Application of Neural Networks in Whispered Speech Recognition

Đorđe T. Grozdić*, Branko Marković, Jovan Galić, and Slobodan T. Jovičić

Abstract — **This paper presents the preliminary results of experimental research of whispered speech recognition that was based on the application of artificial neural networks (ANN). The paper also describes a speech database of words that were spoken in a whisper and normal manner, which was created especially for this study. A part of this database was used for preliminary training and testing of the ANN. Mel Frequency Cepstral Coefficients (MFCC) in normal speech and whispered speech were used as an input to the ANN. The case of speaker dependent recognition was tested and ANNs with optimal topologies have a 97.98% accuracy in speech recognition and 96.21% in whisper recognition for a male speaker. The results for a female speaker were very similar. In the case of whisper recognition, when ANN was trained for normal speech the score of whisper recognition was 75.71% for a male speaker (82.14% for female), and vice versa, when ANN was trained for whisper, normal speech recognition was 82.14% for a male speaker (90% for female).**

Keywords — **Neural networks, speech recognition, whisper recognition, whispered speech database.**

I. Introduction

WHISPERED speech is a specific form of speech that is sometimes used in verbal communication. It is whispered in different situations, for example when we want to make a discreet or an intimate atmosphere in conversation, in the library so as not to disturb other people, or when someone tries to conceal some confidential information from the ears of other people. Whispered speech is often used in criminal activities, especially in telephone conversations where criminals try to disguise their identity. However, in addition to the conscious production of a whisper, whispering may occur due to health problems which appear after rhinitis and laryngitis, but for some people it seems like a chronic disease of larynx structures [1], [2].

By its nature and mechanism of production the whisper is significantly different from usual speech. It is characterized by a noisy structure and the absence of glottal vibrations. Due to the absence of glottal vibrations, whispering lacks the fundamental frequency of voice, intonation contours and, consequently, the other prosodic information. It was found that vowel formants at lower frequencies were shifted to higher frequencies, and that the slope of the spectrum of whisper was much flatter than in speech [3], [4]. In addition, whispered speech has a significantly lower energy as compared to normal speech [2].

The mentioned features of whispered speech are a significant problem in speech technology, especially in speech synthesis, speech recognition, as well as in the identification of the speaker. Therefore, whispered speech is a hot topic in recent research [1], [5]. On the other hand, it is interesting that this type of speech communication, in spite of increased efforts in perception, performs perfectly understandable. The question is: how does whispered speech have such a high intelligibility despite its significant differences regarding normal speech?

There are different approaches, techniques and methods of speech recognition. These techniques are usually based on algorithms of the HMM (Hidden Markov Model), the DTW (Dynamic Time Warping), the ANN (Artificial Neural Network) and their hybrid solutions [6]. Due to the similarity of the ANN with the structure of human brain and its way of speech perception, it was hypothesized that the ANN could yield good results in the recognition of whispered speech. To analyze this hypothesis, a study of the application of the ANN to the recognition of whispered speech in comparison to normal speech was performed.

As an input to the ANN, Mel Frequency Cepstral Coefficients (MFCCs) in normal speech and whispered speech were used. MFCCs are one of the most widely used cepstral features in speech analysis computed from the log-energies in frequency bands distributed over a mel-frequency scale [6].

The paper is organized as follows: in Section 2, a description of speech corpus is given, Section 3 contains a description of feature extraction of speech stimuli used as an input to the ANN, Section 4 describes the characteristics of the ANN, Section 5 presents experimental results, and the Conclusion summarizes the results and indicates directions for further research.

II. Corpus Description

In order to confirm the effectiveness of the ANN in whispered speech recognition, this study uses the Whi-Spe

*The 2012 “Blažo Mirčevski” Award for the young first author of the paper presented at TELFOR 2012.

This paper was supported by the Ministry of Education and Science of the Republic of Serbia under Grant numbers OI-178027 and TR-32032.

Đorđe T. Grozdić, is with the Life Activities Advancement Center, Laboratory for Psychoacoustics and Speech Perception, Belgrade, and Ph. D. candidate at School of Electrical Engineering, University of Belgrade, Serbia (e-mail: djordjegrozdic@gmail.com).

Branko Marković, is with the Čačak Technical College, Čačak, Serbia, and Ph. D. candidate at School of Electrical Engineering, University of Belgrade, Serbia (e-mail: branko333@open.telekom.rs).

Jovan Galić, is with the Faculty of Electrical Engineering, University of Banja Luka, Bosnia and Herzegovina, and Ph. D. candidate at School of Electrical Engineering, University of Belgrade, Serbia (e-mail: jgalic@etfbl.net).

Slobodan T. Jovičić, is with the School of Electrical Engineering, University of Belgrade, and with the Life Activities Advancement Center, Belgrade, Serbia (e-mail: jovicic@etf.rs).

(Whispered Speech) corpus especially developed for this purpose.

Whi-Spe corpus contains 50 whispered/normal paired words: 14 numbers, 6 colors and 30 words. Words were taken from the Serbian emotional speech database GEES [7], which satisfies the basic linguistic criteria of Serbian language (phonemes distribution, syllable composition, accentual structure, consonant clusters). The whispered and normal speech was collected from 5 male and 5 female speakers. Each speaker had read all 50 words ten times in both speech modes, so the Whi-Spe corpus contains 10.000 recorded words. A specific file notation was established to simplify future use of corpus.

The corpus was recorded last year in a quiet laboratory room using an Optimus omni-directional tie-clip microphone at a distance of 25cm from the speaker's mouth for neutral speech and 5cm for whispered speech. Speech data was digitalized using a sampling frequency of 22.050Hz, with 16 bits per sample, in Windows PCM wav format.

In each session, subjects pronounced the whole set of 50 words continuously in both speech modes, and sessions were separated by longer time intervals for several days. The recordings were manually segmented into words and each word was one entry to the Whi-Spe corpus. Quality control of the recordings found various errors, of subjective and objective nature, so pronunciations of some words had to be repeated.

For this experiment a part of corpus, that contains recordings of spoken words (numbers) of two speakers (one male and one female) was used. We considered pronounciation of 14 different numbers. Every number was pronounced 10 times in normal speech and 10 times in whispered speech by each speaker. Accordingly, this part of corpus consists of 140 spoken numbers and 140 whispered numbers per speaker. So, the experiment was designed to analyse speaker dependent recognition of normal speech in comparision to recognition of whispered speech.

III. Features extraction

In spoken word recognition based on ANN we have used a fixed number of frames for MFCC feature extraction. There are two methods to get a fixed number of frames for speech signals if they have different lengths, as in the case of the Whi-Spe corpus: (1) dynamic numbers of sample points over windows, and (2) dynamic windows overlap rates [9]. The first method was used, and each word from the Whi-Spe corpus was segmented into eleven frames that overlap 50%, using the *Hamming* windows. The number of frames (eleven) was determined by the statistical distribution of the number of phones in words from our speech database. The range of the number of phones per word is from 3 to 9 (with the exception of two words that have 12 and 13 phones). The average number of phones per word is 5.58, so the most common case is to have 4, 5, or 6 phones per word. Using 11 frames per word gives an average of one frame per phone in long words, while in short words there are two or three frames per phone. We assumed that this finer temporal and spectral resolution of shorter words should enable their better identification. On the other hand, long words have richer phonetic content.

MFCC extraction procedure is shown in Fig. 1.

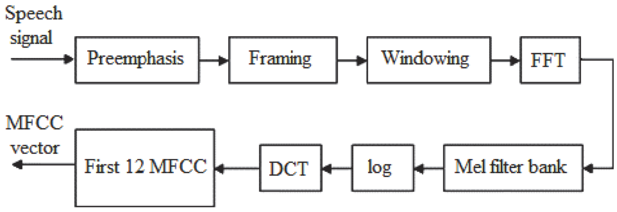

Fig. 1. Block diagram of MFCC extraction procedure.

Every speech frame was represented by three feature vectors: 12 MFCC coefficients, 12 delta MFCC coefficients (MFCC first derivatives) and 12 delta-delta MFCC coefficients (MFCC second derivatives). Finally, each vector consists of 396 coefficients per word, including MFCC and their first and second derivatives, in both normal and whispered speech. This feature extraction procedure was done for every stimuli-word. Obtained vectors, representing the words, were aligned in two matrices of dimensions 396x140 coefficients - one for speech and another for whisper per speaker. Later, these matrices were used as an input database for training and testing of ANNs.

IV. Neural Network

In this paper *Feedforward* ANN with *Back Propagation* algorithm in training process was used. ANN was realized as multilayer perceptrons (MLP) using MATLAB *Neural Network Toolbox* [10]. Two identical networks were formed – one for normal speech and the other for whispered speech. Each network had three layers: the input, the hidden and the output layer as shown in Fig. 2.

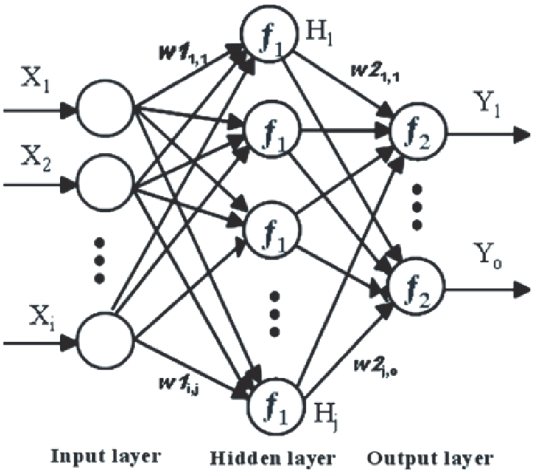

Fig. 2. A three-layer feedforward neural network arhitecture.

The input layer contained 396 input nodes. The output layer contained 14 neurons as the number of words that have to be recognized. The number of hidden neurons was changed during the experiment in order to find ANN topology that has the best performance. *Tansig* (a hyperbolic tangent sigmoid) function was used as the transfer function of neurons.

Neural Network Toolbox has a predefined algorithm and default values that are set for training of neural networks in MATLAB. One of these default settings is a division of training database into three parts: 70% of database for training, 15% for validation and 15% for testing network performances. Data which form these 70%, 15% and the remaining 15% of database are selected stochastically from the entire database. Therefore there is a risk, for example in the case of our experiment in word recognition, that some of the 14 different words are not included at all in the part of database that is used for network training. The consequences of this way of network training are: (1) inability of ANN to recognize a word that is not included in training process, and (2) expressed variability in network performance through training sessions. For this reason, this algorithm in MATLAB is further configured and adapted to the needs of this research.

The database in this study is divided into three parts: 60% of the database is used for training, 20% for validation and 20% for testing. During the division of database it was taken into account that from all ten pronunciations of each word, six word pronunciations are randomly taken for training, other two for validation and the last two for testing of the network. The first part of the database has dimensions 396x84 coefficients, while the second and the third part have dimensions 396x28 coefficients. This kind of database division ensures that the ANN will be trained to recognize each of the 14 words, and variability in network performance after repeated trainings will be significantly reduced.

The networks were trained using the *trainscg* function which is based on the scaled conjugated gradient algorithm. This algorithm was developed by *Martin Moller* and represents a combination of the *Levenberg-Marquardt* algorithm and the scaled conjugated gradient principle [11]. *Trainscg* network training requires a large number of iterations but it significantly reduces the total number of arithmetic operations and the time required to train the network, which makes this function suitable for large neural networks.

The criteria for stopping the training of the network were: the defined maximum number of iterations (1000), the mean square error (0.00), the maximum number of consecutive errors in validation, the so-called *early stopping* method (6), and the minimum gradient (10^{-6}). The analysis of network topology, in terms of the number of hidden layers, showed no significant influence on the results of recognition, so we stuck to the original network structure with three layers.

V. Results

A. Analysis of hidden layer

The first step was to find ANN architecture that has the best performance for this specific purpose. For this reason the starting architectures of two ANNs with 15 neurons in hidden layer were formed. Networks were trained separately - one for normal speech and the other one for whispered speech. The number of hidden neurons in networks was later gradually increased by steps of 15 neurons up to 120 neurons. ANN performances were monitored.

Before the training of ANN, the network has initial values of weighting coefficients of connections between neurons. These values are randomly generated numbers. During the training process, these coefficients are given new values and they are changing in accordance with the database on which the network is trained. When training is completed, weighting coefficients are adjusted and the network is ready for simulation. The simulation is performed by bringing all 140 input vectors to the network and recording the success that is achieved in word recognition. The values of weighting coefficients are then returned to the initial (random) values, so the network could be re-trained. Each considered network topology was trained and simulated in this way 50 times, and the averaged results of their performances are shown in Tables 1 and 2.

TABLE 1: WORD RECOGNITION FOR MALE SPEAKERS IN NORMAL AND WHISPERED SPEECH DEPENDING ON NUMBER OF HIDDEN NEURONS.

Number of neurons	*Normal speech recognition (%)*	*Whispered speech recognition (%)*
15	92.97	91.81
30	94.90	93.29
45	96.58	95.44
60	96.93	93.36
75	97.98	96.21
90	95.47	93.84
105	95.63	95.53
120	96.42	93.65

TABLE 2: WORD RECOGNITION FOR FEMALE SPEAKERS IN NORMAL AND WHISPERED SPEECH DEPENDING ON NUMBER OF HIDDEN NEURONS.

Number of neurons	*Normal speech recognition (%)*	*Whispered speech recognition (%)*
15	93.99	92.01
30	94.48	94.15
45	95.97	94.44
60	94.80	92.64
75	97.46	95.08
90	94.80	94.80
105	94.37	93.86
120	95.62	93.57

The results for male speakers are shown in Table 1 and for female speakers in Table 2. As we can see, word recognition in all cases is above 90%. With an increasing number of hidden neurons the performances of ANNs are getting better. The maximum of word recognition is reached at 75 hidden neurons in both male and female speakers.

This number of neurons is very well matched with the theoretical prediction based on the "geometric pyramid rule" [9]:

$$N_h = (N_i \times N_o)^{1/2} \qquad (1)$$

where N_h is the number of hidden neurons, N_i is the number of input neurons, N_o is the number of output neurons, which in our case gives 75 neurons. With a further increase in the number of neurons, the performances of ANN are slightly decreased.

B. Speech and whisper recognition

The usual problem of the ASR (*Automatic Speech Recognition*) systems occurs at the point when a speaker switches from normal speech to whisper, or vice versa [5]. The problem is related to different train/test scenarios in the ANN application. So we made another experiment that analyzes whispered speech recognition with the ANN that was trained to recognize normal speech, and vice versa, normal speech recognition with the ANN that was trained to recognize whispered speech.

This experiment tested two networks with 75 neurons in the hidden layer which have achieved a recognition of 100% - one trained for normal speech and the other one for whispered speech. Such networks are created and then simulated in speech/whisper and whisper/speech scenarios. The results for both speakers are given in Fig. 3.

In general, the mismatch train/test scenarios, such as speech/whisper and whisper/speech, showed significantly lower recognition scores. For instance, a male speaker had word recognition of 75.71% in the case of speech/whisper scenario and 82.14% in the whisper/speech scenario. For these two scenarios a male speaker showed lower recognition scores than a female speaker, especially in the case of whisper/speech scenario, where a female speaker had a word recognition of 90%.

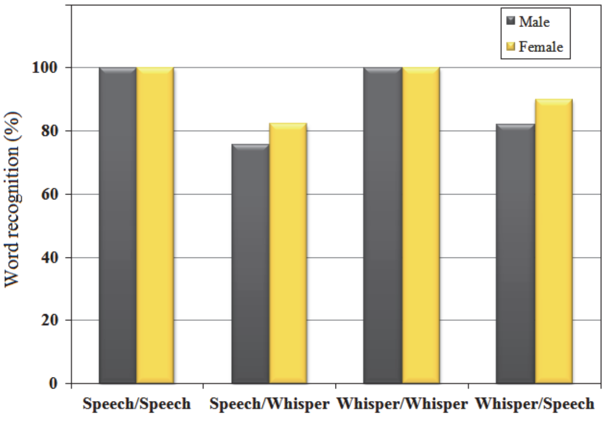

Fig. 3. Results of word recognition rate depending on modality Train/Test.

The relative relations of results in Fig. 3 are in agreement with the results of the experiments performed on the HMM based recognition method [1] and GMM (*Gaussian Mixture Models*) based recognition method [3].

VI. Conclusion

In this study we examined the application of ANN in whispered speech recognition. The motivation for this kind of study and experiments is the fact that whisper is a serious problem for ASR systems despite its perfect understandability in human face-to-face communication.

In these preliminary experiments, it is shown that ANN with MFCC features, extracted from speech and whisper signals, can give high scores in word recognition for both speech and whisper. We found out the optimal number of hidden neurons in ANN that gives the best word recognition scores. Whisper and speech recognition had similar scores - around 97%, while the mismatch train/test scenarios showed significantly worse word recognition. The interesting fact is that whisper/speech scenario had higher recognition scores (86% on the average) than speech/whisper scenario (79% on the average). Further evaluation of these phenomena will be in the focus of our future investigation.

It will be also interesting to find out and compare efficiency in word recognition of popular ASR systems based on different algorithms such as DTW and HMM in case when a speaker switches from normal speech to whisper. We should also examine different speech and whisper features, different methods for their extraction, and other types of ANNs.

References

[1] T. Ito, K. Takeda, F. Itakura, “Analysis and Recognition of Whispered speech,” *Speech Communication*, 2005, pp. 129-152.

[2] S. T. Jovičić, Z. M. Šarić, „Acoustic analysis of consonants in whispered speech,” *Journal of Voice*, 22(3), 2008, pp. 263-274.

[3] C. Zhang, J.H.L. Hansen, “Analysis and classification of Speech Mode: Whisper through Shouted,” *Interspeech 2007*, 2007, pp. 2289-2292.

[4] S. T. Jovičić, “Formant feature differences between whispered and voiced sustained vowels,” *ACUSTICA - Acta Acoustica*, 84(4), 1998, pp. 739-743.

[5] X. Fan, J.H.L. Hansen, “Speaker identification within Whispered Speech Audio Stream,” *IEEE Transactions on Audio, Speech and Language Processing*, 19(5), 2011, pp. 1408-1421.

[6] J. Holms, W. Holms, “Speech Synthesis and Recognition,” Taylor & Francis, London, 2001.

[7] S. T. Jovičić, Z. Kašić, M. Đorđević, M. Rajković, “Serbian emotional speech database: design, processing and evaluation,” *SPECOM-2004*, St. Petersburg, Russia, 2004, pp. 77-81.

[8] S. T. Jovičić, S. Punišić, Z. Šarić, “Time-frequency detection of stridence in fricatives and affricates,” *Int. Conf. Acoustics’08,* Paris, 2008, pp. 5137-5141.

[9] S.-T. Pan, C.-C. Lai, “Using genetic algorithm to improve the performance of speech recognition based on artificial neural network,” chapter in: M. Grimm, K. Kroschel (Eds.), *Robust Speech Recognition and Understanding*, I-Tech Education and Publishing, Vienna, Austria, 2007.

[10] T. Masters, “Practical Neural Network Recipes in C++,” Academic Press, NY, 1993.

[11] H. Demuth, M. Beale, "Neural Network Toolbox User's Guide," The MathWorks, Inc, 2002.