International Journal of
Applied Mathematics,
Electronics and Computers
ISSN: 2147-8228 http://ijamec.atscience.org

Advanced Technology and Science

Original Research Paper

Automatic Voice and Speech Recognition System for the German Language with Deep Learning Methods

Çiğdem Bakır *[1]

Accepted 3rd September 2016

Abstract: In our age, technological developments are accompanied by certain problems associated with them. Security takes the first place amongst such kind of problems. In particular, such biometric systems as authentication constitute the significant fraction of the security matters. This is because sound recordings having connection with the various crimes are required to be analyzed for forensic purposes. Authentication systems necessitate transmission, design and classification of biometric data in a secure manner. In this study, analysis of German language employed in the economy, industry and trade in a wide spread manner, has been performed. In the same vein, the aim was to actualize automatic voice and speech recognition system using Mel Frequency Cepstral Coefficients (MFCC), MelFrequency Discrete Wavelet Coefficients (MFDWC) and Linear. Prediction Cepstral Coefficient (LPCC) taking German sound forms and properties into consideration. Approximately 2658 German voice samples of words and clauses with differing lengths have been collected from 50 males and 50 females. Features of these voice samples have been obtained using wavelet transform. Feature vectors of the voice samples obtained have been trained with such methods as Boltzmann Machines and Deep Belief Networks. In the test phase, owner of a given voice sample has been identified taking the trained voice samples into consideration. Results and performances of the algorithms employed in the study for classification have been also demonstrated in a comparative manner.

Keywords: *Boltzmann Machines, Deep Belief Networks*

1. Introduction

In our day, security problems have been unearthed along with the developments in the technology. Certain studies have been accomplished especially in order to prevent information belonging to some people from being transferred to other people in commercial transactions. Some of such studies are hand script recognition, signature recognition, face recognition, iris recognition and voice recognition [9].

German language belongs to the Germanic branch of Indo-European language family. Approximately 120 million people are speaking German language in the world. In addition, Germany has an important standing in respect of economy, trade, industry and many other fields in the international sense. For this reason German language is used in a quite widespread manner. However, common usage of this language to such an extent, brings security problems for biometric data in this field. Accordingly this calls for the requirement of a secure, fast automatic voice and speaker recognition.

German language comprises of roots (words) and suffixes & prefixes and included in the inflected languages if we consider properties of the German language. German is written using the Latin alphabet and there are 29 letters in its alphabet. An article appears before each noun in German. Words are pronounced as they are written. In addition, it is distinguished from other languages with various developed sound shifts and intonation.

Various studies have been carried out in order for voice and speaker recognition. Jie-Fu et al. have collected voice samples in Chinese from 7 males and 5 females whose ages were ranging between 25 and 45 [8]. Attempts have been made to identify the owner of the voice by trying to analyze these voice samples by means of their tones, vowels, consonants and syllables. Voice samples have been separated into four frequency groups, and each frequency band has been analyzed. However, this study has not been tested for very big data. In addition, intended success was not exactly achieved since it was performed taking its similarities with the English language into consideration.

Tokuda et al. have developed English speech synthesis system using Hidden Markov Model [4]. This system has been developed for speaker recognition and specifies the structure by changing the voice feature. However characteristic feature of the synthesized voice in the study, is pretty low.

Reynolds et al. have implemented SuperSID project in order to enhance performance of speaker recognition systems [1]. Purpose of this project is to develop speaker recognition systems and employ the most suitable features in order to increase its accuracy. However this study failed to completely achieve the acoustic characteristics of the voice and removal of the noise.

Speech has an important plae in communication.Voice recognition study has been carried out for this reason. In this study, a simulation also has been performed in order to solve the voice recognition problem related to security risk. However certain difficulties have got in the way while creating voice database. There was such difficulty ranked first among the others that words were vocalized at different speeds and in different pronunciation by different persons. In addition to that, such reasons as the noise occurred in the environment and voice while recording the voice data, toning effect and syllable stress make

[1] *Computer Engineering Department, Electric-Electronic Faculty, Yildiz Technical University, Campus, 34015, İstanbul/Turkey*
** Corresponding Author: Email: cigdem@ce.yildiz..edu.tr*
Note: This paper has been presented at the 3rd International Conference on Advanced Technology & Sciences (ICAT'16) held in Konya (Turkey), September 01-03. 2016.

voice recognition process difficult [10].
Feature extraction and classification techniques used, were given in the section 2 of the study performed, and experimental study and results were given in the section 3.

2. Feature Extraction Methods

German language is widely used in economy, industry and trade. Therefore, examinations have been made on German language in this study. The study has been realized on a unique data base, which have been formed from the German sound samples, taken from men and women. These sound samples are trained by getting dispersed to various feature vectors with MFDWC, MFCC and LPCC. In the second stage, the feature vectors of the recorded sound signals are trained with classification algorithms, such as Boltzmann Machines and Deep Belief Networks. The gender of the speaker is decided by looking at sound signals at the test data and training data after the system is trained. Furthermore, the classification success in recognizing the gender of speaker has been calculated separately for 1, 3 and 5 and 9 feature vectors and the success of the methods have been presented comparatively by training the feature vectors, obtained from speaking signals with Boltzmann Machines and Deep Belief Networks.

2.1. Mel-frequency Discrete Wavelet Cofficients (MFDWC)

The study in question has been performed, based on a unique database comprising German voice samples collected from men and women. These voice samples were separated into various feature vectors with MFDWC, and trained. MFDWC is a feature extraction method employed in the speech processing. It is used to extract significant information and features by dividing voice data into subsets. Feature extraction steps of MFDWC technique is shown in the Figure 1 [7].

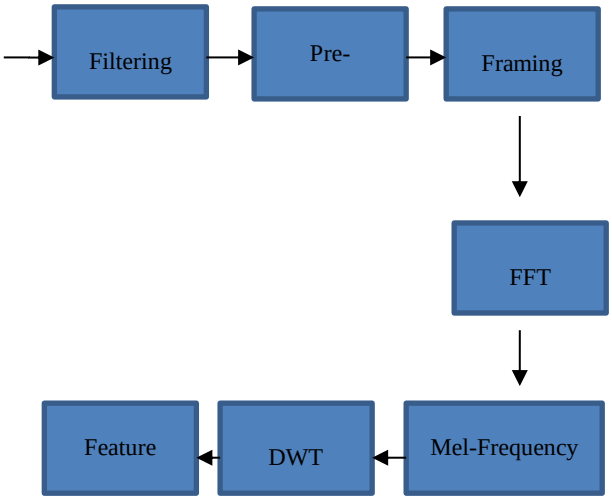

Figure 1. Feature extraction steps of MFDWC

Sample speech signal is shown between the 40-40000 Hz range in the MFDWC feature extraction method. Speech signal is divided into frames after the pre-processing step. Hamming window has been used in this study in order to smoothen the transition of speech samples between the frames. One Mel shows the frequency of voice tone. Mel-scale is scaled between actual frequency of voice signal and estimated voice frequency. For this reason total energy of every frame is calculated.

Classification success in speaker identification has been calculated on an individual basis for MFDWC-3, MFDWC-5 and MFDWC-9 vectors by training the feature vectors obtained from voice signals by means of Boltzmann Machines and Deep Belief Networks.

2.2. Mel Frequency Cepstral Coefficients (MFCC)

MFCC is a feature extraction method, that is used in sound processing. It is used to extract important information and features by dividing the sound data to its subsets. The steps of feature extraction technique of MFCC is indicated in Figure 2[11].

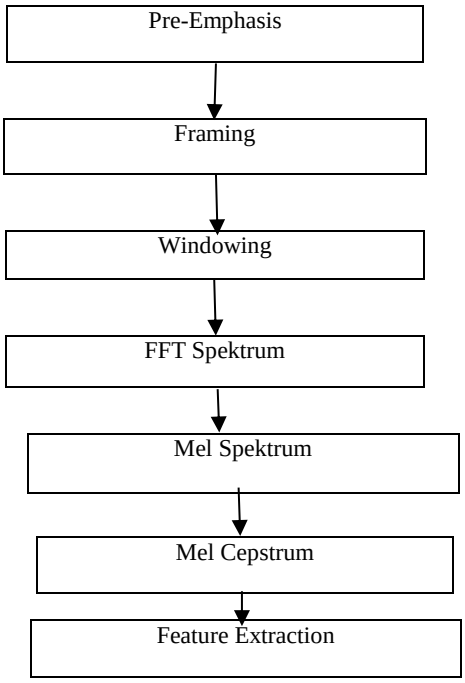

Figure 2. Feature extraction steps of MFCC

Two filters are used in MFCC feature extraction method. The first filter has a linear distribution of frequency values under 1000 Hz and the other has a logarithmic distribution of frequency over 1000 Hz. Pre-emphasis stage is the first stage in obtaining MFFC feature vector. The sound signals, which have high frequency, are passed through a filter at this stage. This way, the energy of the sound is increased at high frequency. The sound signals are analog. The sound signals are converted from analog to digital by getting divided into small frames between 20 and 40 ms during the framing stage and it is divided into N frames.
The sound signal is moved by sliding the sound signal at the windowing stage. This way, the closest frequency lines and the frame, which will come by windowing, that is used are combined The window type, width and sliding amount are determined at this stage. Each of N frames is transmitted from the time space to the frequency space with Fast Fourier Transformer (FFT). The spectral features of sound signals are shown in frequency space.MEL spectrum is obtained by calculating the total weight of these spectral features. This MEL spectrum is formed from triangle waves and are formed by getting passed through a series of filters. MEL spectrum reduces the noise by lowering two neighbour frequencies. The logarithm of signal is taken at the stage of MEL spectrum and the signal is transmitted back again from frequency space to the time space. MEL frequency cepstrum factors are obtained by using DCT (Discrete Cosine Transform) in time space.

2.3. Linear Prediction Cepstral Coefficient (LPCC)

LPCC is a well known and commonly used technique to obtain the characteristic features from sound signals. In this technique, each sample of sound signal is based on the conversion of linear prediction coefficients obtained as a linear weighted total of previous sound signals into cepstral coefficients. This is not a method preferred for sound signals exposed to various environmental effects or noise. LPCC utilizes functions that model the sound path.

LPCC method is obtained by converting LPC coefficients into cepstral coefficients through Fourier conversion. Preliminary process is completed by transmitting speech signal through high filter. Auto correlation characterizes the signal by determining the similarity of each sound signal with itself. This step is materialized in frame of each signal. Signal is analyzed by converting the auto correlation values into LPC parameters by using Levinson-Durbin recursion. In final phase, LPCC parameters are obtained with cepstral analysis [12]. Steps of LPCC feature extraction are provided in Figure 3.

LPCC method is calculated as in Equality 1.1 [13]. a_i LPC coefficients indicate the degree of **p** LPC coefficients.

$$s(n) = \sum_{n=1}^{p} a_i s(n-i) \qquad (1.1)$$

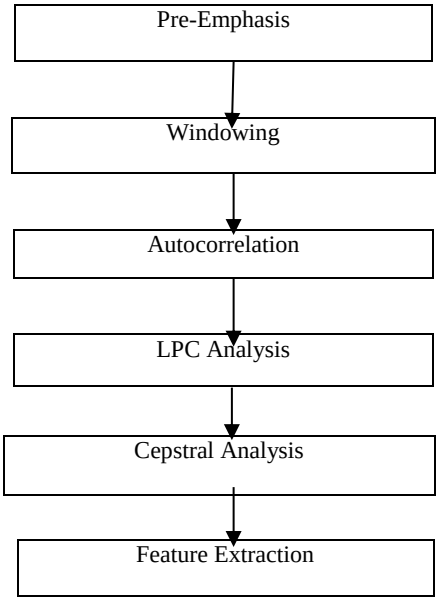

Figure 3. Feature extraction steps of LPCC

3. Classification Methods

Depth learning algorithms are used in certain fields primarily in sound processing in recent years. In particular, depth learning algorithms are developed for the solution of complicated problems. They are comprised of multi layers. Therefore, they may contain numerous hidden information. In this study, two depth learning algorithms are used as Boltzmann machines and Depth belief Networks. Steps of the study are presented in Figure 4.

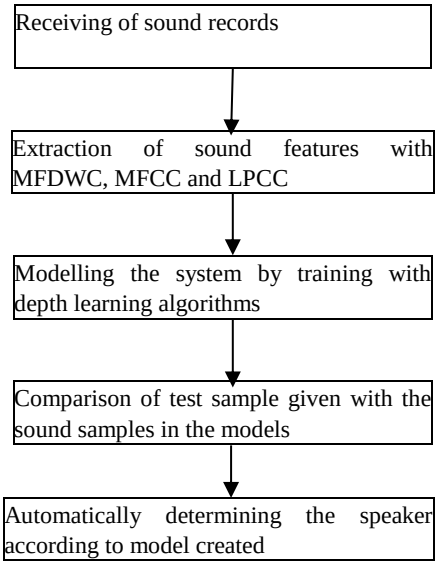

Figure 4. Study steps

3.1. Boltzmann Machines(BM)

Boltzmann Machines(BM) are used in the modelling of units between entry layers and hidden layers. Therefore, parameters learnt in hidden layer are the entry parameters for next BM. So, BM is very significant for deep neural networks.

Boltzmann machines are a stochastic process comprised of recursive artificial neural network. It has multilayer structure. It is used frequently for the solution of searching and learning problems. All artificial neural networks are not linked with each other.

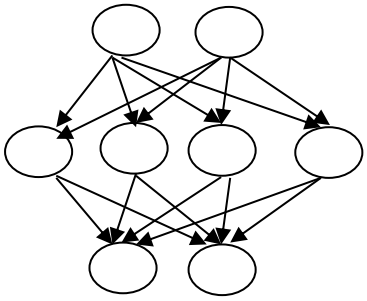

Figure 5. Boltzmann Machines

Restricted Boltzmann Machines (RBM) are the structures limiting the connections to be formed on the layers between units within the entry layer and the units within the hidden layer. All artificial neural networks are not linked with each other. Sample of RBM neural networks are indicated in Figure 5.

RBM is a directional graphical model comprising the hidden and visible units on a layer. All hidden units are indicated as dual graph since they are linked with visible units. Creation of RBM model is indicated in Equality 2.2 [16].

$$p(v,h,\Theta) = \frac{\exp(-E(v,h;\ \Theta))}{Z} \qquad (2.2)$$

In this equality, **v** indicate the visible unit, **h** hidden units, Θ model parameters and E function of model comprised of (v, h, Θ) $Z = \sum_u \sum_h \exp(-E(v, h; \Theta))$ normalization factor.

3.2. Deep Belief Networks (DBM)

Depth Belief Networks (DBM) are the architectures providing modelling of data within hierarchical structure. DBM comprised of subsequent connection of multiple RBM materializes the learning process by subsequently training the RBMs creating its structure. Structure of DBM is indicated in the Figure 6. It is formed of numerous hidden layers. Next possibility of each hidden layer for the data of which entry data is given. Weight of subsequent RBMs are found from hidden activities [14]. This possibility distribution is calculated as in Equality 3.3 [16].

$$p(l = k|h; 0) = \frac{\exp(\sum_{i=1}^{H} \lambda_{ik} + a_k)}{Z(h)} \quad (3.3)$$

Enrty signal in Equality 3.3 is processed through all layers and converted into output multiple distribution. Entry signal is divided into **k** and e λ_{ik} in final layer weights between **k** class and hidden units are indicated.

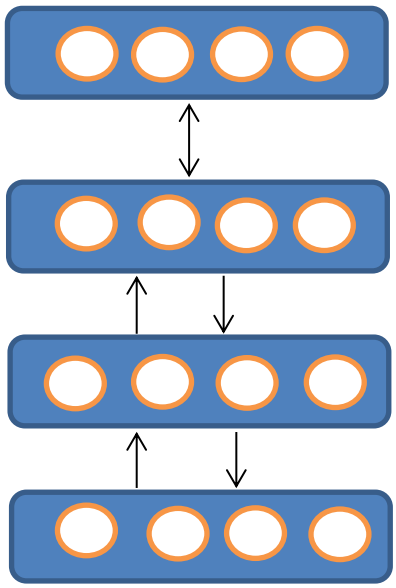

Figure 6. Deep Belief Networks Structure

It is recommended to materialize the learning of DBM network in two phases. In preliminary training phase as first stage, weight parameters modelling the entry data are learnt. Second phase is referred as finetuning and aims for the learning of network parameters classifying the training set [15].

4. Experimental Study and Results

In this study, an authentic and unique German database has been used. The names, surnames, ages, sexes and speeches of persons have been added into this data base. The different number of feature vectors of sound components have been extracted with MFDWC, MFCC and LPCC feature extraction method. In the next stages, the sound samples have been trained by using methods Boltzmann Machines and Deep Belief Networks. The features of the recorded sound samples have been indicated in Table 1.

Table 1. Attributes of used databases

Age Range	Number of speaker	
	Male	Female
18-25 range speakers	15	19
26-40 range speakers	23	16
41 and more speakers	12	15

Feature vectors of voice components with different quantities; have been extracted by means of MFDWC, MFCC and LPCC feature extraction method. Voice samples have been tested by training them, using available feature vectors by means of Boltzmann Machines and Deep Belief Networks methods.
Success rates of speech samples obtained utilizing MFDWC different feature vectors are given for Boltzmann Machines and Deep Belief Networks in the Table 2. Success in the speaker identification increases as the number of words used increases in all techniques employed. Deep Belief Network gave more successful results when compared to Boltzmann Machines.
Success rates of speech samples obtained utilizing MFCC different feature vectors are given for Boltzmann Machines and Deep Belief Networks in the Table 3. Success in the speaker identification increases as the number of words used increases in all techniques employed. Deep Belief Network gave more successful results when compared to Boltzmann Machines.
Success rates of speech samples obtained utilizing LPCC different feature vectors are given for Boltzmann Machines and Deep Belief Networks in the Table 4. Success in the speaker identification increases as the number of words used increases in all techniques employed. Deep Belief Network gave more successful results when compared to Boltzmann Machines.
Success rates of speech samples obtained employing MFDWC, MFCC and LPCC feature vector. Success rates of speech samples obtained employing 9 feature vector, for all feature extraction techniques.
A unique and genuine German language database has been employed in this study. Names, family names, ages, speeches and genders of the persons were added to this database. Feature vectors of voice components with different quantities; have been extracted by means of MFDWC, MFCC and LPCC feature extraction method. Voice samples have been tested by training them, using available feature vectors by means of Deep Belief Networks and Boltzmann Machines. In the testing phase it was determined successfull classification techniques as male or female by available testing example. It has also presented and compared by calculating the success of any method used.

Table 2. Success in classification for MFDWC

Feature vectors/ Methods used	3		5		9	
	Male	Female	Male	Female	Male	Female
Boltzmann Machines	87.62	82.15	74.62	70.82	79.62	76.21
Deep Belief Networks	85.03	83.72	56.75	84.62	87.56	81.62

Table 3. Success in classification for MFCC

Feature vectors/ Methods used	3		5		9	
	Male	**Female**	**Male**	**Female**	**Male**	**Female**
Boltzmann Machines	90.87	91.72	94.67	93.21	97.85	96.94
Deep Belief Networks	94.67	90.86	95.65	95.39	99.91	98.78

Table 4. Success in classification for LPCC

Feature vectors/ Methods used	3		5		9	
	Male	**Female**	**Male**	**Female**	**Male**	**Female**
Boltzmann Machines	93.27	94.84	95.61	94.38	98.83	97.62
Deep Belief Networks	96.71	93.62	97.65	96.05	99.62	99.17

5. Conclusion

Voice recognition plays an important role in our day due to security and many other reasons. Person and speaker identification systems have been developed, being based on an unique database obtained by utilizing German language in this study. Classification success of the methods employed in the study have been calculated separately for men and women, and results are demonstrated in a comparative manner. Deep Belief Networks provided more successful results compared to the Boltzmann Machines when the results are taken into consideration. Speaker recognition system is more successful for men compared to the women. 9 feature extraction is more successful compared to the results obtained utilizing for all other feature extraction techniques.

References

[1] Douglas, Reynold , Walter, Andrews and Joseph, Campbell etc., "The SuperSID Project: Exploiting High-Level Information for High-Accuracy Speaker Recognition", In.Proc. ICASSP, Hong Kong, p.784-787, 2003.

[2] Douglas, Reynolds , Thomas, Quatieri and Robert, Dunn, "Speaker Vrification using Adapted Gaussian Mixture Models", Digital Signal Processing 10, p.19-41, 2000.

[3] Edmondo, Trentin and Marko, Gori, "A survey of hybrid ANN/HMM models for automatic speech recognition", Elsevier Neurocomputing 37, p.91-126, 2001.

[4] Keiichi, Tokuda , Heiga, Zen and Alan, Black, "An HMM-Based Speech Synthesis System Applied to English", Proc.of 2002 IEEE SSW, p.227-230, 2012.

[5] Lihang, Li, Dongqing, Chen and Sarang, Lakare etc, "Image segmentation approach to extract colon lümen through colonic material taggng and hidden markov random field model for virtual colonoskopy", Medical Imaging, 2002.

[6] Lindasalwa, Muda and Mumtaj, Began, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", Journal Computing, vol.2, issue 3,p.138-143, ISBN 2151-9617, 2010.

[7] M., Fahid M. and M.A, Robust Voice conversion systems using MFDWC", 2008 International Symposium on Telecommunications, p.778-781, 2008.

[8] Quan, Jie-Fu, Fan Gang, Zeng F and Robert, Shannon etc., "Importance of tonal envelope cues in Chinese speech recognition", The Journal of the Acoustical Societct of America, vol.104, no.1, p.505-510, 1998.

[9] Seok, Oh and Ching, Suen, "A class-modular feed forward neural network for handwriting recognition", Pattern Recognition, vol.35, issue 1, p.229-244, 2002.

[10] Wouter, Gevaert , Georgi, Tsenov and Valeri, Mladenov, "Neural networks used for speech recognition", Journal of Automatic Control,vol.20,p.1-7,2010.

[11] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, " Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", Jornal of Computing, vol.2, issue 3, p.138-143, ISSN 2151-9617, 2010.

[12] M.Zbancioc and M.Costin, "Using Neural Networks and LPCC to Improve Speech Recognition", International Symposium on Signals, Circuits and Systems, vol.2, pp.445-448, 2003.

[13] O.Eray, "Destek Vektmr Makineleri ile Ses Tanıma Uygulaması", Pamukkale Üniversitesi, 2008.

[14] A.Mohamed, T.Sainath and G.Dahl, "Deep Belief Networks Using Disciminative Features for Phone Recognition, IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011, 2011.

[15] ,G. E. Hinton., ,S. Osindero, and ,Y. Teh, " AFast Learning Algorithm For Deep Belief Nets", Neural Computation,vol. 18, 2006.

[16] A.Mohamed and D.Deng. "Investigation of full-sequence training of Deep Belief Networks", Interspeech2010, pp.2846-2849, 2010.