ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Issue 4, April 2016

Classification of Gender Based On Voice Using Support Vector Machine

Renisha[1],Gomathi Priya[2], Karpagavalli[3], Kaveri Priya [4],Krishnaveni[5]
Assistant Professor, ECE, GCE, Tirunelveli,India[1]
Student[1], ECE, GCE, Tirunelveli,India[2]
Student[2], ECE, GCE, Tirunelveli,India[3]
Student[3], ECE, GCE, Tirunelveli,India[4]
Student[4],ECE, GCE, Tirunelveli,India[5]

Abstract*:* Biometric authentication is an emerging technology that utilize biometric data for the purpose of person identification or recognition in security applications. The speech recognition problem can be modeled as a classification problem, where one wants to get the best degree of separability between classes representing the voice. Gender identification refers to the classical problem of distinguishing active male speech from female speech.Many approaches have been devised to improve the performance of gender identification with relevant strategies. An efficient method comparing Support Vector Machine and Artificial Neural Network methods are implemented here. The MFCC features are extracted and are given to SVM for different kernels and also to ANN and their performance is evaluated.

Keyword: Speech Recognition, Mel-frequency Cepstral coefficient, Support Vector Machine, Artificial Neural Network.

I. Introduction

Speech recognition, more commonly known as Automatic Speech Recognition (ASR), is the process of interpreting human speech in a computer. The main goal of speech recognition is to get efficient ways for humans to communicate with computers. Speech recognition is important, not because it is 'natural' for us to communicate via speech, but because in some cases, it is the most efficient way to interface to a computer. The problems in speech recognition are complexity of human language, noise, echo effect, reverberation, channel variability. Men and women have different voices and the main reason to this is that women have in general shorter vocal tract than men. The fundamental tone of women's voices is roughly two times higher than men's because of this difference. Thus gender classification is highly essential.

II. MEL FREQUENCY CEPSTRAL COEFFICIENT (MFCC)

The most prevalent and dominant method used to extract spectral features is calculating Mel-Frequency Cepstral Coefficients (MFCC). MFCCs are one of the most popular feature extraction techniques used in speech recognition based on frequency domain using the Mel scale which is based on the human ear scale. MFCCs being considered as frequency domain features are much more accurate than time domain features.

Mel-Frequency Cepstral Coefficients (MFCC) is a representation of the real cepstral of a windowed short-time signal derived from the Fast Fourier Transform (FFT) of that signal. The difference from the real cepstral is that a nonlinear frequency scale is used, which approximates the behaviour of the auditory system. Additionally, these coefficients are robust and reliable to variations according to speakers and recording conditions.

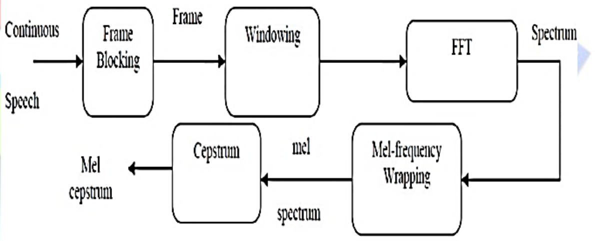


Fig 1.MFCC Extraction

MFCC is an audio feature extraction technique which extracts parameters from the speech similar to ones that are used by humans for hearing speech, while at the same time, deemphasizes all other information. MFCCs use Mel-scale filter bank where the higher frequency filters have greater bandwidth than the lower frequency filters, but their temporal resolutions are the same.

III. MFCC EXTRACTION

A. Frame Blocking

Investigations shows that speech signal characteristics stays stationary in a sufficiently short period of time interval (It is called quasi-stationary). For this reason, speech signals are processed in short time intervals. It is divided into frames with sizes generally between 30 and 100 milliseconds. Each frames overlaps its previous frame by a predefined size. The goal of the overlapping scheme is to smooth the transition frame to frame.

B. Windowing

The second step is to window all frames. This is done in order to eliminate discontinuities at the edges of the frames. If the windowing function is defined as w(n), $0 < n < N\text{-}1$ where N is the number of samples in each frame .The filter coefficients w (n) of a hamming window of length n are computed according to the formula:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1$$

$$= 0 \qquad \text{otherwise}$$

Where N is total number of sample and n is current sample.

C. Fast Fourier Transform

The next step is to take Fast Fourier Transform of each frame. This Transformation is a fast way of Discrete Fourier Transform of each frame. This transformation is a fast way of Discrete Fourier Transform and it changes the domain from time to frequency.

D. Mel Frequency Warping

The human ear perceives the frequencies non-linearly. Researches show that the scaling is linear up to 1kHz and logarithmic above that. The Mel-Scale (Melody Scale) filter bank which characterizes the human ear perceiveness of frequency. It is used as a band pass filtering for this stage of identification. The signals for each frame is passed through Mel-Scale band pass filter to mimic the human ear. As mentioned above, psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency measured in Hz,a subjective pitch is measured on a scale called the melscale. Themel-frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000Hz.As a reference point, the pitch of a 1kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 mels. The relation between frequency of speech and Mel scale can be established as: Frequency (Mel Scaled) = [2595log (1+f (Hz)/700]

IV. SUPPORT VECTOR MACHINE

A. Fundamentals of Support Vector Machines

The Support Vector Machines classification technique stands out by its strong theoretical fundamental, having higher hyperplane, where the linear separation becomes possible. Maximum discrimination is obtained with an optimal placement of the separation plane between the borders of the two classes. A set Hof points x_i∈Rd with i = 1, 2, 3, . is assumed. Each one of the xi belongs to either of two classes labeled $y_i \in \{-1, 1\}$. Establishing the equation of a hyperplane that divides H is the desired goal and for this purpose we have some preliminary definitions. By taking the set H, if linearly separable, there exists w∈Rd and b ∈R to satisfy

$$yi(w \cdot xi + b) \ge 1$$

where i = 1, 2, 3, . . . , n.

The pair (w, b) defines a hyperplane

$$(w \cdot xi + b) = 0$$

This defines a separating hyperplane, leading to the problem of finding the optimal separating hyperplane, to which we try to minimize w as the following

$$\min \frac{1}{2} ||\omega||\text{^}2$$

where $yi(w \cdot xi + b) \ge 1$.

Then converted to a dual problem by Lagrange multiplies

$\max \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_{j(x_I.x_J)}$ This is the general form of SVM. When the data is easily linearly separable, the previous equations are able to classify with minimum error, but when the data is highly nonlinear one needs to use the kernel method, in which the data is put in a higher dimensional plane, where it can be linearly separated. This is possible when we take the dot product of Xi ・Xjand apply another function, validated by the Mercer's Conditions, that in some cases, like the Radial Basis Function (RBF), can place the data in an infinity dimensional space, where the data can easily be separated.

B. Voice Recognition

The voice exists for the human desire of verbalizing its thoughts, emotions and opinions, being part of our identity. From the beginning of its technological and intellectual development, the human beings intended to create machines that were able to produce and understand the human speech. Using voice to interact with automatic systems has a vast field of application. The combination with phone network allows remote access to databases and new services, like, for example, an e-mail check from anywhere on the globe and consultations of flight schedule without needing an operator.

Recently, several methods of Speech Recognition have been proposed using mel-frequency cepstral coefficients and Neural Networks Classifiers. The field of Digital Signal Processing that is the core of many areas in science is to be addressed. From the engineering point of view, signals are functions or series used to carry information from a source to the recipient. The signals specific characteristics depend on the communication used for the transmission. They are processed on the transmission side to be produced and configured, and on the receptor they're decodified to extract the information contained, with maximum efficiency, if possible.

i.Digital Signal Processing

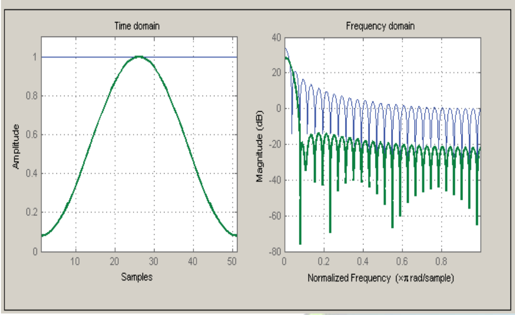


Fig 2 Hamming Window and equivalent SNR

Method that consists in analyzing real world signals (represented by a numerical sequence), extract its features through mathematical tools, in order to extract the essential information. There are many purposes on the matter such as biometric authentication, image processing and recognition and even preventing diseases. Three basic steps are followed which are, sampling, followed by segmentation of words or phonemes and short term analysis by Fourier transform or spectral analysis. After this step, recognition and correct classification of a word is to be performed and for that, there are some existing techniques, which are capable of extracting parameters based on a certain model,and then apply a transformation to represent the signal in a more convenient form for recognition.

ii.Pre-processing of the Speech Signal

The moment the segmentation of the speech is passed through the process of windowing, responsible for 'dividing' the signal with minimum power loss and noise, the speech signal is sampled and segmented into frames and is encoded in a set of mel-cepstral parameters. The number of parameters obtained is determined by the order of mel-cepstral coefficients. The obtained coefficients are then encoded by Discrete Cosine Transform (DCT) in a two dimensional matrix that will represent the speech signal to be recognized.

The process of windowing, hamming windowing for this case, in a given signal, aims to select a small portion of this signal, which will be analyzed and named frame. A short term Fourier analysis performed on these frames is called signal analysis frame by frame. The length of the frame Tf is defined as the length of time upon which a parameter set is valid. The term frame is used to determine the length of time between successive calculations of parameters. Normally, for speech processing, the time frame is between 10ms and 30ms. There's also the superposition of the windowing, which determines where the window will start in order to reduce the power loss, initiating before the previous window reaches its end. Fig shows the plot of a hamming window in time and frequency domains.

iii. SVMs and Biometry

Concerns on widespread use of biometric authentication systems are primarily centered around template security, revocability, and privacy. The use of cryptographic primitives to bolster the authentication process can alleviate some of these concerns as shown by biometric cryptosystems. Support Vector Machines or SVM is one of the most successful and powerful statistical learning classification techniques and it has been also implemented in the biometric field. As for voice recognition, the technique has shown excellent results, hence not only it can generalize, but it can also restrict the parameters if correctly made, leading to a great voice based gender classification system.

iv.Methodology

As a recognition default we proposed the classification and identification of the voice of a speaker by a keyword, in a text-dependent system. The speech signal is sampled and encoded in mel-cepstral coefficients and coefficients of Discrete Cosine Transform (DCT) in order to parameterize the signal with a reduced number of parameters. Then, it generates two dimensional matrices referring to the Discrete Cosine Transform coefficients. The elements of these matrices representing two-dimensional temporal patterns will be classified by Support Vector Machines (SVMs). The innovation of this work is in the reduced number of parameters which lies in the SVM classifier and in the reduction of computational load caused by this reduction of parameters.

After being properly parameterized in mel-cepstral coefficients, the signal is encoded by DCT. One versus all technique is employed to generate SVM. After performing the extraction of the parameters and putting them in the pairs, the Support Vector Machine algorithm is applied in order to generate the hyperplane and classify the new data.

V. Artificial Neural Network

Speaker Recognition is a challenging task and is widely used in many speech aided applications. In order to achieve this task, network can be employed. In machine learning and cognitive science, artificial neural networks (ANNs) are a family of models inspired by biological neural networks (the central nervous systems of animals, in particular the brain) and are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. Humans have the ability to decode the speech signals and understand the information in speech and recognize the speaker. This perception and understanding abilities of humans are needed in many applications such as voice command control, audio archive indexing and audio retrieval etc. The tasks used for speaker recognition are speaker identification, speaker verification or detection and Segmentation and Classification.

A.Tasks for speaker recognition

i. Speaker Identification

Identifies a speaker out of a collection of known speakers using a given voice sample.

ii.Speaker Verification or Detection

Authentication of the speaker is verified by designing a binary decision problem.

iii.Speaker segmentation and classification

Either speech of an individual or when speech of individual is intermixed with other's speech is given as a sample; the desired speech segment must be separated before recognition. This task is useful in multi-speaker recognition problems. Here, the given audio is segmented into homogeneous audio segments and labels are assigned to identify the speaker.

Automatic speaker recognition may be text dependent or text independent. In text dependent category, user gives the text of speech to the system and the knowledge of the phrase is useful for better recognition. In text independent task, the system does not know the phrase spoken in the speech. This will increase the flexibility of recognition System is given in the Figure 3.

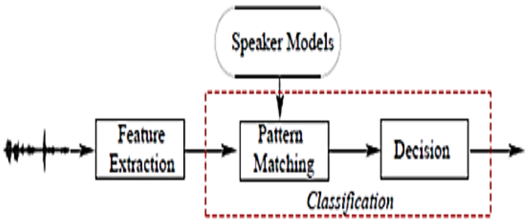


Figure 3. Steps of general speaker recognition problem

Feature Extraction gives the speaker specific information from the given speech signal by performing complex transformations. Different levels of transformations are performed by using semantic, acoustic, phonologic and acoustic features. Pattern matching module compares or matches the extracted features with speaker models. Decision model finds similarity score between the given test sample and the claimed speaker to recognize the speaker.

B.Materials and Methods

Neural Network Toolbox provides algorithms, functions, and apps to create, train, visualize, and simulate neural networks. Classification, regression, clustering, dimensionality reduction, time-series forecasting dynamic system modeling and control can be performed.

The toolbox includes convolutional neural network and auto encoder deep learning algorithms for image classification and feature learning tasks. To speed up training of large data sets, you can distribute computations and data across multicore processors, GPUs, and computer clusters using Parallel Computing Toolbox.

VI. EXPERIMENTS AND RESULTS

To classify the gender, voices of 4 males and 4 females are taken into consideration.40 samples are taken from one person and as a whole 320 samples are taken. For the implementation MATLAB software with Auditory Toolbox is used. MFCC features are then extracted to be given as input to Support Vector Machine and Artificial Neural Network .In Artificial Neural Network, feed forward back propagation model is used.

Among the kernels of SVM, Linear kernel is proved to be the best. In Artificial Neural Network, nftool is employed. On comparing the two methods, SVM proved to be the best. SVM use Structural risk minimization whereas ANN use empirical risk minimization. SVM has global minimum but the ANN has local minimum. ANN handles multiclass classification.

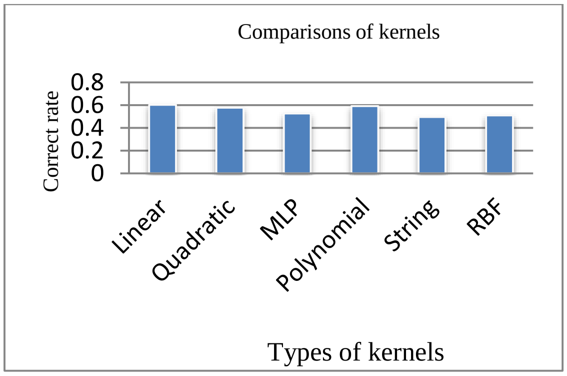


Fig .4: Comparison of kernels

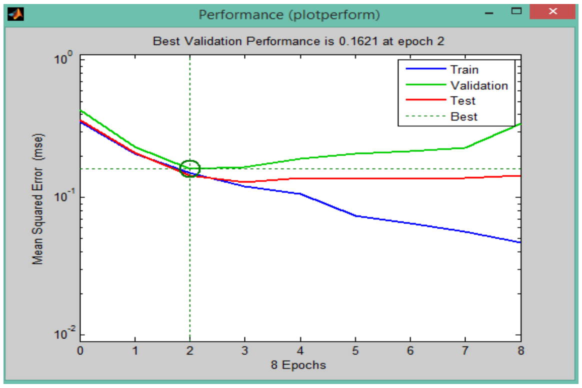
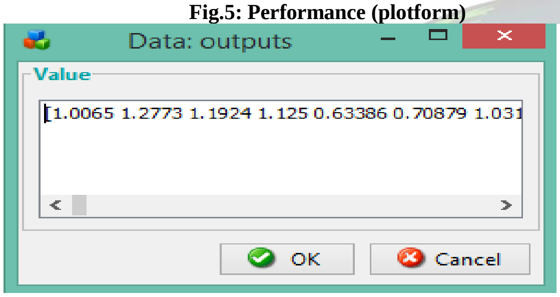


Fig.5: Performance (plotform)



[1.0065 1.2773 1.1924 1.125 0.63386 0.70879 1.031

Fig .6: Sample Data outputs

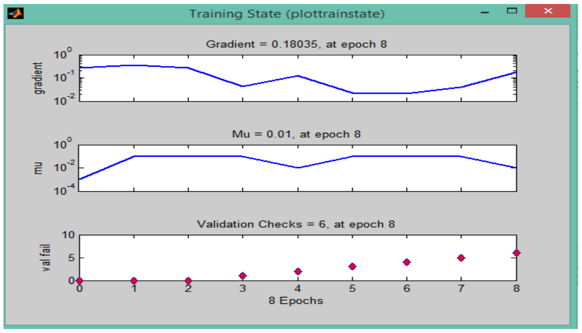


Fig.7: Training state (plot train state)

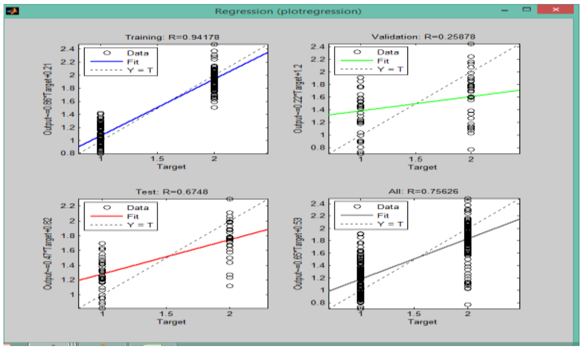


Fig.8:.Regression (plot regression)

VII. CONCLUSION

Gender classification utilizing voice as a input parameter a SVMs to classification, has shown success in general for identifying the speaker. Also the restrictions set by the classifiers, restricts in such a way that prevents false positive to rule over the actual positive results. Analyzing the methodology and applications of SVM, one realizes that it is a technique with excellent response time of computational execution

In order to improve the work, more training data can be given to the system. In order to reduce training and predicting time, one can use the Linear Kernel with little loss of precision and reliability. However, as it has a very wide scope in relation to the classification functions during the learning process of the machines, the SVM ends up compensating for the problem of generalization with the use of more points for classification.

REFERENCES

[1]. Lawrence Rabiner, Biing-Hwang Juang – “Fundamentals of Speech Recognition”

[2]. Wei Han, Cheong-Fat Chan, Chiu-Sing Choy and Kong-Pang Pun – “An Efficient MFCC Extraction Method in Speech Recognition”, Department of Electronic Engineering, The Chinese University of Hong Kong, Hong, IEEE – ISCAS, 2006

[3]. Leigh D. Alsteris and Kuldip K. Paliwal – “ASR on Speech Reconstructed from Short- time Fourier Phase Spectra”, School of Microelectronic Engineering Griffth University, Brisbane, Australia, ICLSP - 2004

[4]. Waleed H. Abdulla – “Auditory Based Feature Vectors for Speech Recognition Systems”, Electrical & Electronic Engineering Department, The University of Auckland

[5]. Pradeep Kumar P and PreetiRao – “A Study of Frequency-Scale Warping for Speaker Recognition”, Dept of Electrical Engineering, IIT- Bombay, National Conference on Communications, NCC 2004, IISc Bangalore, Jan 30 -Feb 1, 2004

[6]. Beth Logan – „Mel Frequency Cepstral Coefficients for Music Modeling’, Cambridge Research Laboratory, Compaq Computer Corporation

[7]. Support Vector Machine Classification of Microarray Gene Expression Data, Michael P. S. Brown William Noble Grundy, David Lin, NelloCristianini, Charles Sugnet, Manuel Ares, Jr., David Haussler

[8]. Text categorization with Support Vector Machines: learning with many relevant features T. Joachims, ECML - 98

[9]. P. Fantinato, Segmentacao de VozbaseadanaAnaliseFractal Transformada Wavelet. Prentice Hall, Outubro 2008.

[10]. L. Rabiner and R. Schafer, Digital Processing of Speech Signals. Prentice Hall, 1978.

[11]. S. Haykin, RedesNeurais:Principio e pratica. Bookman, 2002.

[12]. A. Bresolin, Reconhecimento de vozatraves de unidadesmenores do que a palavra, utilizando Wavelet Packet e SVM, emuma nova EstruturaHierarquica de Decisao. Tese de Doutorado, Natal 2008.

[13]. C. Ding and I. Dubchak, Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics, 2001.

[14]. R. Urena, A. Moral, C. Moreno, M. Ramon, and F. Maria, "Realtime robust automatic speech recognition using compact support vector machines." IEEE Transactions on Audio, Speech, and Language Processing, May 2012

[15]. V. Vapnik and A. Chervonenkis, On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities.Dokl, 1968.

[16]. B. Scholkopf, O. Simard, A. Smola, and V. Vapnik, Prior knowledge in support vector kernels. The MIT Press, 1998.

[17]. P. C. Clarkson and P. Moreno, "Acoustics, speech and signal processing." IEEE International Conference, March 1999

BIOGRAPHY

G Renisha completed her BE degree from Anna University, Chennai in 2008 and ME from PSG College of Technology, Coimbatore in 2010. Presently, she is working as an Assistant professor in the Department of ECE in Govt. College of Engineering, Tirunelveli, Tamilnadu, India and pursuing Ph.d in Anna University, Chennai. Her area of interest is signal processing applications in speech processing.

U.Gomathi Priya B.E., Government College Of Engineering, Tirunelveli.

K.Karpagavalli B.E.,Government College Of Engineering, Tirunelveli.

M.Kaveri Priya B.E., Government College Of Engineering, Tirunelveli.

M.Krishnaveni B.E., Government College Of Engineering, Tirunelveli.