried out with the sample set of Lithuanian native speakers, since there is very little research done around this specific demographic.

3. Research objectives

To identify a speaker from one's voice by classifying a speaker voice feature set extracted from a speaker's voice sample.

4. Research methods

Research methods used in this thesis are the following: literature analysis, theoretical analysis and experimental research. The author applied knowledge of biometric systems, digital signal processing, digital intelligence and pattern recognition theory.

5. The aim and tasks of the research

The aim of the work is to achieve higher classification accuracy for Lithuanian speakers.

With the view to achieve this aim it is necessary to:

1. Explore neural network identification accuracy for Lithuanian speakers comparing it to other methods of classification based on neural networks for voice recognition, and assess the efficiency of different neural networks when performing the task of speaker identification.
2. Create a classification method based on neural networks, which would demonstrate the highest identification accuracy of Lithuanian speakers.
3. To assess the accuracy of the proposed method by comparing it with other methods applied in speaker identification.

6. Scientific novelty

Scientific novelty of the present thesis is supported by the following:

1. It was proposed to use recurrent neural network configuration with 160 neurons and 3 hidden layers of long short-term memory topology to be used for Lithuanian native speaker identification.
2. It was demonstrated how application of this topology resulted in the best identification accuracy for different datasets.
3. Experimental research shows that Lithuanian speaker identification is most accurate when using a bidirectional long short-term memory neural network for classification. The improved accuracy compared to hidden Markov model method is from 3% to 6% higher on all tested datasets. It should be noted that in all experiment cases identification accuracy was always higher when a recurrent neural network was used for classification.

7. Practical significance

The proposed configuration of a neural network may be used to expand or provide better quality of services provided to Lithuanian-speaking users, and to offer individual services in cases where a person’s identity needs to be determined from a voice sample.

8. Statements to be defended

1. The accuracy of speaker identification is statistically significant with the classifier improvement, even if methods for feature extraction remain unaltered.
2. Recurrent neural network with a two-way long short-term memory topology identifies a Lithuanian speaker more accurately than the method based on hidden Markov models.
3. Increase in bi-directional long short-term memory neural network accuracy for Lithuanian speaker dataset classification, depends more on the total number of neurons, while the number of neural network layers is not as significant if the total number of neurons remains similar.

9. Approbation and publications of the research

The main results of the dissertation were published in 8 research papers: 2 papers were published in periodicals, reviewed scientific journals; 3 papers were published in conference proceedings; 3 papers were published in conference abstracts. The main results were presented and discussed at 5 national and 2 international conferences.

10. Outline of the Thesis

This thesis consists of 3 chapters titled Methods for Person Recognition from Voice, Speaker Identification Methods and Experimental Results. It also contains conclusions, a list of literature, and a list of publications. The total volume of the dissertation is 99 pages, 43 pictures, 18 tables and a total of 92 scientific references are cited.

1 METHODS FOR PERSON RECOGNITION FROM VOICE

Speaker identification is a challenging task due to the very 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, the possibility of a person’s recognition by one’s voice is appealing as this task does not require expensive equipment for

data collection compared to other means of biometric identification such as iris scanners or fingerprint readers.

1.1 Speaker identification from voice

The process of speaker identification may be summarized in three general stages (shown in Fig. 1):

1. Voice recording.
2. Feature extraction.
3. Classification model training.

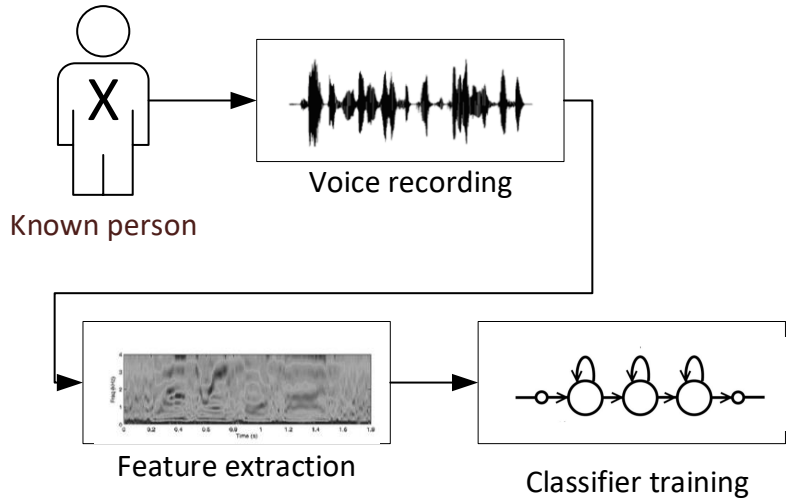

Fig. 1. Speaker enrollment process

In order to identify a speaker, one needs to do the following (shown in Fig. 2):

1. Record a voice sample.
2. Extract features.
3. Classify a speaker against applied models.

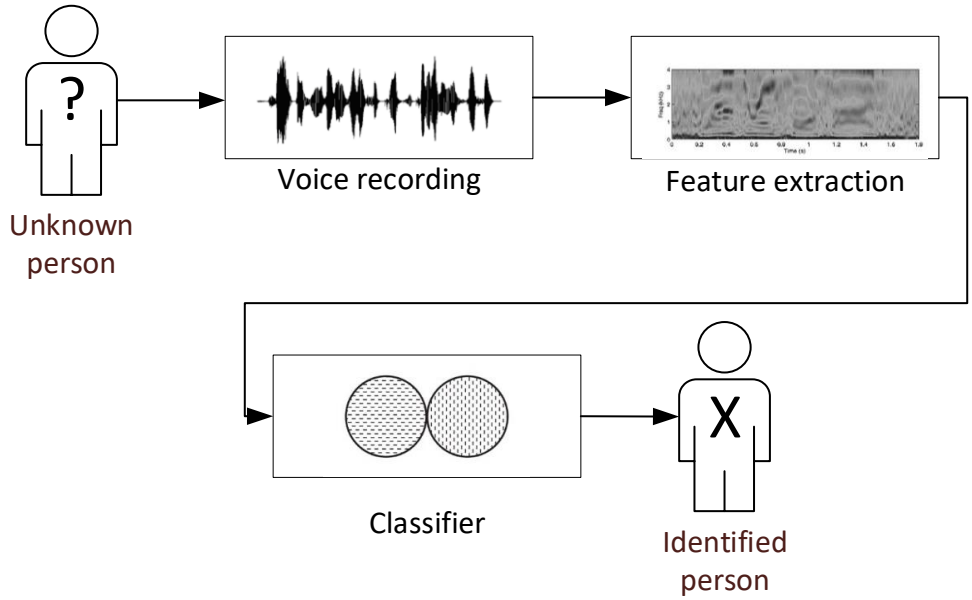

Fig. 2. Speaker recognition process

1.2 Feature extraction

The analysis of scientific literature on feature extraction showed that the audio features should have the following characteristics (Kinnunen, Li, 2009):

- Suitable separation between different speakers.
- Immutable to noise and environment.
- Difficult to impersonate.
- Unaffected by speakers' age or other variables.

There are multiple feature sets that may be used for this task: linear predictive cepstral coefficients, linear spectral frequencies, and perceptual linear prediction, mel frequency cepstral coefficients (Kinnunen, Li, 2009). Other methods include: partial correlation coefficients, log area ratios, formant frequencies, and bandwidths (Naik, 1990; Lipeika, 2005; Kinnunen, Li, 2009).

Mel frequency cepstral coefficients. It was further established that the most popular method for speaker feature extraction is mel frequency cepstral coefficients (MFCC) (Lipeika, 2005; Šalna et al, 2010; Kumar et al, 2011; Togneri, Pullella, 2011). This method is based on the human perception of voice. To supplement the MFCC feature set its first and second order derivatives are extracted additionally (Šalna et al, 2010; Ringelienė et al, 2011; Kumar et al, 2011).

Mel frequency is calculated in accordance with the following workflow:

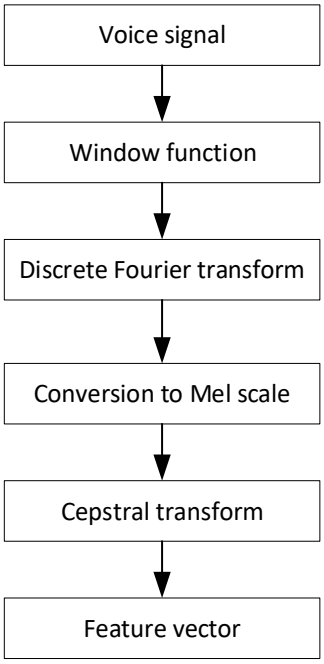

Fig. 3 Mel scale cepstral coefficient calculation workflow

As shown in Fig. 3, a voice signal is split into frames by designated duration and step. Window function is applied to all extracted frames:

$$w(n) = 0{,}54 - 0{,}46\cos\left(\frac{2\pi n}{N_s - 1}\right), \quad (1)$$

where $0 \leq n < N_s$, w(n)=0 in other cases. N_s – window's duration in a number of samples.

Next, for each windowed sample a discrete Fourier transform (DFT) is calculated. Afterwards, Mel filter banks are computed and the result of DFT is converted to Mel scale:

$$M(f) = 2595 log_{10}(1 + \frac{f}{700}). \quad (2)$$

Then, a discrete cosine transform, which is calculated for all frames, is applied. They are called mel frequency cepstral coefficients.

To calculate first and second order derivatives the following formula is used:

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2}, \quad (3)$$

where d_t is the first derivative in time *t*, calculated from MFCC $c_{t-\theta}$ to $c_{t+\theta}$, Θ derivatives window size.

In most cases, 39 coefficients were used in total; these coefficients contain the following features (Šalna, Kamarauskas, 2010; Togneri, Pullella, 2011):

- 12 MFCC;
- 1 MFCC energy element;
- 12 first order derivatives;
- 1 first order derivatives energy element;
- 12 second order derivatives;
- 1 second order derivatives energy element.

1.3 Feature classification accuracy results

Commonly, the starting choice of a classifier is made based on the type of classification, namely, either dependent on a text or independent from an uttered phrase. Text dependent classifiers include: Gaussian mixture models (Reynolds et al, 2000; Ding et al, 2014), neural networks (Nath, Kalita, 2014; Srinivas, Rani, 2014), hidden Markov models (Duda et al, 2000). If the process of recognition is performed by taking a fixed phrase or a spoken text set beforehand, hidden Markov models are usually used. Other methods include Universal Background Model (Reynolds et al, 2000), Joint Factor Analysis (Kenny, 2005), i-vectors (Lei et al, 2014; Richardson et al, 2015).

Research data (Fakotakis et al, 1997; Mahola et al, 2007; Abdallah et al, 2012; Deshmukh, Bachute, 2013; Reynolds, 1995; Reynolds, Rose, 1995; Narayanaswamy, 2005; Meglouli, Khebli, 2015; Aroon, Dhonde, 2017; Zheng et al, 2004; Cheng et al, 2010; Omar et al, 2010; Ganjeizadeh et al, 2014; Bawaskar, Kota, 2015; Jayanth, Roja, 2016) that served as a basis for the comparison is presented in the following table:

Table 1. HMM, GMM, UBM classification method accuracy comparison for speaker identification task

Classification method	**Speaker count**	**Identification accuracy**
HMM (average 97.80%)	5	97.40%
	20	84.50%
	40	100.00%
	630	98.09%
GMM (average 87.30%)	33	92.00%
	44	99.25%
	49	94.50%
	100	from 77.00%
UBM (average 74.12%)	11	96.69%
	22	95.45%
	42	61.90%
	50	67.50%
	463	74.40%
SVM (average 83.53%)	8	95.00%
	10	65.00%
	26	67.10%
	40	95.10%
	50	84.70%

1.4 **Neural networks for speaker acoustic modeling**

In the first chapter classifiers for speaker identification were based on the calculation of likelihood ratio. The other type of a classifier is neural network classifiers. Using the latter, significant breakthrough has been achieved in the field of scientific research over the recent year. Very accurate classification results are being obtained in various fields from image recognition to speech recognition.

There are two major neural network topologies (Krose, Smagt, 1996; Duda et al, 2000):

- Feed-forward neural networks, where data flow has forward direction and there is no feedback loop.
- Recurrent neural networks, where data flow is looped back to previous layer neurons and the information can be propagated backwards.

Examples of feed forward networks contain linear classifiers, multilayer perceptrons (Krose, Smagt, 1996). Recurrent neural networks include "vanilla" topologies (Schuster, Paliwal, 1997), long short-term memory networks (Hochreiter, Schmidhuber, 1997).

1.4.1 Recurrent neural networks

Recurrent neural network classifiers show good results in image and voice recognition tasks. In general, hidden and output layers of these networks may be defined as follows:

$$h_t = \mathcal{H}(W_{xh}x_t + W_{hh}h_{t-1} + b_h), \quad (4)$$

$$y_t = W_{hy}h_t + b_y, \quad (5)$$

where h = (h_1,…,h_T), y=(y_1,...,y_T) are hidden layer and output layer outputs, and W_{xh}, W_{hh}, W_{hy} are weight matrices, b_h, b_y is a bias and $\mathcal{H}$ is an activation function for a hidden layer.

A special case of a recurrent neural network is a bi-directional neural network. In this topology the weights of a hidden layer are calculated in both directions: forwards and backwards in time. This process is defined as follows:

$$\vec{h}_t = \mathcal{H}\left(W_{x\vec{h}}x_t + W_{\vec{h}\vec{h}}\vec{h}_{t-1} + b_{\vec{h}}\right) \quad (6)$$

$$\overleftarrow{h}_t = \mathcal{H}\left(W_{x\overleftarrow{h}}x_t + W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}\right) \quad (7)$$

$$y_t = W_{\vec{h}y}\vec{h}_t + W_{\overleftarrow{h}y}\overleftarrow{h}_t + b_y \quad (8)$$

where $\vec{h}$ is the flow forwards in time, and $\overleftarrow{h}$ stands for the information flow backwards.

1.4.2 Long short-term memory networks

Long short-term memory network is a recurrent neural network which may address the overtraining and exploding gradient problems that occur in the "vanilla" recurrent neural networks (Hochreiter, Schmidhuber, 1997). It uses the concepts of input, output and forget gates.

Inputs x_t are fed in a LSTM cell through an input gate i_t, a forget gate f_t, and an output gate o_t, and are calculated as follows:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}h_{t-1} + b_i) \quad (9)$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}h_{t-1} + b_f\right) \quad (10)$$

$$c_t = f_t c_{t-1} + i_t tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (11)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}h_{t-1} + b_o) \quad (12)$$

$$h_t = o_t tanh(c_t) \quad (13)$$

where σ is a sigmoid function and i_t, o_t, f_t, c_t are input, output, forget gates and cell activation vectors, respectively, W_{xi}, W_{hi}, W_{ci}, W_{xf}, W_{hf}, W_{xc}, W_{hc}, W_{cf}, W_{xo}, W_{ho}, W_{co} are weight matrices, b_i, b_f, b_c, b_o are bias vectors.

A special case of a LSTM is a bi-directional LSTM for which neuron weights are calculated in the same manner as with RNN networks, where calculations are performed in both directions: forwards and backwards in time.

1.4.3 Neural network identification accuracy

Research on speaker identification with neural networks is scarce. The results of research on speaker identification accuracy (Islam et al, 2013; Nath, Kalita, 2014; Chauhan, 2017; Ge et al, 2017; Moreno, Ho, 2003; Chen, Luo, 2009; Nijhawan, Soni, 2014; Sohanty, Swain, 2014, Ding et al, 2014) are summarized in Table 2.

Table 2. MP classification accuracy comparison for speaker identification task

Neural network type	Speaker count	Identification accuracy
Multilayer perceptron	6	from 95.00%
	8	87.50%
	30	50.00%
	83	91.00%
	200	93.29%

There is very limited research data on using LSTM type of neural networks for the task of speaker identification. A phoneme identification accuracy study, which is closest to speaker identification, was carried out by Graves and Schmidhuber (2005). The results of their research are presented in Table 3.

Table 3 Phoneme identification accuracy comparison for various neural networks

Graves, A., Schmidhuber, J., 2005

Neural network topology	Identification accuracy	
	Training sample (3696 phonemes)	Test sample (1344 phonemes)
Multilayer perceptron	67.60%	63.10%
Recurrent neural network	69.90%	64.50%
Bi-directional recurrent neural network	76.00%	69.00%
Long short-term memory	77.60%	66.00%
Bi-directional long short-term memory	78.60%	70.20%

2 SPEAKER IDENTIFICATION METHODS

In order to determine improvement in accuracy a search algorithm workflow is presented in Fig. 4.

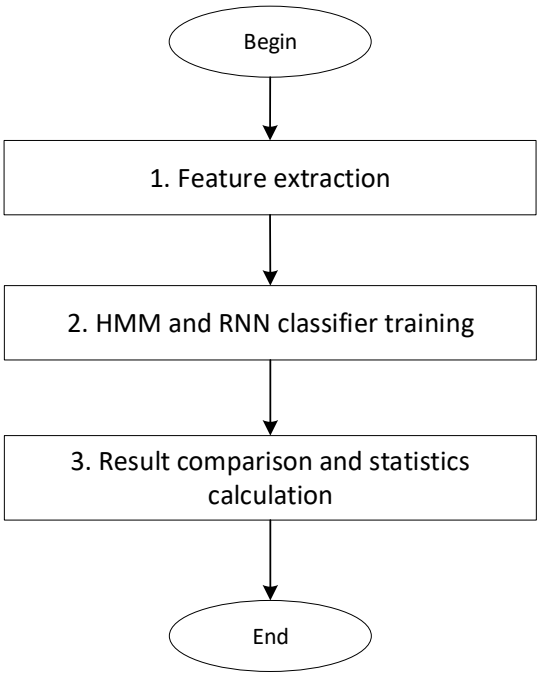

Fig. 4 Identification accuracy testing workflow

1. In order to extract features from speaker voice samples the following steps need to be performed:
    a. Calculation of MFCC for each speaker.
    b. Separation of training and testing samples.
    c. Creation of a list of speakers containing only unique speakers.
    d. Creation of individual lists for each speaker's training and testing samples.
2. Training hidden Markov models and recurrent neural networks, and finding the best model require the following actions:
    a. Configuration of a classifier based on defined hyper-parameters.
    b. Classification of all the samples until a stopping condition is satisfied.
    c. Testing a classifier with testing sample data.
    d. Saving the results for further analysis.

3. To calculate and compare the results for statistical analysis it is necessary to:
    a. Sort and group results.
    b. Calculate statistics.

2.1 Classifier hyper-parameter search

In order to obtain the most accurate identification results, a classifier needs to be fine-tuned for specific hyper-parameters. A grid search method to find the best parameter set is applied.

HMM classification. In the HMM hyper-parameter configuration one of the most significant configuration changes may be achieved by altering the number of hidden states in a model (Fig. 5).

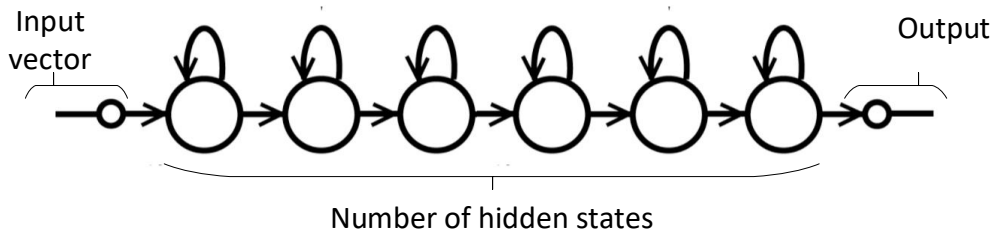

Fig. 5. Hidden Markov model chain

For the initial search 5, 7, 10, 16 and 22 hidden states are taken.

Neural network classification. To create a neural network with highest accuracy the same grid search approach is applied as with HMM parameters.

Principal neural network architecture is shown in Fig. . Here input layer is a MFCC feature vector and a hidden layer contains long short-term memory cells. The output layer is a softmax classifier with an output that provides identification reliability value.

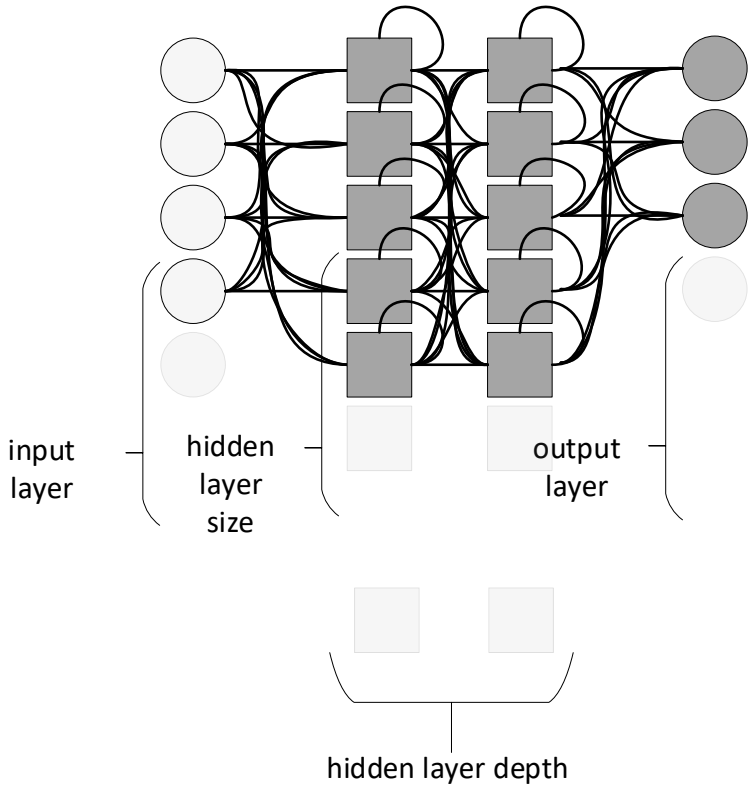

Fig. 6. Neural network generic architecture

Initial hyper-parameter grid for HMM and RNN is provided in Table 4.

Table 4. Hyper-parameter search table.

Step	**Initial parameter**	**Adjusted parameter**
HMM	Number of hidden states 5; 7; 10; 16; 22	Adjusted number of hidden states -2; -1; +1; +2;
RNN 1st step	Architecture LSTM 80;	Architecture BLSTM 80;
RNN 2nd step	Dropout: 0.2; 0.4; 0.6;	Dropout: -0.1; +0.1;
RNN 3rd step	Number of cells: +80; +160;	Number of layers: +1; +2;

2.2 Proposed method for speaker identification

The proposed method for speaker identification uses recurrent neural network structure which has two layers of long short-term memory cells with 160 neurons each and where an output layer is a fully connected layer with

softmax activation. The structure of the proposed neural network is shown in Fig. 7.

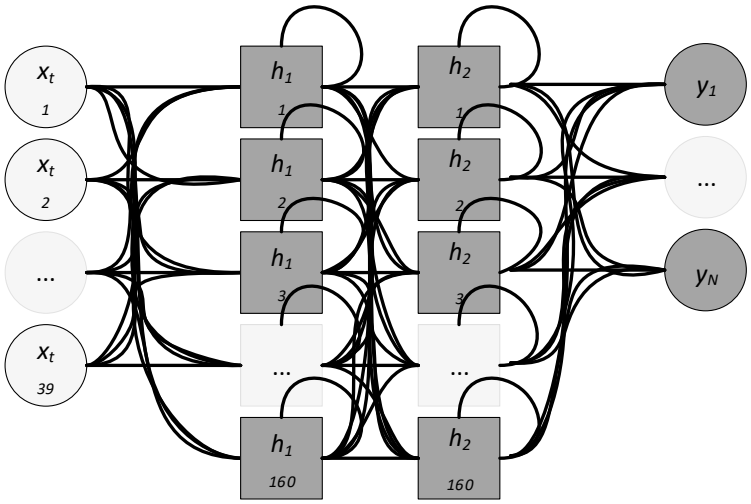

Fig. 7 Suggested neural network architecture

Neural network inputs $x_{t1}, x_{t2}, \ldots, x_{t39}$ have 39 dimension MFCC vectors with time series of t with the step from 1 to 999, which is extracted from each speaker's audio file.

Neurons $h_{1\text{-}1}$, $h_{1\text{-}2}$, ... $h_{2\text{-}160}$: the LSTM cells (Graves, Mohamed and Hinton, 2013) calculated from (9) with input gates (i_t) calculated from (10), forget gates (f_t) calculated from (11), a cell gate (c_t) calculated from (12) and an output gate (o_t) (12)(13).

Output layer $y_1, \ldots, y_N$: a fully connected layer with a softmax activation function, where N stands for all speakers.

Training of proposed neural network. The training process consists of learning of network weights W_{xi}, W_{hi}, and W_{ci} which are initialized according to the methods proposed by Glorot and Bengio (2010). Internal cell weights W_{xf}, W_{hf}, W_{cf}, W_{xc}, W_{hc}, W_{cf}, W_{xo}, W_{ho}, and W_{co} are initialized by the method proposed by Saxe, McClelland and Ganguli (2014).

Bidirectionality of LSTM cells is developed by using RNN bidirectionality principal from the method proposed by Schuster and Paliwal (1997) when a learning procedure has a forward and a backward flow as detailed in chapter 1.4.1.

Network uses sparse categorical cross entropy loss (Boer et al, 2005) function with Adam optimization (Kingma and Ba, 2015).

3 EXPERIMENTAL RESULTS

With the view to conduct a comparison of identification accuracy results based on classification with hidden Markov models versus recurrent neural networks, more specifically, long short-term memory variations, an initial hyper-parameter set and model tuning were developed and a speaker dataset LIEPA was selected.

3.1 Speaker dataset description

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).

Only a certain part of this dataset was taken for the purpose of conducting experimental tests. This allowed running initial accuracy tests more efficiently and in a less time-consuming manner. Hence, the initial experiment has 66 unique speakers, which equals to 66 individual classification sets. The selected speaker subset contains a total of 4691 unique audio samples.

Percentage-wise, 70% of the selected dataset samples are used for training the models, and 30% of samples are used for accuracy testing. 70%/30% splits are done at the speaker level which gives at least 8 unique samples per speaker that are excluded from model training.

The entire dataset is split into 5 equal parts. This allows finding the best classifier configuration within similar samples sets, as outlined in Table 5.

Alternative dataset is created by adding noise to original LIEPA dataset.

Table 5. Full dataset splits.

Partial dataset	**1st set**	**2nd set**	**3rd set**	**4th set**	**5th set**
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 Experimental results

The initial results for the grid search on HMM were obtained. As these results demonstrate, the best performing HMM are those with 3 and 5 hidden states. For RNN the best configurations of initial dataset are BLSTM 240 drop-out 0.3 and 2xBLSTM 160 drop-out 0.0/0.3.

Full dataset accuracy testing. Full list of accuracy results for split datasets is presented in Table 6. Data shown in the table allows determining the best performing models for HMM and RNN configurations, accordingly.

Table 6. Accuracy results for split datasets.

Classifier	**1st set**	**2nd set**	**3rd set**	**4th set**	**5th set**
HMM 3 states	89.29%	82.35%	84.53%	93.10%	90.00%
HMM 5 states	70.75%	60.62%	62.23%	89.76%	77.23%
BLSTM 240 dropout 0.3	90.30%	83.91%	92.48%	97.66%	95.82%
2xBLSTM 160 dropout 0.0/0.3	92.08%	86.45%	92.35%	97.10%	96.39%

The best performing model of HMM (3 hidden states) is compared to the best RNN variant (2xBLSTM 160 drop-out 0.0/0.3) in Fig. 8. Improvement in speaker identification accuracy from 3% to 6% when comparing HMM to RNN was observed.

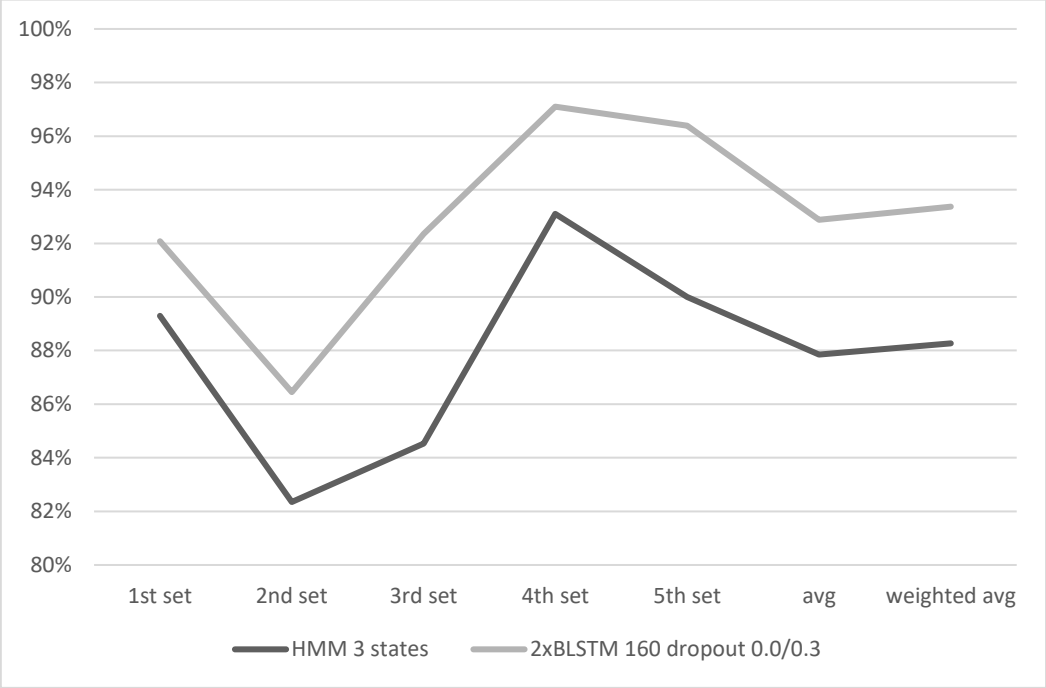

Fig. 8 HMM vs. RNN speaker identification accuracy results comparison for full 5 part dataset

Statistical significance was calculated for all test sets by using McNemar's chi square test (Dietterich, 1998). The results are provided in Table 7.

Table 7. Error and statistical significance analysis for test datasets

Classification result	**1st set**	**2nd set**	**3rd set**	**4th set**	**5th set**
#1 HMM and RNN correct	1085	727	1273	1641	1022
#2 HMM and RNN not correct	38	52	61	24	31
#3 only HMM correct	66	118	99	33	77
#4 only RNN correct	100	129	190	100	91
McNemar's test statistic	6.5602	0.4048	28.027	32.751	1.0059
p-value (1 tail)	0.0052	0.2623	0.0000	0.0000	0.1579
p-value (2 tails)	0.0104	0.5245	0.0000	0.0000	0.3158

Cross validation testing results. In order to verify accuracy of suggested classifiers, a stratified cross validation (Arlot, Celisse, 2010) check was completed and results were presented in Table 8.

Table 8. Accuracy results for split datasets with cross validation

Classifier	**1st set**	**2nd set**	**3rd set**	**4th set**	**5th set**
HMM 3 states	92.27% +/-0.94%	85.75% +/-1.32%	89.41% +/-0.64%	96.37% +/-0.72%	93.41% +/-1.31%
2xBLSTM 160 dropout 0.0/0.3	93.71% +/-0.98%	87.05% +/-2.05%	93.51% +/-0.82%	97.03% +/-0.62%	96.60% +/-1.12%

Cross validation with noisy dataset. To check classifier performance in real case scenarios white noise was added to the original dataset. Identification results are presented in Table 8 and Table 9.

Table 9. Accuracy results for noisy datasets with cross validation

Classifier	**1st set**	**2nd set**	**3rd set**	**4th set**	**5th set**
HMM 3 states	80.06% +/-1.63%	71.61% +/-2.41%	77.35% +/-1.52%	82.50% +/-1.47%	71.95% +/-2.01%
2xBLSTM 160 dropout 0.0/0.3	90.25% +/-0.86%	82.58% +/-0.79%	91.06% +/-1.31%	90.57% +/-1.48%	86.37% +/-1.92%

CONCLUSIONS

1. It was found that the increase in speaker identification accuracy may be achieved by improving a classifier without changing common feature extraction methods. Upon conducting the experiments it was observed that speaker classification performed by applying hidden Markov models was less accurate than by applying recurrent neural networks. The best speaker identification accuracy obtained by using hidden Markov model classifier ranges from 82.35% to 93.10%, while by using neural networks it ranges from 86.45% to 97.10%.
    a. The results obtained by a stratified cross validation with 10 folds for neural network showed increased accuracy of 0.66% to 4.10% in all cases compared to hidden Markov models.

b. With the application of noisy signals speaker identification accuracy with recurrent Markov models increased from 10.19% to 14.42% compared to hidden Markov models.

2. Application of a recurrent neural network with a bi-directional long-term short-term memory topology allows identifying a Lithuanian speaker more accurately compared to the results obtained when a method based on hidden Markov models was applied. Experimentally it was shown on the Lithuanian speaker dataset LIEPA and the improvement from 3% to 6% was achieved on all testing sets compared to HMM classifier. The increase in accuracy was statistically significant for 3 datasets out of 5. When a cross validation method was used the increase in accuracy was observed every time.
3. The accuracy of BLSTM network increases due to the total number of neurons, while the number of layers is not as significant if the number of neurons remains similar. Experiments show that the accuracy of identification is more influenced by the total number of neurons rather than the number of hidden layers. It was noted that in the most accurate configurations, such as BLSTM 160 and BLSTM 160x160, Lithuanian speaker identification accuracy differs only by 0.55%.

ĮVADAS

1. Tyrimo sritis

Darbo tyrimo sritis yra asmens tapatybės identifikavimas naudojant balso įrašus. Tyrimo sritis priklauso informatikos inžinerijos mokslo sričiai, signalų technologijos mokslo šakai. Tyrimo sritis apima keleto uždavinių sprendimą: balso signalų apdorojimą, diktoriui identifikuoti tinkamų požymių radimą, diktoriaus balso požymių klasifikavimo uždavinį. Šiame darbe koncentruojamasi į diktoriaus balsą aprašančių požymių klasifikavimo problemą. Šios problemos sprendimas reikalauja daugelio skirtingų algoritmų panaudojimo. Kiekvienas algoritmas slepia savyje nemažus iššūkius: nuo balso pavyzdžių surinkimo, unikalių požymių išskyrimo iki požymių klasifikavimo ir jų priskyrimo konkrečiam diktoriui.

2. Problemos aktualumas

Asmens identifikavimas pagal balsą yra vienas iš patogiausių asmens biometrinio identifikavimo būdų. Problema yra aktuali, nes egzistuojantys diktoriaus identifikavimo algoritmai ir sistemos neatpažįsta diktoriaus 100 proc. tikslumu, tad atliekant identifikavimą visuomet lieka klasifikavimo klaidos tikimybė. Pagrindinė to priežastis yra ypač didelis signalo variatyvumas. Tai, kad kiekvieno diktoriaus kiekvienas ištarimas yra vis kitoks, kad balsas gali būti įtakotas ligos ar bandymų pamėgdžioti identifikuojamą asmenį, yra tik keli tokio variatyvumo pavyzdžiai. Diktoriaus balso nepastovumas yra pagrindinė priežastis, kodėl, nepaisant nemenkų pastangų, šioje srityje diktoriaus identifikavimo negalima laikyti tobulai išspręstu uždaviniu ir tema išlieka aktuali moksline prasme – diktoriaus identifikavimo tikslumą galima padidinti.

Diktoriaus identifikavimo tikslumas ypač aktualus, kai reikia atpažinti asmenį pagal jo balsą. Kadangi nėra sistemų, identifikuojančių diktorių 100 proc. tikslumu, jų pritaikymas yra apribotas. Pavyzdžiui, diktoriaus identifikavimas yra aktualus kriminalistikoje, siekiant atpažinti frazę ištarusį asmenį, tačiau neturint pakankamai aukšto tikslumo patikimumo, pritaikymas šioje srityje yra ribotas. Kitose srityse, kuriose yra leidžiama didesnė tikslumo

paklaida, identifikuojant pašnekovą galėtume pateikti daugiau personalizuotos, asmeninės informacijos, suteikti papildomų paslaugų. Anot Bimbot ir kt. (2004), diktoriaus identifikavimo uždavinys yra labai aktualus norint apsaugoti bet kokio tipo sąsajas ar autentifikuojant asmenį. Todėl bet koks diktoriaus identifikavimo tikslumo padidinimas suteiktų galimybę plačiau panaudoti identifikavimą pagal balsą minėtose srityse.

Iš kitos pusės, diktoriaus identifikavimo uždavinys taip pat yra sudėtingas dėl atpažinimui reikalingų įvairių metodų panaudojimo komplekso. Kad galima būtų sėkmingai identifikuoti diktorių, reikia surinkti jo balso pavyzdžius, išskirti šių balso pavyzdžių požymius ir atlikti klasifikavimą.

Kadangi diktoriaus identifikavimas yra klasifikavimo problema, sprendžiant klasifikavimo uždavinius gerai žinoma „Nėra nemokamų pietų“ (angl. *no free lunch*) teorema, kuri sako, kad kiekvienas klasifikatorius gali būti pakankamai tikslus tik klasifikuojant tam tikros srities specifinį uždavinį (Goodfellow, Bengio ir Courville 2016). Taigi, sprendžiamos diktoriaus identifikavimo problemos aktualumas yra dar didesnis, nes darbe eksperimentas yra atliekamas su lietuviškai kalbančių diktorių garsynu, o didelio pavyzdžių kiekio diktoriaus identifikavimo tyrimų su lietuvių kalba beveik nėra atlikta.

3. Tyrimo objektas

Disertacijos tyrimo objektas – diktoriaus identifikavimas pagal akustinį balso signalą, taikant rekurentinius neuroninius tinklus.

4. Tyrimo metodai

Disertacijos uždaviniai yra sprendžiami naudojant literatūros apžvalgos, teorinės analizės ir eksperimentinio tyrimo metodus. Teorinei analizei ir tyrimams panaudotos biometrikos sistemų, skaitmeninio signalų apdorojimo, skaitmeninio intelekto bei atpažinimo teorijos žinios.

5. Tikslas ir uždaviniai

Darbo tikslas − pasiūlyti klasifikavimo metodą, skirtą lietuviškai kalbančio diktoriaus identifikavimui pagal balsą, paremtą rekurentiniais neuroniniais tinklais.

Darbo tikslui pasiekti sprendžiami uždaviniai:

1. Ištirti neuroninių tinklų efektyvumą lietuviškai kalbančio diktoriaus identifikavimui, palyginant su kitais diktoriaus atpažinimo metodais. Įvertinti skirtingų neuroninių tinklų tipų efektyvumą diktoriaus atpažinimo uždaviniui.
2. Pasiūlyti geriausią rekurentinių neuroninių tinklų topologiją lietuviškai kalbančio diktoriaus identifikavimui.
3. Įvertinti pasiūlyto klasifikavimo metodo tikslumą skirtingomis sąlygomis, naudojant skirtingos kokybės signalus.

6. Mokslinis naujumas

Šio darbo mokslinį naujumą sudaro tokie elementai:

1. Sukurta rekurentinio neuroninio tinklo konfigūracija su 160 neuronų ir trijų paslėptų sluoksnių dvikrypčio ilgos trumpalaikės atminties topologijos konfigūracija, skirta identifikuoti lietuviškai kalbantį diktorių pagal jo balsą.
2. Parodyta, kad tokia topologija leidžia pasiekti aukščiausią diktoriaus identifikavimo tikslumą skirtingiems duomenų rinkiniams.
3. Eksperimentiniais tyrimais nustatyta, kad lietuviškai kalbančių diktorių identifikavimui naudojant dvikrypčio ilgos trumpalaikės atminties rekurentinio neuroninio tinklo klasifikatorių, gaunami nuo 3 proc. iki 6 proc. tikslesni diktoriaus identifikavimo rezultatai, nei atlikus akustinį modeliavimą su paslėptaisiais Markovo modeliais.

7. Darbo rezultatų praktinė reikšmė

Pasiūlyta neuroninio tinklo klasifikatoriaus konfigūracija gali būti naudojama siekiant pagerinti lietuviškai šnekantiems vartotojams teikiamų paslaugų kokybę ar išplėsti jų pasiūlą, teikti individulių paslaugų pasiūlymus, kai asmens tapatybę reikia nustatyti naudojant balsą.

8. Ginamieji disertacijos teiginiai

1. Lietuviškai kalbančio diktoriaus identifikavimo tikslumą galima statistiškai reikšmingai padidinti, turimam diktoriaus akustinį signalą aprašančių požymių rinkiniui parenkant tinkamesnį klasifikatorių.

2. Neuroninio tinklo klasifikatoriaus topologija, naudojant dvikryptės ilgos trumpalaikės atminties rekurentinį neuroninį tinklą, lietuviškai kalbančius diktorius identifikuoja tiksliau, nei klasifikuojant diktorius, naudojant paslėptuosius Markovo modelius.
3. Nustačius dvikrypčio ilgos trumpalaikės atminties rekurentinio neuroninio tinklo topologiją, naudojant nedidelį sluoksnių skaičių, su kuria stebimas geriausias lietuviškai kalbančio diktoriaus identifikavimo tikslumas, ir toliau keičiant tik paslėptų neuronų sluoksnių gylį, bet paliekant panašų bendrą neuronų skaičių, gaunamas tik nežymiai besiskiriantis identifikavimo tikslumas.

9. Darbo rezultatų aprobavimas

Disertacijos rezultatai buvo paskelbti 2 periodiniuose ir 6 kituose leidiniuose.

Disertacijos rezultatai pristatyti šiose tarptautinėse konferencijose:

1. BIS 2017 „20th International Conference on Business Information Systems“. Pranešimas: „Identifying Lithuanian Native Speakers Using Voice Recognition“. 2017-06-29, Poznanė, Lenkija.
2. BIS 2016 „19th International Conference on Business Information Systems“. Pranešimas: „Speaker Authentication System Based on Voice Biometrics and Speech Recognition“. 2016-07-07, Leipcigas, Vokietija.

Skaityti pranešimai respublikinėse konferencijose:

1. International Conference „Electrical, Electronic and Information Sciences“ 2018. Pranešimas: „Diktoriaus identifikavimas naudojant BLSTM tipo neuroninio tinklo konfigūraciją“. 2018-04-26, Vilnius, Lietuva.
2. 9th International Workshop „Data analysis methods for software systems“. Pranešimas: „Speaker Recognition Using Deep Neural Networks“. 2017-11-30, Druskininkai, Lietuva.

3. 8th International Workshop „Data analysis methods for software systems“. Pranešimas: „Speaker Recognition Using Deep Neural Networks“. 2016-12-01, Druskininkai, Lietuva.
4. 7th International Workshop „Data analysis methods for software systems“. Pranešimas: „GMM-UBM Enhancement with DNNs for Automated Speaker Recognition Systems“. 2015-12-03, Druskininkai, Lietuva.
5. Informacinės technologijos 2015. 20-oji tarpuniversitetinė magistrantų ir doktorantų konferencija. Pranešimas: „Asmens balso panaudojimas autentikavimui“. 2015-04-24, Kaunas, Lietuva.

10. Disertacijos struktūra

Disertacijos apimtis 99 lapai, 18 lentelių, 43 paveikslai. Disertacijoje remtasi 92 literatūros šaltiniais. Disertaciją sudaro įvadas, 3 skyriai: Asmens atpažinimo pagal balsą metodai, Diktoriaus identifikavimo algoritmai, Eksperimentinio tyrimo rezultatai, išvados ir literatūros sąrašas.

IŠVADOS

1. Eksperimentais nustatyta, kad diktoriaus identifikavimo tikslumą galima padidinti naudojant ilgos trumpalaikės atminties rekurentinius neuroninius tinklus, palyginus su diktoriaus identifikavimui dažniausiai naudotais paslėptais Markovo modeliais. Diktoriaus identifikavimo tikslumo pagerinimas pasiektas naudojant tuos pačius požymių rinkinius, taigi vien dėl neuroninio tinklo klasifikavimo metodo naudojimo. Geriausias diktoriaus identifikavimo tikslumas naudojant paslėptuosius Markovo modelius svyruoja tarp 82,35 proc. ir 93,10 proc., tuo tarpu naudojant rekurentinius neuroninius tinklus svyruoja tarp 86,45 proc. ir 97,10 proc su tais pačiais duomenimis.
    a. Pakartojus tyrimą naudojant atsitiktinio kryžminio validavimo metodą iš 10 pjūvių, gauti rezultatai su neuroniniu tinklu visais

atvejais yra geresni nuo 0,66 iki 4,10 proc. nei naudojant paslėptuosius Markovo modelius.

b. Naudojant užtriukšmintus signalus, diktoriaus identifikavimo tikslumas naudojant rekurentinius Markovo modelius padidėja nuo 10,19 iki 14,42 proc., lyginant su paslėptais Markovo modeliais.

2. Tai leidžia teigti, kad pasiūlyta neuroninio tinklo klasifikatoriaus topologija, naudojant dvikryptės ilgos trumpalaikės atminties rekurentinį neuroninį tinklą lietuviškai kalbančių diktorių identifikavimui, yra pranašesnė negu paslėptais Markovo modeliais paremti diktoriaus identifikavimo metodai: naudojant rekurentinį neuroninį tinklą diktoriaus identifikavimo tikslumas padidėja nuo 3 proc. iki 6 proc., lyginant su tiksliausiu klasifikatoriumi, naudojančiu paslėptuosius Markovo modelius. Gauti rezultatai yra statistiškai reikšmingi ir statistinis reikšmingumas yra stebėtas trijose bandymų aibėse iš penkių. Naudojant kryžminio validavimo metodą, geresni rezultatai stebėti visų tyrimų atvejais.
3. Eksperimentiškai nustatyta, kad tarp įvairių konfigūracijų ilgos trumpalaikės atminties rekurentinių neuroninių tinklų (pavyzdžiui, tarp geriausius rezultatus parodžiusių dvikrypčio ilgos trumpalaikės atminties 160 neuronų tinklo ir dvikrypčio ilgos trumpalaikės atminties dviejų sluoksnių 160x160 tinklo topologijos), lietuviškai kalbančio diktoriaus identifikavimo tikslumas skiriasi tik 0,55 proc.

AUTORIAUS PUBLIKACIJŲ SĄRAŠAS DISERTACIJOS TEMA

Straipsniai publikuoti periodiniuose recenzuojamuose leidiniuose:

1. Dovydaitis, L. ir Rudžionis, V. 2018, „Building bi-directional LSTM neural network based speaker identification system“, *Computational Science and Techniques*, vol. 6, no 1, p. 574–580. DOI: 10.15181/csat.v6i1.1579
2. Dovydaitis, L. ir Rudžionis, V. 2018, „Speaker identification accuracy improvement using BLSTM neural network“, *Indian Journal of Computer Science and Engineering (IJCSE)*, vol. 9, no. 2, p. 31–37. DOI: 10.21817/indjcse/2018/v9i2/180902017

Straipsniai recenzuojamuose konferencijų leidiniuose:

1. Dovydaitis L. ir Rudžionis V. 2017, „Identifying Lithuanian Native Speakers Using Voice Recognition“ in *Abramowicz W. (eds) Business Information Systems Workshops BIS 2017. Lecture Notes in Business Information Processing*, p. 79–84. DOI: 10.1007/978-3-319-69023-0_8
2. Dovydaitis L., Rasymas T. ir Rudžionis V. 2017, „Speaker Authentication System Based on Voice Biometrics and Speech Recognition“ in *Abramowicz W., Alt R., Franczyk B. (eds) Business Information Systems Workshops BIS 2016. Lecture Notes in Business Information Processing*, p. 79–84. DOI: 10.1007/978-3-319-52464-1_8
3. Dovydaitis, L. ir Rudžionis, V. 2015, „Asmens balso panaudojimas autentikavimui“ in *Informacinės technologijos 2015, 20-oji tarpuniversitetinė magistrantų ir doktorantų konferencija, Kaunas, Lietuva, balandžio 24 d., 2015 : konferencijos pranešimų medžiaga,* Kaunas, Lietuva, p. 147–151.

Santraukos konferencijų leidiniuose:

1. Dovydaitis, L. ir Rudžionis, V. 2017, „Speaker Recognition Using Deep Neural Networks“ in *Data analysis methods for software systems: 9th*

International Workshop, Druskininkai, Lithuania, November 30 – December 2, 2017, Druskininkai, Lithuania, p. 15.

2. Dovydaitis, L. ir Rudžionis, V. 2016, „Speaker Recognition Using Deep Neural Networks“ in *Data analysis methods for software systems: 8th International Workshop, Druskininkai, Lithuania, December 1–3, 2016*, Druskininkai, Lithuania, p. 18.
3. Dovydaitis, L. ir Rudžionis, V. 2015, „GMM-UBM Enhancement with DNNs for Automated Speaker Recognition Systems“ in *Data analysis methods for software systems: 7th International Workshop, Druskininkai, Lithuania, December 3–5, 2015*, Druskininkai, Lithuania, p. 18.

TRUMPOS ŽINIOS APIE AUTORIŲ

Darbo patirtis	
Datos	2016.12 – Dabar
Profesija arba pareigos	Debesijos sprendimų architektas
Darbovietės pavadinimas	Microsoft
Datos	2016.05 – 2016.12
Profesija arba pareigos	IT Sprendimų architektas
Darbovietės pavadinimas	CSC
Datos	2012.07 – 2016.01
Profesija arba pareigos	Infrastruktūros sprendimų vadovas
Darbovietės pavadinimas	Blue Bridge UAB
Dėstymo patirtis	
Datos	2015.01 – 2017.06
Įstaigos pavadinimas	ISM Vadybos ir ekonomikos universitetas
Datos	2015.01 - 2017.07
Įstaigos pavadinimas	Vilniaus Universiteto Verslo Mokykla
Išsilavinimas	
Datos	2013.10 – 2017.09
Kvalifikacija ir įstaiga	Doktorantūros studijos, Vilniaus universitetas
Datos	2011.09 – 2013.06
Kvalifikacija ir įstaiga	Verslo informatikos magistras, Vilniaus universitetas

Laurynas Dovydaitis

RESEARCH ON THE ACCURACY OF LITHUANIAN SPEAKER'S IDENTIFICATION USING RECURRENT NEURAL NETWORKS

Summary of Doctoral Dissertation
Technological Sciences
Informatics Engineering (07 T)
Editor Silvija Naranovič-Skurdauskienė

Laurynas Dovydaitis

LIETUVIŠKAI KALBANČIO DIKTORIAUS IDENTIFIKAVIMO NAUDOJANT GRĮŽTAMOJO RYŠIO NEURONINIUS TINKLUS TIKSLUMO TYRIMAS

Daktaro disertacijos santrauka
Technologijos mokslai
Informatikos inžinerija (07 T)
Redaktorė Dagnė Pakalniškytė-Alseikienė