International Current Journal of Engineering and Science (ICJES)

Volume 1, Issue 1, pp 1-5, December-2022

Available Online at: www.icjes.com

Gender Identification from Speech Signals Using Mel Frequency Cepstral Coefficients Based Feature Extraction

[1]A. M. Ibrahim, [2]Dr. Abdallah Ghallab Eldin

[1,2]Faculty of Engineering, Computer and Systems, Ain Shams University, Egypt

***Abstract* - Gender identification of a speaker, which is the everyday distinguishing speech's characteristic. It can effortlessly be identified by an individual who hears it. It is substantially vital for many applications to identify gender information driven from signals of speech. With the help of gender recognition, the systems which are dependent on gender are defined. Proper gender identification can increase the efficiency and robustness of any gender-dependent system. In this research, Identification of gender is developed using MFCC coefficients and other acoustic properties taken from signals of speech with GMM. Testing was conducted using Surftech's Free American English Dataset (SLR45) with speech from ten speakers (five females and five males). So here we are determining the gender of a speaker using MFCC and other acoustic features and GMM and five other types of machine learning algorithms (Neural Network, Decision tree, Random Forest SVM and Gradient boosting) for classification of gender. The results achieved show that GMM and gradient boosting perform better using MFFCC and other acoustic features.**

Keywords: Machine Learning, SVM, Gradient boosting, Neural Network, Mel-frequency cepstral coefficients, Gaussian mixture model.

I. INTRODUCTION

Any listener with a speech of content may easily understand speech characteristics, for instance, age, the accent of gender, and the emotional state of a speaker. In a speech, the most distinct feature is a speaker's gender. Automated gender identification is helpful in many areas. Preliminary information of gender of the speaker in automated gender identification provides the means to define models which depend on gender, and gender-dependent models provide comparatively good results the other system [1, 2].

In gender identification, conversion of speech has to be done first, which is technically known as features, .so feature selection is a driving factor, amongst others improving the success rate of the system. With the advancement in technology, there are many ways to denote speech signals. Different features are applied. Frequency pitch, the format of frequency and energy [3,4]. There are various classification methods to model speech signals transformed into feature vectors: GMM Gaussian Mixture Model, SVM Support Vector Machine, (HMM) Hidden Markov Model and (NN) Neural Network are usually applied in the identification of speech [5,6,7].

In the fundamental frequency is a mainly used feature for recognition of a speaker's gender [8]. The fundamental frequency is entirely dependent on the quality of the voice. It can be collected from portions of the voice of speech. In the pitch frequency of males and females, there is an intersection point; hence fundamental frequency is not sufficient for gender recognition. Till now, exploration of different features and classification has been done. In research using fundamental frequency and the first two formants, 88.42% accuracy is obtained [3]. In this research, the Fault rate is considerably decreased by making use of features of MFCC and fundamental frequency [5]. Finally, we are doing the experiment and determining the MFCC's and other extracted voice features success feature vector in recognition of gender recognition is analyzed for ten English speakers (five females and five males) the result achieved is 95% and we have also used the same extracted for classification of gender, five algorithms (SVM, Decision tree, Random Forest Gradient boosting and Neural Network) were used here for classification gender.

II. FEATURE EXTRACTION

Signal of speech not only contains information, but it holds information about the age of an individual, gender of a speaker, emotional state of a speaker. Determination of essential features is done in the first phase of this identification system. Once the essential features are decided then in next stage the conversion of the speech signal into values of measurement which contain typical features and shorter variability denoting a characteristic of speech is done. The conversion methods are parametric and non-parametric models.

International Current Journal of Engineering and Science (ICJES)

Volume 1, Issue 1, pp 1-5, December-2022

Available Online at: www.icjes.com

Humans' production mechanism defines a speech production model in parametric models; it generally uses linear predictive coding speech analysis for this aim. The non-parametric method uses MFCC. It is based on the mechanism of perception of human speech. in this research, other essential voice features such as mean frequency, mode frequency, first quartile, third quartile are also extracted using Fast Fourier Transform, and these extracted voice features combined with MFCC are used as feature vector to train GMM model to identify the gender. The extracted features using FFT is given in the below figure 1.

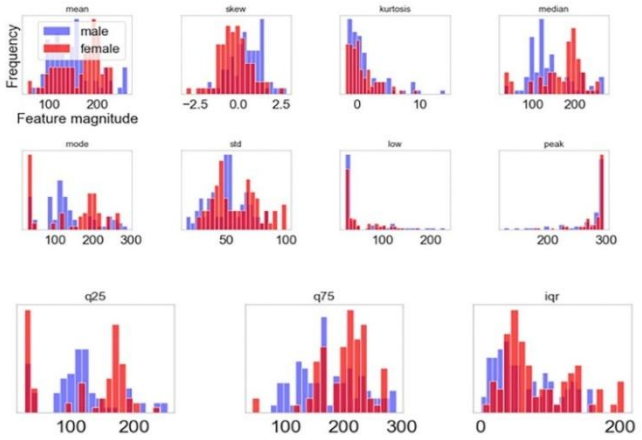

Figure 1: Acoustic Properties with Distribution

A) Mel-frequency cepstral coefficients:

Mel-frequency cepstral coefficients is considered to be a commonly applied feature in speech as well as in recognition of the speaker [9,10]. In an experimental study, volkman and steven found out that the system of human hearing senses frequencies in range 1 KHherz and above algorithmically. Below, the relationship between me frequency and actual frequency is given in Equ 1.

$$mel(f) = 2595 \times \log_{10}\left(1 + \frac{f}{700}\right) \qquad (1)$$

In Mel-frequency cepstral coefficients speech is converted into parameters. Block diagram for MFCC is given below.

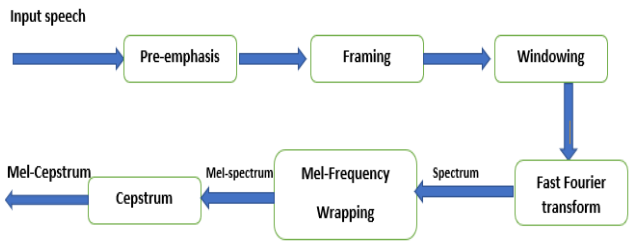

Figure 2: MFCC

B) Pre-emphasis

Compensating of the higher frequency being suppressed in producing mechanism of human sound is the importance of pre-emphasis. Digital filter having a particular function transfer is applied for pre-emphasizing speech. Here arithmetical Equation for high pass filter is given in Equ 2.

$$y(n) = x(n) - q \times x(n-1)$$

$$Y[n] = x[n] - a * x[n-1],\ a \approx (0.95 - 0.97) \quad (2)$$

In this research we have taken a=0.97

C) Framing and Windowing

In framing, the signal of speech of input is split up into frames having a duration lesser than the duration of the window, and as a signal of speech is not fixed .so signal properties changes swiftly, it makes it impossible for Fourier-transform to apply for the signal of speech. The MFCC approach, like all speech analysis methods, is used on short segments where the voice exhibits stationary acoustic properties. These segments are usually 20-30 milliseconds long, with a 10-15 millisecond shift along with the signal. As a result, each frame incorporates a bit of the preceding frame. In general, a hamming window is desirable. The Hamming window is written as,

$$w[n] = 0.54 - 0.46 * \cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \quad (3)$$

D) Frequency Spectrum

FFT transforms speech signals from the domain of time to the domain of frequency once they have been separated into analysis windows. The amplitude spectrum is a notation that represents the frequency distribution of a voice signal.

E) Mel-warping

A set of filters arranged with regard to mel scale linearly is applied to transform the resulting amplitude of spectrum into mel scale. This set comprises 50 percent overlapping triangular band pass filters. Between 20 and 30 is chosen as the filter coefficient.

F) Mel Spectrum and Cepstrum

At this point, the mel spectrum is calculated by multiplying the amplitude spectrum of the signals by the applied honey filters and calculating the logarithm of the power in each filter. The DCT in Eq (4) is used back to the

International Current Journal of Engineering and Science (ICJES)

Volume 1, Issue 1, pp 1-5, December-2022

Available Online at: www.icjes.com

time domain when the logarithmic coefficients of the honey spectrum are real integrals.

The generated coefficients are referred to as a consequence (MFCC).

$$\widetilde{C_n} = \sum_{k=1}^{K} \log(\widetilde{S_k}) \cos(n(k-\frac{1}{2})\frac{\pi}{K}, n = 0,1,\dots,K,\ k = 0,1,\dots,K \quad (4)$$

Mel spectrum coefficients are represented by, where k=1,2,...,K. The average logarithmic energy is retrieved from the feature vector since the first component acquired as a significance of conversion represents it.

III. GAUSSIAN MIXTURE MODEL

The GMM modeling's fundamental aspect is finding a set that collects the mean vector and mixture's weight from each accent's training speech utterance. Samples of training are taken from the American English corpus by Surfingtech, having utterances from ten speakers. Using training samples, different feature vectors could be extracted to use it to identify gender here MFCC feature is used for each test sepal signal. Then GMM testing can be applied for every speech signal and calculating the maximum likelihood score and then comparing both scores, and finally predicting a speaker's gender based on maximum likelihood score. Here we have used the maximum likelihood method to estimate the parameters of the Gaussian mixture model.

The maximum likelihood is calculated using below equation

$$\boldsymbol{P(X|\lambda)} = \sum_{\boldsymbol{k=1}}^{K} W_k\, P_k(X|\mu_k, \sum_k) \quad (5)$$

Where $P_k(X|\mu_k, \sum_k) = \frac{\mathbf{1}}{\sqrt{\mathbf{2\pi}|\sum_k|}} e^{\frac{1}{2}(X-\mu_k)^T \Sigma^{-1}(X-\mu_k)}$ is distribution of Gaussian.

$\boldsymbol{\lambda}$=training data
μ is the mean
$\sum$=co variances of matrices
W_k=wights
K=index of GMM components

The recognition of gender is done in 3 steps. First, we have to retrieve the essential voice features, then computing the likelihood of belonging to a particular gender, and lastly making the comparison of both scores and deciding the gender of a speaker.

3.1 Parameter Estimation

There are several approaches for estimating appropriate Gaussian Mixture Model parameters for feature's vectors distribution derived from speech. Maximum likelihood (ML) estimate is the most well-known and well-established of these approaches. From the available training data, the goal of Maximum likelihood estimation is to discover the parameters which maximize the Gaussian Mixture likelihood.

$$\boldsymbol{P(X|\lambda)} = \prod_{t=1}^{T} p(\overrightarrow{x_t}|\boldsymbol{\lambda}) \quad (6)$$

This equation is a nonlinear function of the λ parameter, and it is difficult to maximize it directly. As a result, parameter estimation of Maximum likelihood is accomplished by an iterative aspect process known as expectation maximization (EM).

3.2 EM Algorithm

The Expectation-Maximization algorithm is a generic approach for estimating ML parameters from data sets that are incomplete or have some missing data. The EM algorithm has two fundamental uses. The first arises when there are actually some missing values. Missing values might be the result of a malfunction with the observation process or constraints. The second situation comes when solving the likelihood function analytically is challenging, but we can simplify it by supposing that missing or concealed parameter exists. In pattern recognition, the letters are more extensively utilised.

Assume that X is not a complete set of data and z= (z1,...,zk) is a random variable which exists, with K-dimensional indicating the identification of the mixing components that create X. The mixing model is written as.

$$P(\vec{x}|\lambda) = \sum_{k=1}^{K} a_k p_k(\vec{x}|z_k, \lambda) \quad (7)$$

Where $p_k(\vec{x}|z_k, \lambda)$ are densities of components described by λ each, λ is a pointer variable with only one and the rest zero, and a_k=p(z_k) are the mixture weights whose total $\sum_{k=1}^{K} a_k$ is 1.It denotes the likelihood of any picked $\vec{x}$vector randomly created by the k[th] components.

The EM method begins with an initial estimate of λ and updates λ it till convergence is achieved.

Random or heuristic approaches can be used to select initial parameters and weights. There are Expectation and Maximization steps in each repetition.

Expectation -Step: The probability w_{ik} of each data point to $\overrightarrow{x_i}$ $1 \le i \le N$belonging to$1 \le k \le K$ is calculated using the current λ values. The Eq .8 mixing component (10). The total of mixing weights is equal to $\sum_{k=1}^{K}$ wik = 1 - for each data

International Current Journal of Engineering and Science (ICJES)

Volume 1, Issue 1, pp 1-5, December-2022

Available Online at: www.icjes.com

point; $\overrightarrow{x_i}$. in this stage, a NxK mixed weight Matrices is created, with sum of every row equal to 1.

$$w_{ik} = p(z_{ik} = 1|\overrightarrow{x_i}, \lambda) = \frac{p_k(\overrightarrow{x_i}|z_k, \lambda_k).a_k}{\sum_{m=1}^{K} p_m(\overrightarrow{x_i}|z_m, \lambda_m).a_m} \quad (8)$$

Maximization-Step: To find and make calculation of new parameters values, first compute the total of mixture weights $N_k=\sum_{k=1}^{K}$ wik for all components. This is the sum of data points which are actually used. That belong to the k^{th} component It is possible to create a new mixture weight, mean, and covariance matrix. Equations (9), (10), and (11) were used to calculate the results.

$$a_k{}^{new} = \frac{N_k}{N} \qquad 1 \le k \le K \quad (9)$$

$$\overrightarrow{\mu_k}^{new} = \left(\frac{1}{N_k}\right) \sum_{i=1}^{N} w_{ik}.\overrightarrow{x_i} \qquad 1 \le k \le K \quad (10)$$

$$\Sigma_k{}^{new} = \left(\frac{1}{N_k}\right) \sum_{i=1}^{N} w_{ik}.\left(\overrightarrow{x_i} - \overrightarrow{\mu_k}^{new}\right).\left(\overrightarrow{x_i} - \overrightarrow{\mu_k}^{new}\right)^T \quad (11)$$

After all of the new parameters have been calculated, repeat the process. Step Expectation is returned, and member weights are computed and updated. The parameterization procedure is still in progress. A pair of Expectation and Maximization steps is required. As though it were a single iteration each cycle concludes with an equation (12) defines the logarithmic likelihood value as calculated, and if there were no significant changes noticed at the end Convergence is achieved after a certain number of iterations.

$$logl(\lambda) = \sum_{i=1}^{N} logp(\overrightarrow{x_i}|\lambda)$$

$$= \sum_{i=1}^{N} (log \sum_{k=1}^{K} a_k p_k(\overrightarrow{x_i}|z_k, \lambda_k)) \quad (12)$$

The Gaussian distribution for the k^{th} mixed component is $p_k(\overrightarrow{x_i}|z_k, \lambda_k)$

IV. EXPERIMENTAL RESULTS

In this research, automated identification of the gender of a speaker is proposed. There are two main stages. Firstly, the system is entirely trained by using sentences of a known gender speaker. Secondly, the testing is done by using the vocalized sentences of the unknown speaker. We have used a free dataset American English corpus by Surfingtech, having utterances from a considered set of ten speakers (five females and five males) in the Training stage of gender identification of speaker and two Gaussian mixture model models are built, one for male and another female feature vector are MFCC coefficients and other acoustic voice properties extracted using FFT. Five different machine learning algorithms (SVM, Gradient boosting, Neural Network, Decision tree, Random Forest) for classification of gender the obtained result is given in table 2.

Table 1: Confusion matrix for GMM model

	Gender expected being female	Gender expected being Male
Gender guessed being Female	563	28
Gender guessed being Male	21	346

By considering above table, we can calculate the accuracy of our system

- Precision for of gender being female = 563 / (563 + 28) = 0.95%
- Precision for gender being = 376 / (376 + 21) = 0.94%
- Accuracy of identification system = 939 / 988 = 0.95%

Table 2: Result for classification of gender

Model	Training dataset	Testing dataset
Gradient boosting	95.8	93.7
Neural Network	94.1	93.7
SVM	92.7	92.1
Decision Tree	1.000	89.8
Random Forest	98.9	93.3

V. CONCLUSION

In this research work, the type of speaker was determined by generating the MFCC coefficient and other acoustic properties of the sound from the speaker formulation using a Gaussian mixture model. Five different machine learning classification algorithms are used for the classification of gender. GMM and Gradient boosting algorithm give excellent results. For feature scope, the performance can be improved and may provide better result by making use of normalization of GMM UBM-GMM.

REFERENCES

[1] Alex Acero and Xuedong Huang, "Speaker and Gender Normalization for Continuous-Density Hidden Markov Models", *in Proc. of the Int. Conf. on Acoustics, Speech, and Signal, IEEE,* May 1996.

[2] C. Neti and Salim Roukos, "Phone-specific gender dependent models for continuous speech recognition", *Automatic Speech Recognition and Understanding Workshop (ASRU97), Santa Barbara, CA,* 1997.

[3] R. Vergin, A. Farhat and D. O'Shaughnessy, "Robust gender-dependent acoustic-phonetic modeling in continuous speech recognition based on a new automatic male/female classification", *Proc. of IEEE Int. Conf. on Spoken Language (ICSLP)*, pp. 1081-1084, Oct. 1996.

[4] S. Slomka and S. Sridharan, "Automatic gender identification optimized for language independence", *Proc. of IEEE TENCON'97,* pp. 145-148, Dec. 1997.

[5] E.S. Parris and M.J. Carey, "Language Independent Gender Identification", *ICASSP*, pp 685-688, 1996.

[6] Ting, H, Yingchun, Zhaohui, W., "Combining MFCC and Pitch to Enhance the Performance of the Gender Recognition", *IEEE*, 2006.

[7] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech and Audio Process.*, 3 (1), 72–83, Jan. 1995.

[8] M. H. Sedaaghi, "A Comparative Study of Gender and Age Classification in Speech Signals", *Iranian Journal of Electrical & Electronic Engineering*, Vol. 5, No. 1, pp. 1- 12, March 2009.

[9] L. Rabiner and B.-H. Juang, "Fundamentals of Speech Recognition, Englewood Cliffs (N.J.)", *Prentice Hall Signal Processing Series*, 1993.

[10] J. R. Deller, J. H. L. Hansen, J. G. Proakis, "Discrete-Time Processing of Speech Signals", *IEEE Press, Piscataway (N.J.)*, 2000.

***** End of the Article *****

