

## SINGIDUNUM UNIVERSITY

### Department of Postgraduate Studies and International Cooperation

Referring to the Decision of the Department for Postgraduate Studies of Singidunum University, Belgrade, as well as on the basis of consideration of PhD thesis of the candidate **Ashrf Nasef**, under the title: "**Speech Recognition in noisy environment using Deep Learning Neural Network**" hereby we submit the following:

## REPORT ON EVALUATION OF THE DOCTORAL DISSERTATION

### 1. Biography of the PhD Candidate

Basic biographical information on the candidate are as follows:

**Name:** Ashrf Nasef

**Nationality:** Libyan

**Date and place of birth:** 09.08.1982. Zliten

**Marital status:** Married

**Phone Number:** +38161 6330936

**Address:** Kostolacka 25A/1 Belgrade, Serbia

**E-mail:** ashraf8259@yahoo.com, ashrfnasef@gmail.com

**Education:** Primary School: "Ahmad Rafek Almhdwe" Zliten, Libya, 1989-1995

Preparatory School: "Ahmad Rafek Almhdwe" Zliten, Libya, 1995-1998

High School: "Zliten Secondary School" Zliten, Libya, 1998-2001

**Undergraduate:** “Elmergib University” Zliten, Libya. 2001-2004, (Bachelor degree in Computer Science)

**Postgraduate:** Faculty of technical sciences, University Novi Sad. Master in Electrical and Computer Engineering, 2010-2012

**Work:** Elmergib University Faculty of Arts and sciences, Zliten Libya 2007-2009

**Publications:**

- **Conference paper:** Ashrf Nasef, Marina Marjanović. Development of an Open Source Based Software Tool for Arabic Text Stemming and Classification, International Scientific Conference on ICT and E-Business Related Research - Sinteza 2016.
- **Conference paper:** Ashrf Nasef, Marina Marjanović, Angelina Njeguš. Optimization of the speaker recognition in noisy environments using a stochastic gradient descent, Proceedings of International Scientific Conference on Information Technology and Data Related Research, Apr, 2017.
- **Journal Paper:** Ashrf Nasef, Marina Marjanović-Jakovljević, Angelina Njeguš. Stochastic gradient descent analysis for the evaluation of a speaker recognition. Analog Integrated Circuits and Signal Processing, Volume 90, Issue 2, pp 389–397, ISSN: 0925-1030 (Print) 1573-1979 (Online), Springer, February, 2017.

## 2. Research scope and problem

The speech signal contains the information about the language (acoustic phonetic symbols), prosody (intonation signals), gender (vocal tract and pitch - frequency of voiced sounds), age, accent (formants), speaker's identity, emotion and health [1], [2], [3]. While the aim of speech recognition is to recognize the spoken words in speech, speaker recognition identifies the speaker by recognizing the spoken phrase, and verifies the speaker [4]. The abilities of decoding the speech signals, understanding the linguistic and speaker information in speech, and recognizing the speaker, are needed in many speech aided applications, such as access control, access to confidential information, voice command control, transaction authentication, and audio archive indexing [5].

Speaker recognition systems, generally, perform the three main tasks: speaker identification, speaker verification/detection, and speaker classification [4]. Identification takes a speech utterance of an unknown speaker and compares with predefined speaker model of valid users. Feature matching process finds the best match that is used to identify the unknown speaker. In speaker verification, the unknown speaker first claims identity, and then one-to-one matching is done, i.e. the claimed model is used for identification. If the match is above a predefined threshold, the identity claim is accepted. The basic model for speaker recognition system is shown in Fig. 1.

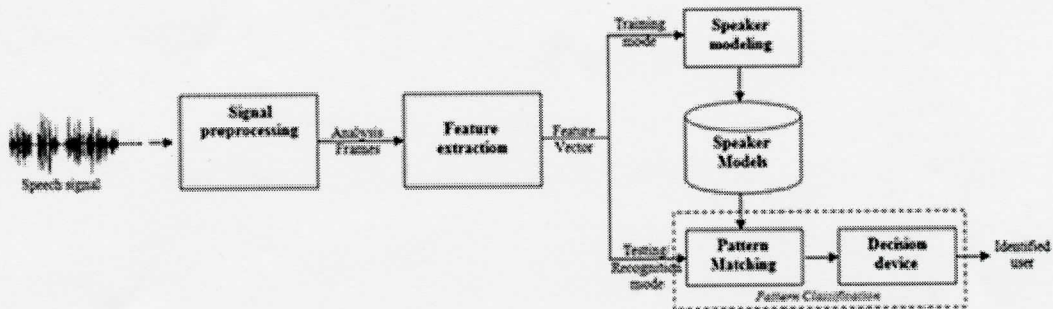


Figure 1. General speaker recognition system architecture.6

In order to recognize the speaker, the first step in speaker recognition system is to convert the speech waveform, using digital signal processing tools to a set of features for further analysis. During signal preprocessing, sometimes when removing an unwanted information, some useful information can be lost. After preprocessing of the speech signal in the signal modeling, the next step is parameterization of speech signal, which is called feature extraction. The aim of this process is to produce a meaningful representation of the speech signal. Some main tasks of feature extraction are the process of converting the speech signal to a digital form (signal conditioning), measuring important characters of the signal (signal measurement), augmenting these measurements with derived measurements (signal parameterization), and statistical modeling. For parametrically representing the speech signals there are several common methods, such as Mel-Frequency Cepstrum Coefficients, Filter Bank Energy analysis, and others [7].

The result of feature extraction is the sequence of acoustic vectors that are extracted from an input speech of each speaker and provide a set of training vectors (patterns) for that speaker [8]. The next step is pattern/feature matching that identifies the unknown speaker by comparing extracted features with the set of known speakers. The pattern classification measures the similarity of the input feature vectors, and groups the patterns that share the same properties. The pattern classification results in whether to accept or reject a speaker [9].



### **3. Related Work**

Recently, researchers have started to explore several different strategies for using Deep Neural Networks (DNN) for speech recognition, speaker recognition, and spoken language recognition tasks [4], [10], [11]. Among the first results, in 2000, was the use of Deep Belief Networks, which are based on restricted Boltzmann machines, and deep autoencoders [12]. However, training DNN with big number of hidden layers with autoencoders, has shown to be quite difficult task [13]. From 2006, dimensionality of data with autoencoder networks was reduced by gradient descent which is used for fine-tuning the weights [14]. Furthermore, this approach has branched into a major variants, such as batch gradient descent, stochastic gradient descent, and mini-batch gradient descent [15]. When dealing with continuous speech recognition, the Recurrent Neural Networks (RNN) were proposed [16].

One of the recent advances in DNN, that improve its performance, optimization, and prediction quality, are rectified linear units, and dropout (to overcome overfitting) [17]. However, there are still few challenges that need to be addressed. Sutskever et al. (2013) showed the importance of momentum-accelerated Stochastic Gradient Descent (SGD) that uses well-designed random initialization.<sup>18</sup> Le et al. (2011) introduced more sophisticated optimization methods such as Limited memory Broyden-Fletcher-Goldfarb-Shanno (BFGS), and conjugate gradient, that simplify, and speed up the process of pretraining deep algorithms [15]. Senior, et al (2013) trained DNN for large vocabulary speech recognition with minibatch SGD by using a variety of learning rate schemes. They show that adequate choice of learning rate schemes leads to faster convergence, and lower word error rates [19].

In this paper, we present the model of SGD with and without dropout rate applied on speech dataset. We analyze the different combinations of its parameters such as learning rate, hidden and input layer dropout rate.

Additionally, the speech recognition performance fluctuation with different background noise levels is analyzed and the impact of employing of different coding techniques in speech recognition is determined.

### **4. The Aim of the Research.**

New wave of consumer-centric applications, such as speech recognition in order to interact with mobile devices and home entertainment systems. In order to improve automatic speech recognition in real environment (the noisy environments), the major contribution of the thesis are as follows:

Analysis the speech recognition performance fluctuation with different background noise levels.  
Determining the impact of employing coding techniques in speech detection recognition.

Many researches have shown that deep learning neural network methods have the best performance in comparison with other classifiers. However, those methods with many parameters require a lot of tunings in order to optimize the performance in different supervised learning tasks.

In this thesis, we show that picking a good combination of parameters can significantly improve the performance of deep learning Neural Network methods in automatic speaker recognition even in a noisy environment. Additionally, improvement of the speech recognition by using deep convolutional networks on spectrograms will be considered.

## **5. The methodology of scientific research**

First, the effect of external influences on speech recognition in noisy environments with different levels of Signal to Noise Ratio will be examined. Second, the speech recognition performance by tuning various parameters in Stochastic Gradient Descent DNN algorithms will be analyzed. Third, speech recognition by using deep Convolutional Neural Networks on spectrograms for a spoken language identification task will be considered in order to improve the performance of the previous tasks.

## **6. The content of the thesis**

This thesis contains 6 main chapters. In the first chapter, introduction about motivation and speaker recognition is given. Second chapter contains the State-of-The-Art of the speaker recognition area. Chapter 3 describes the feature extraction. Chapter 4 explains the working principle of Neural Network methods. In chapter 5, experiments are presented and the chapter 6 the conclusion remarks are given.



## 7. Scientific contribution

Recent researches in the field of automatic speaker recognition have shown that methods based on deep learning neural network provide better performance than other classifiers based on hidden Markov models and Gaussian mixture models. On the other hand, those methods with many parameters require a lot of tunings in order to optimize the performance in different supervised learning tasks. The goal of this thesis is to show that selecting appropriate value of parameters can significantly improve the performance of deep learning neural network methods in automatic speaker recognition. The reported study introduces an approach to automatic speaker recognition based on deep neural networks and the stochastic gradient descent algorithm. It particularly focuses on three parameters of the stochastic gradient descent algorithm: the learning rate, and the hidden and input layer dropout rates. Additional attention was devoted to the research question of speaker recognition under noisy conditions. Thus, two experiments were conducted in the scope of this thesis. The first experiment was intended to demonstrate that the optimization of the observed parameters of the stochastic gradient descent algorithm can improve speaker recognition performance under no presence of noise. This experiment was conducted in two phases. In the first phase, the recognition rate is observed when the hidden layer dropout rate and the learning rate are varied, while the input layer dropout rate was constant. In the second phase of this experiment, the recognition rate is observed when the input layers dropout rate and learning rate are varied, while the hidden layer dropout rate was constant. The second experiment was intended to show that the optimization of the observed parameters of the stochastic gradient descent algorithm can improve speaker recognition performance even under noisy conditions. Thus, different noise levels were artificially applied on the original speech signal. The obtained results show that dropout optimization can significantly enhance the performance of stochastic gradient descent method in automatic speaker recognition even under noisy conditions. It is also shown that selecting an appropriate value of the learning rate is also a very important task, since for some values of this parameter, the performance of the method is negatively affected.

## 8. References.

- [1] DOUGLAS E. COMER, "Computer Networks and Internets", Fifth Edition, Pearson Education, Inc. Upper Saddle River, New Jersey, 2009.
- [2] Netgear, "Wireless Networking Basics", Netgear, Inc., USA, 2005.
- [3] Walter Goralski, "Differences In Addressing Between IPv4 and IPv6", Juniper Networks, 2014.
- [4] Rick Lehtinen, "Computer Security Basics", 2nd Edition, O'Reilly, 2006.

- [5] Kintu Zephernia, "Migrating to IPv6", KTH Information and Communication Technology, 2012.
- [6] Mohammad Mirwais Yousafzai, Nor Effendy Othman and Rosilah Hassan, "Toward IPv4 to IPv6 Migration within a Campus Network", *Journal of Theoretical and Applied Information Technology*, 77(2): 209-217, 2015.
- [7] Amer Nizer Abu Ali, "Comparison Study Between IPv4 & IPv6", *International Journal of Computer Science Issues*, 9(3): 314-317, 2012.
- [8] Rajinder Singh, "A comparison of Security Features of Ipv4 and Ipv6", *International Journal of Advanced Research in Computer Science and Software Engineering*, 5(7): 826-828, 2015.
- [9] Emre Durdađı and Ali Buldu, "IPV4/IPV6 security and threat comparisons", *Procedia - Social and Behavioral Sciences*, Elsevier, 2(2): 5285-5291, 2010, DOI:10.1016/j.sbspro.2010.03.862.
- [10] Monjur Ahmed, Alan T Litchfield and Shakil Ahmed, "VoIP Performance Analysis over IPv4 and IPv6", *I.J. Computer Network and Information Security*, 11:43-48, 2014, DOI: 10.5815/ijcnis.2014.11.06.
- [11] Fred Baker, "IPv4/IPv6 Coexistence and Transition", *IETF Journal*, 4(3): 16-17, 2009.
- [12] Guoqiang ,Bruno Quoitin and Shi Zhou, "Phase Changes in the Evaluation of the IPv4 and IPv6 As-Level Internet Topologies", *Computer Communication Journal*, Elsevier B.V., 34(5): 649-657, 2011, DOI:10.1016/j.comcom.2010.06.004.
- [13] Dan Wing, "Network Address Translation: Extending the Internet Address Space", *IEEE Internet Computing*, 14(4): 66-70, 2010, DOI: 10.1109/MIC.2010.96.
- [14] Sun Microsystems, "Making the Transition from Ipv4 to Ipv6 (Reference)", in *Ipv6 Administration Guide*, edited by Sun Microsystems Inc., 2003.
- [15] Shalini Punithavathani and Shery Radley, "Performance Analysis for Wireless Networks: An Analytical Approach by Multifarious Sym Teredo", *The Scientific World Journal*, 2014: 1-9, 2014, DOI: 10.1155/2014/304914.
- [16] Lisa Phifer, "Measure wireless network performance using testing tool iPerf", *SearchNetworking.com*, retrieved on 3/5/2015, <http://goo.gl/a6P4eE>.
- [17] Srikanth Sundaresan, Nick Feamster and Renata Teixeira, "Measuring the Performance of User Traffic in Home Wireless Networks", in *Passive and Active Measurement*, edited by Jelena Mirkovic and Yong Liu, Springer International Publishing, 8995: 305-317, 2015, DOI: 10.1007/978-3-319-15509-8\_23.

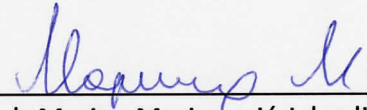
## 9. Opinion of the Committee Members

Na osnovu iznetog, članovi Komisije su zaključili da kandidat mr Ashrf Nasef ispunjava sve uslove iz zakonskih propisa i odgovarajućih zahteva iz opštih akata Univerziteta „Singidunum“ u Beogradu za odbranu doktorske disertacije. Stoga članovi Komisije predlažu Veću Departmana za posle diplomске studije i međunarodnu saradnju Univerziteta „Singidunum“ u Beogradu da kandidatu mr Ashrfu Nasefu, odobri izradu doktorske disertacije pod radnim naslovom Speech

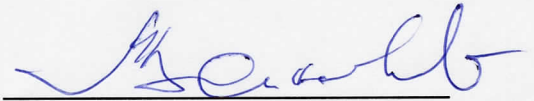
Recognition in noisy environment using Deep Learning Neural Network ", a za mentora rada se predlaže prof. dr Marina Marjanović-Jakovljević.

Belgrade, 29.05.2017.

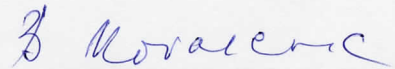
Committee Members



prof. dr Marina Marjanović-Jakovljević,  
mentor  
Singidunum University



Prof. dr Mladen Veinović, member  
Rector of Singidunum University



Prof. dr Branko Kovačević, member  
Faculty of Electrical Engineering  
University of Belgrade