International Conference on Smart Data Intelligence
(ICSMDI 2021)

Gender Recognition and Classification of Speech Signal

***Bhagyalaxmi Jena*[1], *Anita Mohanty*[2], *Subrat Kumar Mohanty*[3]**

[1]*Department of Electronics and Communication Engineering, Silicon Institute of Technology, Bhubaneswar*
[2]*Department of Electronics and Instrumentation Engineering, Silicon Institute of Technology, Bhubaneswar*
[3]*Department of Electronics and Communication Engineering, College of Engineering Bhubaneswar, Bhubaneswar*
E-mail address: bjena@silicon.ac.in, anita@silicon.ac.in, skmohanty0509@gmail.com

ABSTRACT

Human speech comprises different sounds for the purpose of speaking, singing, and expressing emotions and ideas, etc. which is generated by the vibration of the vocal cord of human beings. The rate of vibration of vocal cord is termed as pitch which is a frequency domain parameter. Vocal tracts of males are mostly longer in comparison to vocal tracts of females which leads to differences in human speech for males and females. With increase in demand of Human Computer Interaction (HCI) systems, speech processing plays an important role in enhancing HCI systems. The development of gender recognition systems finds its applications in gender based virtual assistants, telephonic surveys and voice controlled automation systems. Research works were mostly done in either frequency domain or time domain. In this work, speech signal analysis was based on both time and frequency domain. Different speech parameters were generated by short-time, statistical and spectral analysis. The differences in parameters was used as a working principle for the gender model to recognize the gender of the unknown user. The classifier model based on Genetic Algorithm, Gaussian Mixture Model (GMM) is known to have an accuracy of 70% with complex training. So, the classifiers used in this work were KNN and SVM. After training and testing, the accuracy of the system was found to be 80 percent on average.

Keywords: STAM, STE, ZCR, KNN, SVM

1. Introduction

Human speech is formulated by the vibration of vocal cord of human being consisting of different sounds for speaking, singing, expressing emotions and ideas, etc .The vocal cord being an important source of sound with human voice as a part of human sound synthesis [3] [5]. With advancing technology and more and more usage of human computer interaction based systems, the speech processing plays an important role in enhancing the HCI system [2] [4]. This project is based on analysis of speech signals and its different parameters to devise a gender classifier. The main aim of the gender classifier is to predict the gender of the speaker by analyzing different parameters extracted from the voice sample [1]. Earlier investigation focuses on frequency domain analysis using pitch or cepestral coefficients [7] [9]. This classification process and analysis mainly concentrates on short time and statistical analysis of the speech signals. The analysis also includes comparison of short-time average magnitude and short-time energy of male and female voice samples. This overall analysis is implemented with help of MATLAB programming. The short-time and statistical analysis was performed on all the voice samples collected and the features extracted were compared to establish a working principle for the gender classifier from speech signal [6] [8]. The goal of classifier is to design a gender model which could recognize gender of the input speech signal in recognition phase [10]. The recognition accuracy of this system is 80 percent on average.

2. Speech Database

Recording for 300 different males and females of age group 20-22 were done. The sentence recorded was “A Quick Introduction for Scientists and Engineers”. The recording was done in recording specific software Sony Sound Forge. The speech signals were recorded with sampling frequency of 44100 Hz in .wav format in Mono Channel. The noise reduced audio signal were saved in two different folder under the name of male and female for the male and female speech files after reducing noise to form Speech Database. The plots for 20 male and 20 female were tested and plotted.

3. Statistical Analysis

Mean is sum of the amplitudes of the samples of the speech signal x, starting from the first sample to the last sample divided by the total number of samples N of the speech signal. The average or mean of the speech signal can be given mathematically by Eq. 1,

$$\mu = \frac{1}{N}\sum_{i=1}^{N} S(i) \qquad (1)$$

Where, μ= Mean of the speech signal

N = Number of speech samples

S = Speech signal

Table 1. Mean of Male and Female Speech Signal

SL No.	Mean	Label
1.	1.29E-04	Female
2.	8.10E-07	Male
3.	6.51E-05	Male
4.	3.77E-05	Female
5.	5.45E-05	Female
6.	7.29E-05	Male
7.	5.12E-04	Female
8.	8.07E-07	Male
9.	3.30E-05	Female
10.	4.85E-07	Male

Variance of speech signal obtained in statistical analysis is a measurement of the spread between amplitudes of samples in the signal. Basically, it determines how far each sample's amplitude is from the mean value and therefore with respect to other samples of the speech signal. The variance of the speech signal can be given mathematically by Eq.2,

$$\text{Variance } = \frac{1}{N}\sum_{i=1}^{N} S(i)^2 - \left(\frac{1}{N}\sum_{i=1}^{N} S(i)\right)^2 \qquad (2)$$

Where, Variance =Variance of the speech signal

N = Number of speech samples

S = Speech signal

Table 2. Variance of Male and Female Speech Signal

SL No.	Variance	Label
1.	0.01424689	Female
2.	0.01430108	Male
3.	0.00731786	Male
4.	0.01220707	Male
5.	0.016568828	Female
6.	0.008568367	Male
7.	0.008231633	Male
8.	0.04022475	Female
9.	0.012413181	Female
10.	0.017245887	Female

4. Technique

4.1 Short-Time Average Magnitude (STAM)

Short-Time Average Magnitude (STAM) helps in the detection of the start point and the end point of the speech signal. In Speech Transmission Systems that multiplex several conversations, STAM can help detect the boundaries of speech so that the pauses need not be sent. The mathematical expression for Short-Time Average Magnitude of a speech signal is given by Eq.3,

$$M_n = \sum_{k=-\infty}^{\infty} |S(k)|W(n-k) \tag{3}$$

Where, Mn= Short-Time Average Magnitude

S = Speech Signal

W = Window Signal

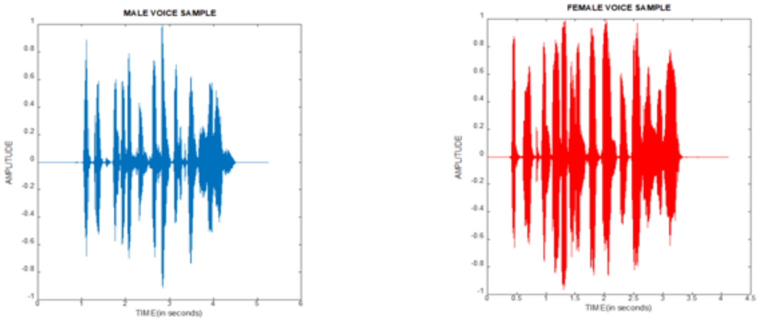

Fig. 1 Input Male and Female Speech Signal

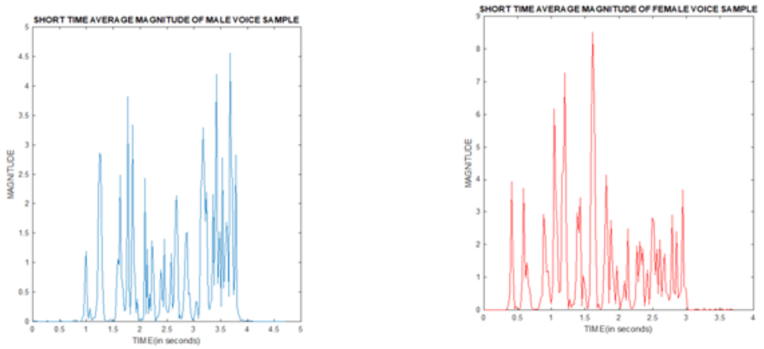

Fig. 2 Output of STAM of Male and Female Speech Signal

4.2 Short Time Energy (STE)

Short-Time Energy (STE) is the energy content of the signal in time domain. Such measures can help segment speech into smaller phonetic units. The large variation in amplitude between voiced and unvoiced speech, as well as smaller variations between phonemes with different manners of articulation, permit segmentations based on energy in gender recognition systems. For isolated word recognition, energy feature can add accurate determination of the end points surrounded by pauses. The mathematical expression for Short-Time Energy of a speech signal is given by Eq.4

$$E_n = \sum_{k=-\infty}^{\infty}[S(k)W(n-k)]^2 \tag{4}$$

Where, E_n = Short-Time Energy

S = Speech Signal

W = Window Signal

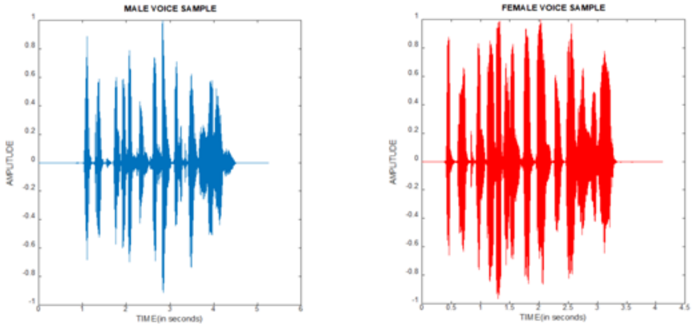

Fig. 3 Input Male and Female Speech Signal

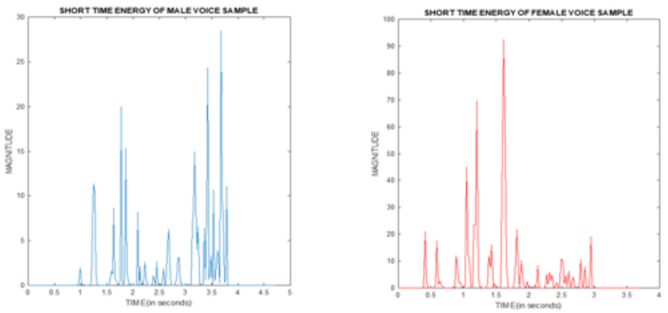

Fig. 4 Output of STE of Male and Female Speech Signal

5. Classifier

Gender classifier is the basic unit of any automatic speech recognition system. A gender classifier is embedded with programs which compare various speech parameters for determining the gender of the speaker.

5.1 KNN

Often abbreviated KNN stands for k-nearest-neighbor, is an approach used for classification of data by estimating probability of belongingness of a data point to be an associate of one among different classes depending on the label of the points nearest to belongs to. KNN Algorithm have its basics from feature similarity technique i.e. how closely out-of-sample features resemble our training set determines how we classify a given data point. The appropriate value of k is determined by the size of training set; in most cases , higher k value eliminates the interference on the classification, but may minimizes the boundaries between different classes. The k-nearest-neighbor is also an instance of a ”lazy learner” algorithm as until any data set query is done, it does not build any model. The flow chart of KNN algorithm is given in Fig.5.

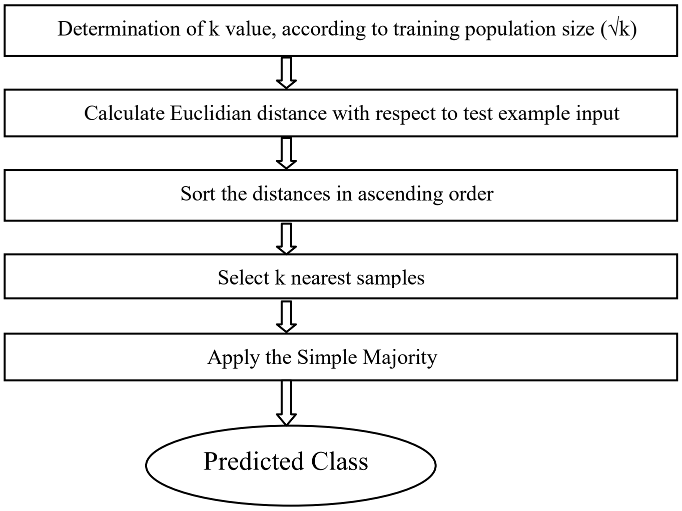
Determination of k value, according to training population size (√k)
Calculate Euclidian distance with respect to test example input
Sort the distances in ascending order
Select k nearest samples
Apply the Simple Majority
Predicted Class

Fig. 5 Flow Chart of KNN Algorithm

Table 3. Predicted Output of KNN Model

SL No.	Actual Output	KNN Output
1.	Male	Male
	Male	Male
3.	Female	Male
4.	Male	Female
5.	Male	Male
6.	Female	Male
7.	Female	Female
8.	Female	Female

The accuracy of KNN Model from above Table 3 was found out to be 62.5%.

5.2 SVM

SVM stands for support vector machine is basically a machine learning algorithm that analyses data for the purpose of classification and regression analysis. A SVM is also known as support vector network (SVN). SVM designs linear model depending on the support vectors which forms the basis to estimate decision function [11]. For linearly separable training data set, SVM estimates the optimal hyper plane which can separate the points of different classes without error. SVM through a non-linear mapping, maps the input patterns into higher dimension feature space. For linearly separable data, a linear SVM is mostly used to classify the data sets. In case of non-linearly separable data, kernel functions are used to make it linearly separable data.

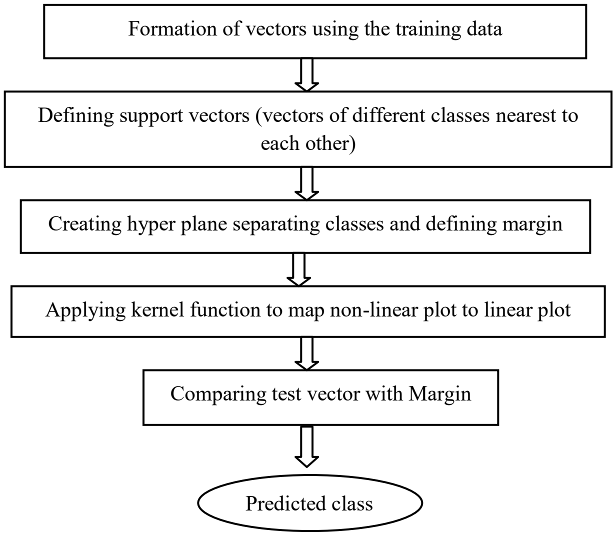
Formation of vectors using the training data
Defining support vectors (vectors of different classes nearest to each other)
Creating hyper plane separating classes and defining margin
Applying kernel function to map non-linear plot to linear plot
Comparing test vector with Margin
Predicted class

Fig. 6 Flow Chart of SVM Algorithm

The input patterns which lie on or near to the margins which are exaggerated are called the support vectors. In SVM, each data set comprises of a pair of input features and their expected output value. Supervised learning algorithm examines the training data set and formulates the required kernel function, which helps to map new outcomes. Mostly, Metrics of SVM is the minimization and generalization of upper bound error by expanding the margin which separates the data set from the hyper plane. The flow chart of SVM algorithm is shown in Fig.6.

Table 4. Predicted Output of SVM Model

SL No.	Actual Output	SVM Output
1.	Male	Male
2.	Male	Male
3.	Female	Male
4.	Female	Female
5.	Male	Male
6.	Male	Male
7.	Female	Female
8.	Female	Female

The accuracy of SVM Model from above Table 4 was found out to be 87.5%.

6. Conclusion

The gender recognition of speech signal using KNN and SVM was found to be more precise in SVM which was found to have an accuracy of 87.5% in comparison to KNN with an accuracy of 62.5%. From the statistical and short time analysis of speech signal, it was found that parameters like mean, variance, STAM, STE for female had higher value in comparison to male.

References

[1] Burkhardt F, Ballegooy M, Englert R, Huber R. An emotion-aware voice portal. In Proc. ESSP, 2005; 123-131.

[2] Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W F, Weiss B. A database of German emotional speech. In Proc. Interspeech. 2005; 1517- 1520.

[3] Sigmund M. Gender distinction using short segments of speech signal. International Journal of Computer Science and Network Security. 2008; 8(10):159–162.

[4] Chaudhari S J, Kagalkar R M. Methodology for gender identification, classification and recognition of human age. International Journal of Computer Applications. 2015; 975:8887.

[5] Wu K, Childers D G. Gender recognition from speech. Part i: Coarse analysis. The journal of the Acoustical society of America. 1991; 90(4):1828–1840.

[6] Livieris I E, Pintelas E, Pintelas P. Gender recognition by voice using an improved self-labeled algorithm. Machine Learning and Knowledge Extraction. 2019; 1(1):492–503, 2019.

[7] Pahwa A, Aggarwal G. Speech feature extraction for gender recognition. International Journal of Image, Graphics and Signal Processing. 2016; 8(9):17.

[8] Duda R O, Hart P E, Stork D G. Pattern classification. John Wiley & Sons. 2012.

[9] Hasan M R, Jamil M, Rahman M. Speaker identification using Mel frequency cepstral coefficients. Variations.2004; 1(4).

[10] Vogt T, André E. Improving automatic emotion recognition from speech via gender differentiation. In Proceedings of the Language Resources and Evaluation Conference, Genoa, Italy. 2006; 1123–1126.

[11]Jena B, Mohanty A,Mohanty S.K,Gender Recognition of Speech Signal using KNN and SVM,ICICNIS 2020.