Automatic Speaker Gender Identification for the German Language

C. Bakir

Abstract— **Authentication systems necessitate transmission, design and classification of biometric data in a secure manner. Moreover, in voice process of biometric can be obtained successful results by determining gender of speaker. In this study, the aim was to designed system taking German sound forms and properties for automatic recognition gender of speaker. Approximately 2658 German voice samples of words and clauses with differing lengths have been collected from 50 males and 50 females. This voice samples includes more than one word as a word. Features of these voice samples have been obtained using MFCC (Mel Frequency Cepstral Coefficients). Feature vectors of the voice samples obtained have been trained with such methods as Hidden Markov Model, Dynamic Time Warping and Artifical Neural Network. In the test phase, gender 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.**

Index Terms—***Speaker Gender Recognition System; Hidden Markov Model; Dynamic Time Warping; Artificial Neural Networks.***

I. Introduction

NOWADAYS, the security problems have started to pop out along with the development of the technology. The security issue comes at the helm of these problems. Especially, the biometric systems, such as identity authentications, form the most significant part of the security issue. Thus, forensic sound examinations of the audio records, which are the subject of various crimes, are required. Some studies have been made in order to prevent the leakage of information, which belong to the persons, particularly in commercial transaction to other persons. Hand Writing Recognition, Signature Recognition, Face recognition, Voice recognition and Iris Recognition form some of these studies [6].

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 made to recognize the sound and speaker. But, only in a small part of these studies, the gender recognition study, which is based on sound signal, was able to be made. The voice recognition systems are reviewed as independent from the speaker and dependent in two parts. In case the sound record, which is used in training and test stage, is the same of the speaker, it means dependent to the speaker, if not it means independent from the speaker.

Perry and his colleagues have tried to determine the gender of 10 male 10 female speakers, whose ages varied between 4-16 by looking at acoustic sound features. They have observed the formant frequency of two vowels, which don’t blend and basic frequency with the age group differences of the speakers. The formant frequency has reflected the characteristic features of the sound and relations between the age differences and gender differences. But, this study has been realized on an English data base [7].

Parris and Carey have realized a study, independent from the text in order to determine the gender of the speaker. They have tried to develop a new method by combining acoustic sound features and tone frequency of the speaker. They have normalized the data with Linear Discriminant Analysis method –LDA- on OGI data base and trained it with Saklı Markov Model [11].

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 [1]. 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.

C. BAKIR, is with Department of Computer Engineering University of Yildiz Technical University, Istanbul, Turkey, (e-mail: cigdem@ce.yildiz.edu.tr).

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 [2]. 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 [3]. 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.

Reynolds et al. have attempted to substantiate speaker identification and verification using Gaussian Mixture Model (GMM) method [8]. Attempts have been made to determine speaker verification in the system according to the probability distribution. 11 different hypotheses have been developed for this probability distribution. Data used in the study has been extracted from telephone conversations.

Speech has an important place 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 the 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 voice recognition process difficult [9].

Feature extraction and classification techniques used, were given in the section 2 of the study performed, experimental study in the section 3 and conclusion were given in the section 4.

II. METHOD

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 MFCC. In the second stage, the feature vectors of the recorded sound signals are trained with classification algorithms, such as DTW, HMM and ANN. 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 MFCC-1, MFCC-3 and MFCC-5 and MFCC-9 and the success of the methods have been presented comparatively by training the feature vectors, obtained from speaking signals with DTW, HMM and ANN.

A. *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 1[10].

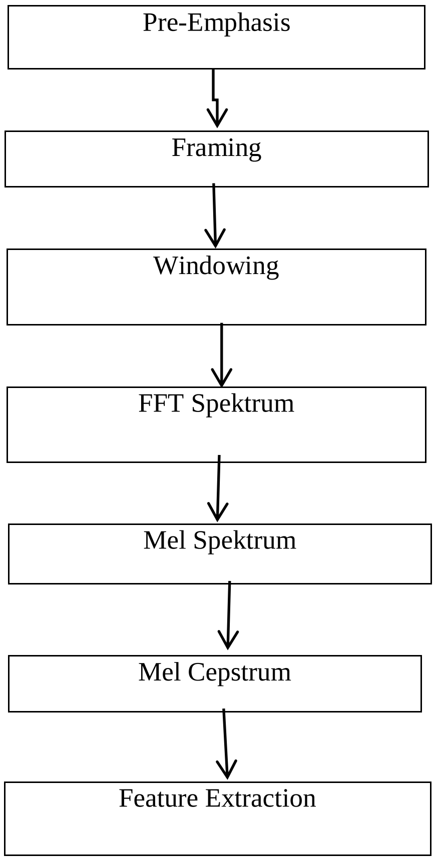

Fig.1 -MFFC feature extraction steps

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.

B. Artifical Neural Network (ANN)

ANNs have a very wide fields of application up to automotive, banking, defense industry, electronics, entertainment, finance, insurance, manufacture, oil and gas, robotics, telecommunication and transportation industry.

Artificial neural networks are information systems which mirror human brain function, and classify the data through learning. They have been developed, being based on a principle of human brain functioning. In other words; ANNs have been developed with a logic similar to the biological neural networks, and are data processing structures connected to each other with weights.

ANNs comprise of input layer, output layer and hidden layers. Data is received into neural networks through input layer. And it is transferred to outside through output layer. Layers between input and output layers constitute hidden layers.

Neurons in the feed-forward neural networks are connected just in the forward direction [6]. Each layer of neural network contains the connection of next layer and these connections are not in the backward direction. In a sense, there is a hierarchical structure between neurons, and the neurons located in one layer can only communicate data to the next layer. Structure of a feed-forward ANN is shown in the Figure 2.

Backward propagation network shows how to train a neuron [7]. Trainer is a sort of learning. Network is maintained both with the sample inputs and expected outputs when the trainer method is employed. Expected outputs are compared with actual outputs for the networks the inputs of which are given. Error is calculated in case the expected outputs are used, and weights of various layers are adjusted in the backward direction from output layer to input layer. In other words, it is given for both input data and output data. Network updates its coefficients in order to obtain the expected output.

ANN is the most widely used method. In this algorithm, error in the output layer is calculated at the end of each iteration, so this error is transmitted to all neurons in the direction from output layer to input layer, and weights are readjusted according to the error margin. Such error margin is distributed to the previous neurons located before the said neuron in proportion to their weights.

Layers are located one after another in a multilayer artificial neural network. Outputs of neurons in a layer will be given as their weights, to the input of next layers, and these weight are used in the calculation of outputs for the next layer. Weights of the hidden layer between input and output layers are calculated [7].

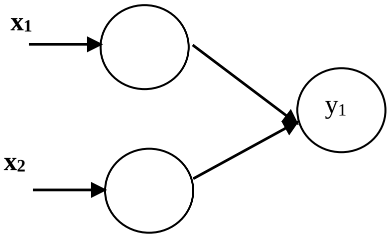

Fig. 2. Feed-forward neural network

C. Hidden Markov Model (HMM)

A lot of studies have been carried out with regard to the Hidden Markov Models (HMM) in many fields from past to today. HMM has been used in a wide manner in face recognition, speech recognition, voice recognition, hand script recognition, human body motion recognition, bioinformatics, estimation of gene, cryptanalysis, protein structure and sequence, DNA sequence and pattern recognition.

In Hidden Markov Model (HMM) the aim is to try to estimate future situations that will likely occur in cases when the existing situations are given as an input to the system. HMM is a stochastic process since it generates different output whenever it is operated. In addition, system in Markov models may move from its own state to another state according to the probability distribution, or remain in the same state. Probabilities occurred in the states are called as transition probabilities. States are not seen by the observer as distinct from HMM normal Markov model. However, transition subject to the states may be observed. HMM speaker recognitions systems comprise of the following steps [5].

S= { S_1, S_2,..., S_Q } shows current status of the speech signals generated where there are Q numbers of states.

- Initial state probabilities is determined in a discrete time, t. (π= { P_r (S_i |t=0, S_i ε S})
- Transition probabilities are calculated according to the current states. a_{ij} = (P_r(S_j t in time t| S_i in time t-1), S_i ε S, S_j ε S))
- F, which is the number of features observed, is determined.
- Probability distribution of speech signal will be calculated in this way. (b_x= {b_i (x)= P_r (x(S_i), S_i ε S, x ε F})
- HMM generated is demonstrated by λ =(a,b,π).

D. Dynamic Time Warping (DTW)

Dynamic Time Warping (DTW) finds out to which speaker the voice signal given belongs, by calculating the similarity between the time-variant two speech signals. The most optimal time curve can be identified between two signals with this method.

$$Q=q_1,q_2,\ldots,q_i,\ldots,q_n \quad (1)$$

$$C= c_1, c_2,\ldots, c_j,\ldots, c_m \quad (2)$$

Q and C in the equation 1 and equation 2 demonstrate two distinct speech signals; n and m show the lengths of these speech signals [4]. In this case, the ratio of similarity between Q and C signals is calculated using Euclid length as in the equation 3 [4].

$$d(q_i,c_j)= (q_i,c_j)^2 \quad (3)$$

A matrix (i,j) is generated for Q and C. Accumulated distance matrix is calculated using this matrix.

$$D(i,j)=\min[D(i-1,j-1),D(i-1,j),D(i,j-1)]+d(i,j) \quad (4)$$

III. EXPERIMENTAL STUDY

In this study, an authentic and unique German data base 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 MFCC feature extraction method. In the next stages, the sound samples have been trained by using methods of ANN, HMM and DTW. The features of the recorded sound samples have been indicated in Table 1.

In Table 2, the success ratios of the speech samples, obtained by taking MFCC-9 Feature Vector, for ANN, HMM and DTW have been given. As much as the number of the words, used in speeches, are increased, all techniques, that have been used for success of recognizing the gender of speaker has also increased. Saklı Markov Model has given more successful results compared to other techniques. 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 MFCC feature extraction method. In the next stages, the sound samples have been trained by using methods of ANN, HMM and DTW. During the test stage, it has been tried to be defined whether the test sample is a male or a female with the used methods. Furthermore, the success of all methods have been calculated and presented comparatively.

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 MFCC feature extraction method. Voice samples have been tested by training them, using available feature vectors by means of ANN, HMM and DTW methods. In the testing phase it was determined male or female by available testing example. It has also presented and compared by calculating the success of any method used.

In Table 2, the success ratios of the speech samples, obtained by taking MFCC-1, MFCC-3, MFCC-5 and MFCC-9 feature vectors for ANN, HMM and DTW have been given. As much as the number of the words, used in speeches, are increased, all techniques, that have been used for success of recognizing the gender of speaker has also increased. HMM has given more successful results compared to other techniques. The tone frequency is between 120-200 Hz in women and 60-120 Hz in men. The determining the sex in women from tone frequency is more successful compared to the men.

TABLE I
ATTRIBUTES OF USED DATABASES

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

TABLE II
SUCCESS OF THE GENDER SPEAKER USED METHODS

Used Feature Vectors	ANN (%)		HMM (%)		DTW (%)	
	Female	Male	Female	Male	Female	Male
MFCC-1	72.03	68.44	71.63	69.63	69.63	67.35
MFCC-3	74.02	71.87	76.35	75.33	76.01	71.93
MFCC-5	81.63	79.37	84.99	82.37	79.22	75.54
MFCC-9	85.69	80.36	98.34	97.02	87.37	86.33

IV. CONCLUSION

The sound recognition has a great importance from the angle of security and many other reasons. In this study, a system, aimed at determining the gender of the speaker, has been developed on the unique database, obtained by using German language. Classification success of the methods, used in the study, has been calculated separately in men and women and the results have been presented comparatively. When we looked at the results; it has been seen that HMM method has given more successful results compared to other classification methods. Furthermore, the speaker gender recognition system is more successful in women compared to the men. MFCC-9 feature extraction has rather more success as per the results, obtained by using 1,3 and 5 feature vectors.

REFERENCES

[1] 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, pp.505-510, 1998.

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

[3] 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, pp.784-787, 2003.

[4] Lindasalwa, Muda and Mumtaj, Began, “Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques”, Journal Computing, Vol.2, No.3, pp.138-143, ISBN 2151-9617, 2010.

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

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

[7] Theodore L. Perry, Ralph N. Ohde,a) and Daniel H. Ashmead, ” The acoustic bases for gender identification from children’s voices”, J. Acoust. Soc. Am. 109 (6), pp.2988-2998, 2001.

[8] Douglas, Reynolds, Thomas, Quatieri and Robert, Dunn, “Speaker Verification using Adapted Gaussian Mixture Models”, Digital Signal Processing 10, pp.19-41, 2000.

[9] Wouter, Gevaert, Georgi, Tsenov and Valeri, Mladenov, “Neural networks used for speech recognition”, Journal of Automatic Control, Vol.20, pp.1-7, 2010.

[10] 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, No.3, pp.138-143, ISSN 2151-9617, 2010.

[11] Eluned, Parris, Micheal, Carey, “Language Independent Gender Identification”, Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on, Vol.2, pp.685-688, 1996.

[12] 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.

[13] Seok, Oh and Ching, Suen, “A class-modular feed forward neural network for handwriting recognition”, Pattern Recognition, Vol.35, No.1, pp.229-244, 2002.

BIOGRAPHIES

CIGDEM BAKIR was born in İstanbul. She received the B.S. degrees in computer engineering from the University of Sakarya, in 2010 and the M.S. degree in computer engineering from Yildiz Technical University, İstanbul, in 2014.

Since 2012, she was a Research Assistant with the Yildiz Technical University. Her research interests include recommendation systems, data mining, image processing and biomedical signal processing.