International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VII, July 2019- Available at www.ijraset.com

Speaker Recognition and Gender Identification using Artificial Neural Network and Support Vector Machine

Ayush Krishna[1], Dr. Neelu Jain[2]
[1]*Mtech. Student,* [2]*Professor, Electronics and Communication Engineering Department, Punjab Engineering College (Deemed to be University), Chandigarh*

Abstract: ***Identity of a person via voice is one of the most interesting techniques used for user identification. Accuracy of identification process depends on the number of feature vectors and the number of speakers. This paper aims to develop a system able to identify a person and gender from his speech. Recognition relies on English words as a recognition phrase. For speaker recognition speech features used are Mel Frequency Cepstral Coefficients (MFCCs), Energy, Pitch and Zero Crossing Rate while for gender recognition only MFCCs, Energy and pitch are used. Artificial Neural Network (ANN) and Support Vector Machine (SVM) are used as classifiers. Experimental results demonstrated that the proposed system returns 97% accuracy rate for gender identification and 95% accuracy rate for speaker recognition.***
Index Terms: Speaker Recognition, Gender Recognition, MFCC, Pitch, Energy, Artificial Neural Network, Support Vector Machine

I. INTRODUCTION

Speaker identification is one of the most complicated tasks in speech recognition problems. It generally depends on the production of speech of the speaker with physiological and behavioral characteristics. These characteristics rely on the speech generation (voice source) and the envelope behavior (vocal and nasal tract) [1]. All speaker recognition systems contain two main phases: (*i*) feature extraction, and (*ii*) recognition. During the first step, a training vector is generated from the speech signal of the phrase spoken by the user. These training vectors are stored in a database for subsequent use in the recognition phase. During the recognition phase, the system tries to identify the unknown speaker by comparing the extracted features from phrase with the ones from a set of known speakers.

The success of a speaker recognition system depends on extracting the right speaker dependent features which should be invariant in the articulation dynamics. The selection of a feature should be such that it must have large variation between speakers and small variations between different sessions of the same speaker, it must be robust against noise and channel effects and hard to mimic or reproduce [2]. The earliest speaker recognition technique [3] proposed the use of long term averages of acoustic features such as spectrum representation and pitch [4]. However these methods require long speech utterances and therefore much speaker dependent information was lost.

The linear predictive coefficient (LPC) parameters modeled on the vocal tract characteristics represents the vocal tract resonance of the acoustic spectrum [5] [6], however these parameters are severely affected by noise. The LPC Residual signal representing the speakers glottal information [7], helps improve the speaker recognition rate to some extent. The use of filter banks for calculating features provides more robustness and accuracy to the system. MFCC modeled on the filter bank technique is the most popular feature for speaker recognition because of its advantages over the other features and seems difficult to beat in practice [2]. Once feature set is created, a speaker model is trained and stored into the system database [2]. The use of Neural Networks (NN) for classification is considered for many applications. NN's model the decision function which best which best discriminates a speaker from a known set instead of training individual models to represent a particular speaker [8]. This approach provides each speaker with his/her personalized NN which is activated only by their utterance.

In this paper, we focus on developing robust speaker recognition and gender identification system using feature fusion technique on MFCC, pitch and energy features on two different classifiers Artificial Neural Network and Support Vector Machines.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VII, July 2019- Available at www.ijraset.com

II. PROPOSED AUTOMATIC SPEAKER RECOGNITION AND GENDER IDENTIFICATION SYSTEM

Both the speaker recognition and gender identification system consists of three main parts which are speaker database collection, feature extraction and feature matching[9,10]. The steps are shown in figure 1.

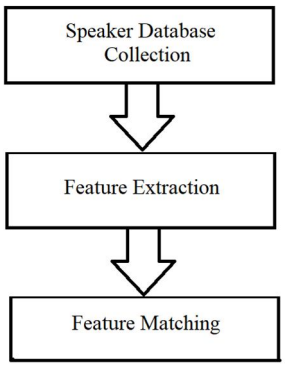

Fig 1.Steps in speaker recognition and gender identification

A. Database Collection

This is the first step of speaker recognition system development. To collect the database for biometric speaker recognition Session Variability need to be taken into consideration. Session variability also known as Inter-session variability refers to all the phenomena causing variations in two recordings of same speaker. In other words, two speech samples recorded by the same person are mismatched and could not be recognized by the system. There are several factors responsible for inter-session variability such as Transmission channel, Transducer characteristics, Environmental noise etc.

B. Feature Extraction

While dealing with speech signals, generally, it is preferable dividing signals into frames to obtain stationarity. Speech signal is usually do not have stationarity over a long time-span, but for a small duration it is stationary. This is due to reason that glottal system cannot vary instantaneously. Speech is stationary approximately in window of 25 ms. The following features are extracted from each speech sample.

1) *Mel-frequency cepstral coefficients (MFCC):* MFCC coefficients are extracted for all the available speech samples of all speakers according to the procedure given in figure 2.

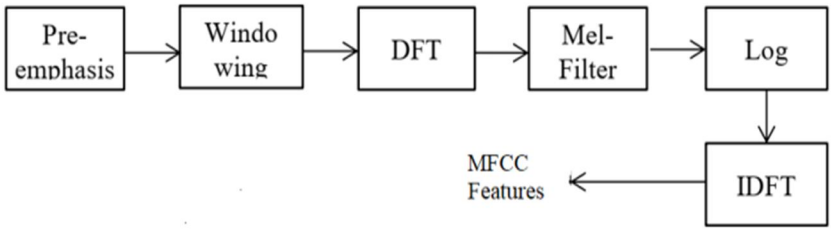

Fig2: MFCC Feature Extraction Workflow

Pre-emphasis is a high-pass filtering process for amplifying the energy at high frequencies. In the windowing process the samples of the signal are multiplied with a window function. This is done for minimizing any signal discontinuities, effectively slicing the signal into discrete segments. A popular choice is the Hamming window because it prevents any sharp edges like rectangular windows. DFT then converts a sequence interchangeably in time and frequency domain. It is given by equation 1.

$$X_K = \sum_{n=0}^{N-1} x_n . e^{-j\frac{2\Pi kn}{N}} \qquad (1)$$

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VII, July 2019- Available at www.ijraset.com

The DFT calculations only pertain to a linear frequency scale, therefore, we have to apply a process called frequency warping, in which the spectrum frequencies have to be converted to smaller numbers using the logarithmic Mel scale. In order to achieve this, a filter bank can be built as presented in Fig3, and effectively map the DFT frequency bin centers. This filtering is also known as the Mel-Spectrum defined as equation 2 and is shown in figure 3,

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (2)$$

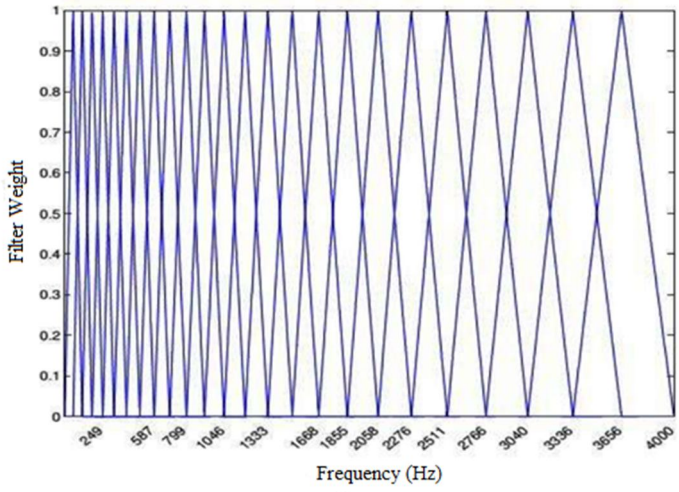

Fig3. Triangular Mel-Filter Banks for a 24-Filter System

Inverse Discrete Fourier Transform of Mel-Spectrum is then computed, yielding the MFCCs. First 12 values of the cepstrum contain the meaningful information to provide unique characteristics of the waveform. These are used as MFCC features.

2) Pitch: The voiced signals are generated by the vibration of vocal folds and sub glottal air pressure. The time duration of one glottal cycle is called as the pitch period and pitch is the fundamental frequency which is the reciprocal of the pitch period. Pitch estimation can be performed following method:

a) The digitized speech signal is hamming windowed to convert it into a suitable frame size.

b) The windowed signal is transformed into frequency domain by using Fast Fourier Transformation algorithm.

c) Logarithm is computed on the absolute values of the signal in the frequency domain.

d) Inverse Fast Fourier Transform is applied on the signal computed to convert it into Cepstral domain and the first signal peak is considered as the pitch frequency.

3) Zero Crossing Rate: In case of speech signals, a zero crossing is the moment when the successive samples have different algebraic signs. Now the ZCR is the number of times in a given time interval/frame that the zero crossing occurs in a speech signal. If the number of zero crossings is more in a given signal, then the signal is changing rapidly and accordingly the signal may contain high frequency information and if the number of zero crossing is less, then the signal is changing slowly which means the signal may contain low frequency information.

The zero crossing can be calculated using the equation 3.

$$Z_n = \sum_{m=-\infty}^{\infty} |sgn[x(m)] - sgn[x(m-1)]| w(n-m) \quad (3)$$

where

$$sgn[x(m)] = 1 \quad x(m) \geq 0$$
$$= -1 \quad x(m) < 0 \quad (4)$$

And

$$w(n) = \frac{1}{2N}, \quad 0 \le n \le N-1$$
$$= 0 \quad Otherwise \qquad (5)$$

4) *Energy:* The short time energy of a speech signal is defined as shown in equation 6.

$$E_n = \sum_{m=n-N+1}^{n} x^2(m) \qquad (6)$$

Which means the short-time energy at sample n is simply the sum of squares of the N samples n-N+1 through n. Short time energy of a speech signal is calculated from each frames of the signal.

5) *Kurtosis:* Kurtosis is the mostly used feature, which is the relation between the middle fourth order moment and the square of the middle second order moment and is given by the equation 7.

$$k = \frac{E((x-\mu_1)^4)}{E((x-\mu_1)^2)^2} \qquad (7)$$

C. Feature Matching

The state-of-the-art feature matching techniques used in speaker recognition and gender identification includes Artificial Neural Network (ANN) and Support Vector Machine (SVM).

1) *Artificial Neural Network (ANN):* The most elementary unit of a neural system is a *neuron* in both biological and artificial networks. Synapses correspond to the connections between neurons and are responsible for transmitting information (stimulus). As a neuron can be connected to many other neurons, several stimuli can accumulate in a neuron. For an artificial neural network (ANN), the stimuli can be assumed as the incoming signal xi and the synapses as the connections wi. In practice, wi are represented as weights that scale the incoming inputs according to their importance. These weighted inputs accumulate inside the neuron and some function of the sum is given as an output, y. This function, $\phi(.)$ is called the *activation function.* In general, there is also a bias (threshold) term Γ for each neuron. In mathematical terms, the output is given by Equation 8.

An example schematic of a simple NN structure can be seen in figure 4.

$$y = \phi\left(\sum_{i=1}^{n} x_i w_i - \Gamma\right) \qquad (8)$$

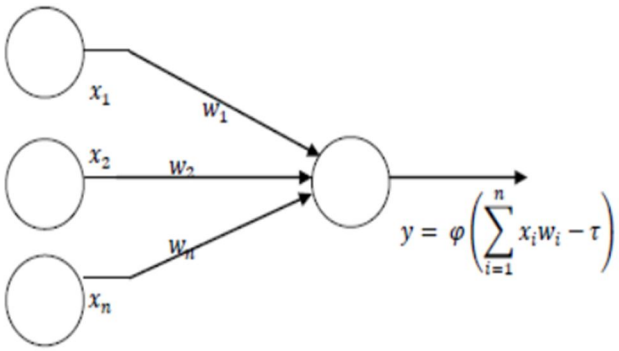

Fig4: A Simple Neural Network Structure

A regular neural network consists of several layers each containing several units called neurons. The first and the last layers are called the input layer and the output layer respectively. The layers in between these two are called the hidden layers. The total number of layers and the number of units in each layer affects the expression power of NN. A NN is said to be fully connected if each neuron in a layer is connected to every other neuron in the next layer. Figure 5 corresponds to this type of networks as there are no missing connections between neurons. Otherwise, the NN is said to be partially connected.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VII, July 2019- Available at www.ijraset.com

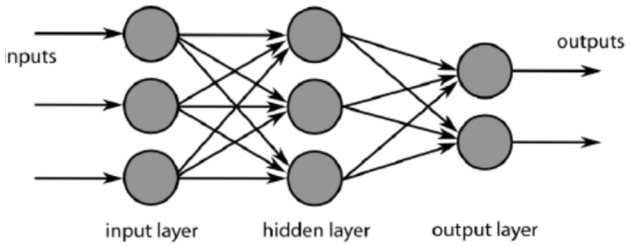

Figure 5: Neural Network structure

2) *Support Vector Machine:* In machine learning, support-vector machines (SVM) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

III. RESULTS

The speaker recognition and gender identification system has been implemented in Matlab2018a on windows10 platform.

A. Gender Identification System

There are two databases used for testing the gender identification system. First an own database is created, the speakers selected for creating database belong to different regions of India i.e. they have different native language. Speakers are made to speak a single phase “Please enter now”. The other database is TIMIT dataset. The identification rate is defined as the ratio of the number of speakers identified to the total number of speakers tested. Table 1 shows the average identification rate for gender identification system using TIMIT dataset.

Table 1 Gender Identification System Average Accuracy

Number of speech samples			Average Identification Accuracy (%)	
Male	Female	Total	SVM	ANN
100	100	200	90	85.4
150	150	300	93.56	92.73
200	200	400	95.42	94.2
250	250	500	96.08	95

Table 1 shows that increasing the number of speech samples increases the recognition accuracy but it also has a disadvantage that it leads to increase in the time taken to train the classifier. So there is a tradeoff between accuracy and speed. Table 2 shows comparative analysis of the accuracy rate using own dataset and TIMIT dataset.

Table 2 Gender Identification Accuracy using TIMIT and own dataset

Number of speech samples	Accuracy with TIMIT dataset		Accuracy with own dataset	
	ANN	SVM	ANN	SVM
10	70	60	60	50
15	73.33	62	70.33	53.33
20	80	76	75	60

Table 2 indicates that the accuracy of the system is more for the TIMIT dataset. This can be because of more noise present in the samples in own dataset. The TIMIT dataset is created in a recording room with high quality equipment which prevents any noise from being recorded. On the other hand own dataset can definitely have some noise or jitter which may be a cause for lower accurate results.

B. Speaker Recognition System

The speaker recognition system recognizes the speaker from their speech samples. The database for this system contains a single phrase "Please enter now" spoken by 20 speakers spoken some number of times. For determining the number of speech samples of each speaker an experiment is conducted with varying number of speech samples which is tabulated in table 3.

Table 3 Variation in the accuracy with number of speech samples

Number of speech samples of each speaker	Recognition Accuracy (%)
3	78
5	84
7	89
10	94

Table 3 shows that increasing the number of speech samples increases the recognition accuracy of the system. So using 10 speech samples as standard value in all the experiments.

The speaker recognition systems employs ANN with 3 hidden layers and 5 neurons and a SVM classifier. The identification rate is defined as the ratio of the number of speakers identified to the total number of speakers tested. An average identification accuracy for speaker recognition is given in table 4.

Table 4 Speaker Recognition System Average Accuracy

Number of speakers	Average Accuracy (%)	
	ANN	SVM
10	96.25	97
15	96	94.5
20	95.16	94.3

Table 4 indicates that the accuracy of the system decreases as the number of speakers increases. This may be compensated by increasing the number of speech samples of each speaker. This will eventually lead to increase in training time. So there is always a tradeoff between number of samples and training time.

REFERENCES

[1] D A Reynolds, "An Overview of Automatic Speaker Recognition Technology", IEEE International Conference on Acoustics, Speech And Signal Processing (ICASSP), Volume 4, May 2002, pp. IV-4072—IV-4075.
[2] Tomi Kinnunen, Haizhou Li, "An Overview of Text Independent Speaker Recognition: From Features to Supervectors", Speech Communication, Elsevier, Volume 52, 2010, pp. 12-40.
[3] S. Furui, F. Itakura, and S. Saito, "Talker recognition by longtime averaged speech spectrum", Electron., Commun. in Japan, Vol. 55-A, No. 10, pp. 54-61, 1972.
[4] J. Markel, B. Oshika, and A. Gray, Jr., "Longterm feature averaging for speaker recognition", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-25, pp. 330-337, Aug. 1977.
[5] Xugang Lu, Jianwu Dang, "An Investigation of dependencies between Frequency Components and Speaker Characteristics for Text Independent Speaker Identification", Speech Communication, Elsevier, Volume 50, 2008, pp. 312-322.
[6] Ankita N. Chadha, Jagannath H. Nirmal, Pramod Kachare, "A Comparative Performance of Various Speech Analysis- Synthesis Techniques", International Journal of Signal Processing Systems, vol. 2,No. 1, June 2014, pp.17-22.
[7] He, J., Liu, L, "On the use of features from prediction residual signals in speaker identification", In: Proc. EUROSPEECH95, Madrid, Spain,Vol. 1, pp. 313–316, 1995.
[8] A.K. Jain, J. Mao and K.M. Mohiuddin, "Artificial Neural Networks: A Tutorial", IEEE Computer, pp. 31-44, Mar. 1996.
[9] Zhong-Xuan, Yuan & Bo-Ling, Xu & Chong-Zhi, Yu. (1999). "Binary Quantization of Feature Vectors for Robust Text-Independent Speaker Identification" in IEEE Transactions on Speech and Audio Processing, Vol. 7, No. 1, January 1999. IEEE, New York, NY, U.S.A.
[10] F. Soong, E. Rosenberg, B. Juang, and L. Rabiner, "A Vector Quantization Approach to Speaker Recognition", AT&T Technical Journal, vol. 66, March/April 1987, pp. 14-26