© 2019 IJSRSET | Volume 6 | Issue 1 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology
DOI : https://doi.org/10.32628/IJSRSET196110

A Hybrid Approach to Gender Classification using Speech Signal

M. Yasin Pir[1], **Mohamad Idris Wani**[2]

[1]Department of Computer Applications, Govt. Degree College, Pattan, Baramulla, Jammu and Kashmir, India

[2]Department of Electrical Engineering, Jamia Millia Islamia, New Delhi, India

ABSTRACT

Speech forms a significant means of communication and the variation in pitch of a speech signal of a gender is commonly used to classify gender as male or female. In this study, we propose a system for gender classification from speech by combining hybrid model of 1-D Stationary Wavelet Transform (SWT) and artificial neural network. Features such as power spectral density, frequency, and amplitude of human voice samples were used to classify the gender. We use Daubechies wavelet transform at different levels for decomposition and reconstruction of the signal. The reconstructed signal is fed to artificial neural network using feed forward network for classification of gender. This study uses 400 voice samples of both the genders from Michigan University database which has been sampled at 16000 Hz. The experimental results show that the proposed method has more than 94% classification efficiency for both training and testing datasets.

Keywords : Wavelet, Neural Networks, Stationary Wavelet Transform, Sampling, Power Spectral Density and Gender Classification.

I. INTRODUCTION

Speech is the basic inherent means of communication for humans and it is a combination of physiological and behavioural biometrics. Speech not only contains the information to be convened but also contains information about the speaker itself and due to the advancement in voice recording technology, this information about the speaker can be quantified from a speech signal of a person. This information can be used in number of applications ranging from gender classification to person identification. The process of recording speech samples of persons, sampling it and extracting different parameters from it is called speech processing and comes under the umbrella of signal processing.

Basis of classifying the gender of a person from their speech signal arises due to the fact that both male and female voice have quantifiable features such as power, frequency, power spectral density, pitch etc. and almost all these features lie in a particular range for a particular gender as compared to another gender and hence the task of classifying the gender of a person can be summed up to be a two class problem.

Automated Gender classification from speech signals is a two-class problem where in different quantifiable parameters of a speech signal are extracted using various techniques and using these parameters a classifier is trained to recognize the gender of a person from his/her speech. One of the main advantages of using speech for automated gender classification is the amount of ease with which this biometric feature can be obtained and exposes the minimum biometric data to a system employing gender classification.

Neural Networks are simplified models of the biological nervous systems and therefore have drawn their motivation from the kind of computing performed by a human brain [1].Neural Networks are efficient information processing systems with advantages of parallelism, adaptivity, and fault tolerance with real time operation [2]. Such networks come under the domain of soft computing techniques which produce an approximate result. Neural Networks have become very popular in the past decade and are showing promising signs of tackling the problems posed by current information processing systems. Artificial Neural Networks have multiple processing units called neurons and these neurons are interconnected with each other by links and these links are modified by their respective connection weights. If a neuron in a particular layer is not connected to any neuron in the same layer or any neuron in the preceding layer it is a Feed Forward Network else a Recurrent Network.

Learning is a process by which an ANN makes itself adjustable to the stimulus by making changes in the free parameters of the ANN; this adaptability is brought about by making change either in the connection weights (parameter learning) or by changing the entire structure of the ANN (Structure Learning) [2]. There are mainly three types of learning:

A. Supervised Learning: This type of learning is also known as learning with a teacher, and for each input pattern the desired (target) pattern is presented hence the network is being taught exactly what should be the output.
B. Unsupervised Learning: In this type of learning, the network is not presented with any type of information about what should be the output, however similar types of data patterns are grouped together for specifying how the member of each group looks like.
C. Reinforced learning: It is similar to supervised learning but where in the systems is trained to do a particular job and learns on its previous experiences and outcome while doing a similar kind of a job.

The manner in which neurons are structured and the way in which neurons are connected to other neurons in the same layer and the neurons in the preceding and succeeding layers forms the architecture of the network. There are much architecture proposed in the literature and each has its advantages and shortcomings.

Due to massive parallelism of neurons, Artificial Neural Networks (ANN's) can be employed in a number of applications ranging from linearization to pattern solving. ANN's because of their high adaptability and learning capabilities can be used be pattern classification.

Wavelet analysis shares some common features with Fourier analysis but wavelet analysis has the advantage of capturing features in the time series that vary across both time and frequency. Wavelets represent mathematical functions that can decompose data into different frequency components. After decomposition, every component is studied with a resolution matched to its scale. These functions are generated by the dyadic dilations and integer shifts of a single function called a *mother wavelet*. The time-frequency localization is the key feature of wavelets and most of the energy of the wavelets is restricted to a finite time interval and its Fourier transform is band limited. When compared to Fourier analysis, the advantage of the time-frequency localization is that wavelet analysis varies the time-frequency aspect ratio, producing good quality time resolution and poor-quality frequency resolution at higher frequencies and a good quality frequency resolution and poor-quality time resolution at lower frequencies. This approach is reasonable when the signal on hand has low frequency components for long durations and high frequency components for short durations and the signals that we encounter in most economic applications are frequently of this type.

The minimum requirements imposed on a function$\psi(t)$to qualify for being a *mother wavelet* are that $\psi(t) \in L^2(\mathbb{R})$ (space of square integrable functions) and also fulfils a technical condition, usually referred to as the *admissibility condition*

$$0 < C_\psi = \int_0^\infty \frac{|\psi^\wedge(\xi)|^2}{|\xi|} d\xi < \infty, \qquad \text{-- (1.1)}$$

Where $\psi^\wedge(\xi)$ is the Fourier transform of $\psi(t)$ [3]. To endorse that $C_\psi < \infty$, a wavelet function must satisfy the conditions $\int_{-\infty}^{\infty} \psi(t)dt = 0$ and$\int_{-\infty}^{\infty} |\psi(t)|^2 dt = 1$. These two conditions mean that:

1. $\psi(t)$must be an oscillatory function with zero mean and
2. The wavelet function has unit energy.

There can be two types of wavelets within a given function or family depending on the normalization rules where father wavelet is represented by **(1.2)** and mother wavelet is represented by **(1.3)** and j = 1……, *J* and *J*-level wavelet decomposition can be written as [4]:

$$\phi_{j,k} = 2^{-j/2}\phi\left(t - 2^j k/2^j\right) \qquad \text{--- (1.2)}$$

$$\psi_{j,k} = 2^{-j/2}\psi\left(t - 2^j k/2^j\right) \qquad \text{--- (1.3)}$$

Wavelet transform can be broadly classified into two types:

1. Discrete Wavelet Transform (DWT)
2. Continuous Wavelet Transform (CWT)

CWT is an implementation of the wavelet transform using arbitrary scales and almost arbitrary wavelets. The wavelets used in this case are not orthogonal and the data obtained by this transform are highly correlated. On the other hand, DWT is an implementation of the wavelet transform using a discrete set of the wavelet scales and translations obeying some defined rules and decomposes the signal into mutually orthogonal set of wavelets [5].

Unlike Fourier analysis, a signal is decomposed into a set of mutually orthogonal wavelet basis functions in DWT and they differ from sinusoidal basis functions since they are spatially localized i.e., non-zero over only part of the total signal length. Wavelet functions are dilated, translated and scaled versions of a common function, known as the mother wavelet. Since DWT is invertible, the original signal can be completely recovered from this representation. DWT is not just a single transform as in case of Discrete Fourier Transform, rather a set of transforms, each with a different set of wavelet basis functions. Two of the most common are the Haar wavelets and the Daubechies set of wavelets.

Wavelet Transform has found applications is various fields such as;

a. Image compression
b. Edge and corner detection
c. Denoising
d. Filter designing
e. ECG (Electrocardiogram) Analysis

And there are many more applications of wavelet transform and to mention all of them is not in the scope of this work.

DWT gives a sparse representation for most of the signals, and hence most of the natural signals are captured by a subset of DWT coefficients and is typically much smaller than the original signal. Hence the transform has the same number of coefficients as the original signal but most of them are closer to zero in value hence a compressed form of signal can be obtained and with a high signal quality. However, the CWT is a highly redundant transform [6].

During the study Daubechies wavelet transform was used and the reason for using this transform is that it has more coefficients for every vanishing moment while as Haar wavelet transform only has two coefficients for every vanishing moment.

II. BRIEF LITERATURE REVIEW

Automatic gender identification from speech is an important problem with many applications including speaker identification, speaker segmentation, and personalizing human-machine interactions.

Due to the recent advancements in the techniques of voice recording, it has become possible for researchers to quantify the parameters of voice, both time related as well as frequency related. Gender identification on the basis of a speech signal is an automated process of identifying the gender of a speaker from its speech signal.

Automated gender classification from speech signals may be time domain based or frequency domain based; when a speech signal is considered directly for taking measurements i.e. obtaining information about the speaker it is called time domain analysis; whereas if the frequency content of a speech signal is used to form a spectrum for evaluating the information about a speaker, it is called frequency domain analysis [7]. Identification of a person's gender from his or her speech mainly revolves around the difference in the values of power and frequency content of the two genders. Ali M.S. et.al. proposed a system which uses PCM (Principal Component Analysis) for encoding the speech signal in digital domain, then the signal was subjected to Discrete Fourier Transform (DFT) for computing the frequency information of the signal. However, the speech signal consists of only real points; the proposed system makes use of Real Point DFT for increasing efficiency [8]. The system further proposes to use Short Time Fourier Analysis of a windowed signal for producing a reasonable feature space for recognition. The system recorded an accuracy of 80%.

Harb H. et.al. proposed a system for identifying the gender of a person from their speech signal. For signal analysis the system uses Fast Fourier Transform (FFT) with a Hamming window of 30ms width and a 20ms overlap. The spectrum is further filtered conforming to the Mel Scale to obtain a vector of 20 Spectral coefficients every 10ms [9]. Then in each window mean and variance of the MFSC vectors are calculated, and for forming a feature vector, 40 such vectors are concatenated. The feature vector is then normalized so as to ensure that classifier captures the relation between the frequencies in the spectrum and not the frequencies themselves. A Neural Network has been used as a classifier. The system recorded a total accuracy of about 92%.

Martin A. F. et.al. have presented one of the most novel works in the areas of gender identification using speech signals. The system proposed multi speaker detection with multiple linguistics [10]. In this study rather than considering a single speaker, multi speaker analysis were taken into consideration for tackling practical environment with greater efficiency.

Khan A. et.al. proposed a gender classification system using fuzzy logic [11], the fuzzy system is trained using different parameters such as power amplitude, total harmonic distortion and power spectrum. Although the study shows satisfactory results but the problem with fuzzy logic is that as the rule base increases, the complexity of the system increases and it does not include the added benefit of learning.

Other hybrid systems have also been proposed in the literature, although they show a promising solution to the problem but at the same time increases the system complexity considerably and also no attention has been given to tackle the problem of noise. Also, these systems require greater computational time because of greater complexity [12 , 13].

Prabha M. et.al. proposed a system which uses energy for classifying the gender; the signal transformation is done using FFT and the system recorded an efficiency of 93.5% [14]. A machine learning algorithm used in [15] recorded an accuracy of 92.5%. During the preprocessing phase, the windowing and pre-emphasis were used to for noise cancellation.

III. METHODOLOGY

The proposed technique includes application of 1-D Stationary wavelet transform (DWT) on voice samples. We use Daubechies wavelet function at different levels to decompose the voice sample for denoising, analysis, and feature extraction. The resultant wavelet coefficients are stored in a file. During the training phase, the artificial neural network is trained using feed forward network with 250 voice samples in the dataset that includes values of corresponding coefficients with label 0 for male and label 1 for female images. Similarly, 150 voice samples of both genders from the dataset are used in testing phase. Finally, gender classification rate is calculated in terms of Mean Squared Error (MSE) and Region of Convergence (ROC). Hence, gender classification is achieved in two main steps. In the first step, features are extracted from the voice samples using Daubechies wavelet and in the second step, the selected features are classified using artificial neural network. Following are the steps of the proposed technique:

1. Read all voice samples one by one.
2. Apply the 1-D Stationary wavelet transform (DWT) to each of the voice samples for decomposition.
3. Take the denoised and reconstructed dataset of all voice samples along with the label (0 for male, and 1 for female).
4. Train and test the network using the denoised speech signals.
5. Calculate the gender classification rate.

The framework of the proposed model is depicted in Figure 1.

IV. RESULTS & DISCUSSION

This study uses 400 voice samples of both the genders from Michigan University database which has been sampled at 16000 Hz. The database is divided into two segments one for training and the other for testing. The speech samples of the above-mentioned database were subjected to Daubechies wavelet transform at different levels and during decomposing and reconstruction, it was observed that an optimum result of decomposition and reconstruction was obtained at level 4 of the Daubechies wavelet using unscaled white noise Stationary Wavelet Transform for 1D signal. A number of reconstructed signals are concatenated to form a feature vector so as to ensure proper training of the classifier.

Speakers from both genders speak several words such as ‘had’, ‘have’ etc. depicting the variation in speech signals.

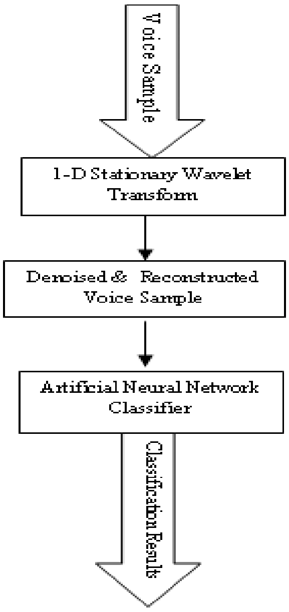
Voice Sample
1-D Stationary Wavelet Transform
Denoised & Reconstructed Voice Sample
Artificial Neural Network Classifier
Classification Results

Figure 1 : Framework of Proposed Model

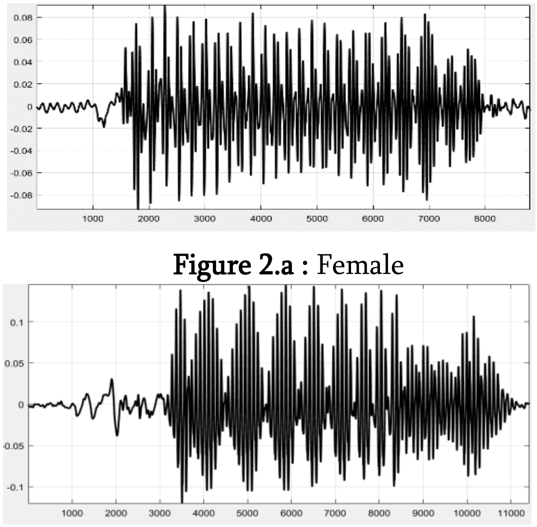

Figure 2.a : Female

Figure 2.b : Male

Figure 2 : Denoised Signal of a female and male saying word “heard”.

The classifier is trained using the denoised feature vectors and using optimum number of epochs and training of samples is essential so as to ensure the classifier works in a proper manner.

For obtaining better results and better understanding of the system a feature vector consisting of only single words such as “heard” “had” etc. were used for training the classifier, but while training with these feature vectors consisting of voice samples of different speakers of both genders the system would lag behind in efficiency as the system started converging at 40% or 50%.

However, when a feature space was created using different words spoken by both male and female gender, the classifier started working efficiently and yielded way better results. The reason for better efficiency when using different words is that the network gets more and more accustomed to the variability of the human speech samples and hence performs efficiently. Also, the network was trained at different number of iterations and various ratios of voice samples were used for training, testing and validations. The system shows an overall accuracy of 94% when 70% of the samples are used for training 15% for testing and 15% for validation.

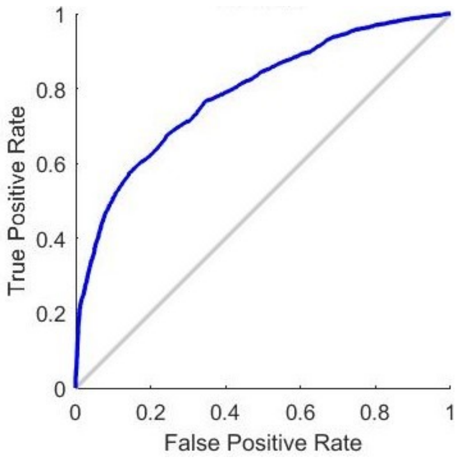

Figure 3. b: ROC for ‘head’

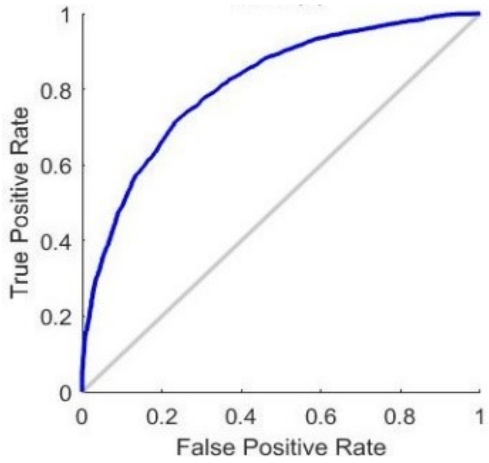

Figure 3.c: ROC for ‘heard’

Figure 3 : ROC Of system trained using feature vectors containing only single words spoken by different speakers; a: is ROC for had, b: for head and c: for heard

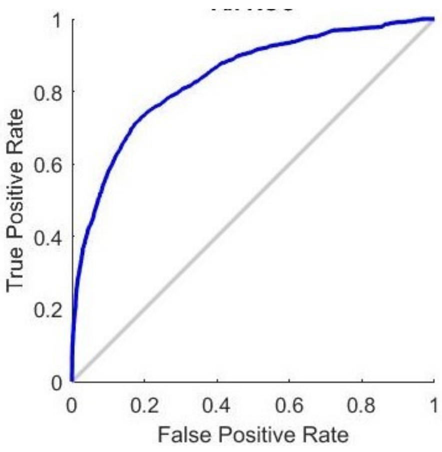

Figure 3. a : ROC for ‘had’

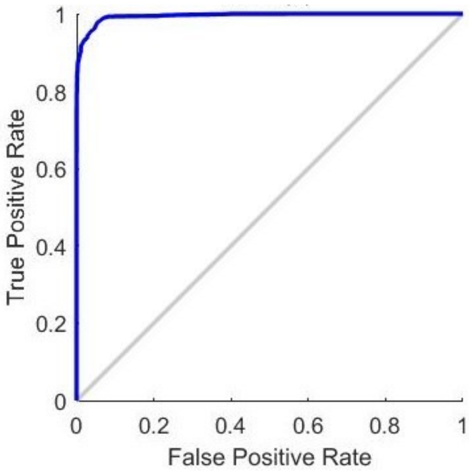

Figure 4 : All ROC; overall ROC of the system trained using feature vectors consisting of different denoised words.

V. CONCLUSION

The main goal was to develop a gender recognition system using speech signal. This study proposes a hybrid approach of Gender Classification by combining Wavelet Transform and Artificial Neural Network wherein 1D Stationary Wavelet Transform is used for denoising and subsequently reconstruction of voice samples and Artificial Neural Network is used as classifier Model. The proposed hybrid approach of gender classification has better classification rate than most of the other techniques proposed in the literature in terms of accuracy acquired and simplicity. The proposed model was implemented in Matlab and the average recognition accuracy is 94 %.

VI. REFERENCES

[1]. Rajasekaran S. and Vijaylakshmi G. A.(2003). "Neural Networks, Fuzzy Logic and Genetic Algorithms: Synthesis and Applications", PHI.

[2]. Sivanandam S. N. and Deepa S. N(2011). "Principles of Soft Computing", 2nd Edition, Wiley.

[3]. Crowley, P.(2007). "A Guide to Wavelets for Economists. Journal of Economic Surveys", 21 (2), pp. 207–267.

[4]. Kaijian H., LeanY., Kin K.L.(2012). "Crude oil price analysis and forecasting using wavelet decomposed ensemble model", Journal of Energy and Exergy Modelling of Advance Energy Systems, 46(1), pp.564–574.

[5]. Chang S. G., Yu B., Vetterli M. (2000). "Adaptive wavelet thresholding for image denoising and compression", IEEE Trans. Image Processing, 9(9), pp. 1532-1546.

[6]. Brassarote G.O.N., Souza E.M., Monico J.F.G. (2018). "Non-decimated Wavelet Transform for a Shift-invariant Analysis", Trends in Applied and Computational Mathematics,19(1), pp. 93-110.

[7]. Wani M. I, Farooqi B, Wani N, Mehraj (2018). "Speech Based Gender Classification", Emerging Trends and Innovations in Electronics and Communication Engineering - ETIECE-2017, 5(1), e-ISSN: 2348-4470 .

[8]. Ali M. S, Islamand M. S, Hossain M. A. (2012). "Gender Recognition System Using Speech Signal", International Journal of Computer Science, Engineering and Information Technology, IJCSEIT, 2(1),ISSN: 2231-0711, pp. 118-120.

[9]. Harb H, Chen L.(2003)."Gender Identification Using A General Audio Classifier", Dept. Mathématiques Informatique, Ecole Centrale de Lyon, France.

[10]. Martin A. F, Przybocki M. A.(2001). "Speaker recognition in a multi-speaker environment", 7th European Conference on Speech Communication and Technology, (Eurospeech 2001), Denmark, pp. 787–90.

[11]. Khan A, Kumar V, Kumar S.(2017), "Speech Based Gender Identification Using Fuzzy Logic" International Journal of Innovative Research in Science, Engineering and Technology 6(7), ISSN(Online): 2319-8753, pp. 14344-51.

[12]. Khanum S, Sora M.(2015). "Speech based Gender Identification using Feed Forward Neural Networks", National Conference on Recent Trends in Information Technology, International Journal of Computer Applications , pp 5-8.

[13]. Meena K, Subramaniam K, and Gomathy M.(2013). "Gender Classification in Speech Recognition using Fuzzy Logic and Neural Network", The International Arab Journal of Information Technology, 10(5), pp. 477-485.

[14]. Prabha M, Viveka P, Sreeja G. B.(2016). "Advanced Gender Recognition System Using Speech Signal", IJCSET , 6(4),pp. 118-120.

[15]. Kaur D.(2014). "Machine Learning Based Gender Recognition and Emotion Detection",

International Journal of Engineering Sciences & Emerging Technologies, IJESET, 7(2), ISSN:22316604,pp.646-651.

Cite this article as :

M. Yasin Pir, Mohamad Idris Wani, "A Hybrid Approach to Gender Classification using Speech Signal", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), ISSN : 2456-3307, Volume 6 Issue 1, pp. 17-24, January-February 2019.

Available at doi :

https://doi.org/10.32628/IJSRSET196110
Journal URL : http://ijsrset.com/IJSRSET196110