# Speech Recognition to Distinguish Gender and A Review and Related Terms

Jagannath Jadhav[#1]
[#] M Tech Student,ECE,MITE,Modabidri
MITE,Modabidri, Karnataka, India

Sahana Devali[*2]
*Assistant Professor,ECE,MITE,Moodabidri
MITE,Modabidri, Karnataka, India

***Abstract*: Speech recognition is a process where a speech file is recognized against the stored speech data set. Speech recognition is the ultimate goal concerned with science, technology, and engineering of discovering patterns and extracting potentially useful or interesting information automatically or semi-automatically from speech data. It analyzes the data set according to the classifier and predicts results accordingly. In this scenario, a predicted output is one which matches the most with the data base. Several kinds of classifier have been used in this scenario. This paper represents different sections of the speech recognition process and classification methods are also discussed. This paper aimed to design an efficient system for extracting useful information from speech signal and use the same information to design the classifier which is capable of differentiating male and female voices.**

***Keywords: Speech signal processing, speech recognition, feature extraction, classifiers, recognition process.***

## I. INTRODUCTION

Information mining from speech signal or Speech recognition can be defined as an activity that extracts some new nontrivial information contained in large databases. The goal is to discover hidden patterns, unexpected trends or other subtle relationships in the data using a combination of techniques from machine learning, statistics and database technologies. This new discipline today finds application in a wide and diverse range of business, scientific and engineering scenarios. . A typical speech can be broadly defined as speech with emotional content, speech affected by alcohol and drugs, speech from speakers with disabilities, and various kind of pathological speech. Individual speech can vary because of different timing and amplitude of the movement of speech articulators. Physical mechanism of speech undergoes changes, which can affect the nasal cavity resonance and mode of vibrations of the vocal cords. A series of environmental variables like background noise, reverberation and recording condition have also to be taken into account. In essence every speech production is unique and this uniqueness makes the automatic speech processing quite difficult. Speech recognition project aimed to design an efficient system for extracting useful information from speech signal and use the same information to design the classifier which is capable of differentiating male and female voices.

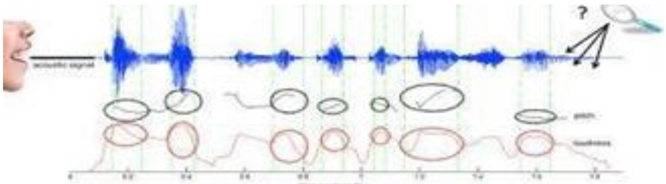

Fig. 1. speech signal.

The dynamic requirements of automated systems have pushed the extent of recognition system to consider the precise way of command rather to run only on command templates. The idea correlates itself with the speaker identification at the same time recognizing the gender of speaker.This can also be used in the spoken dialogue system e.g. at call centre applications where the support staff can handle the conversation in a more adjusting manner. At the other hand the signal processing tools like MATLAB and pattern recognition researcher's community developed the variety of algorithms (e.g. ANN, SVM) which completes needed resources to achieve the goal of recognizing gender from speech.

## II. GENERAL PRINCIPLES OF SPEECH SIGNAL PROCESSING

The whole processing block chain common to all approaches to speech processing is shown in figure 5. The first step in the processing is the speech pre-processing, which provides signal operation such as digitalization, pre-emphasis, frame blocking, and windowing. Digitalization of analog speech signal starts the whole processing. The microphone and the analog to digital (A/D) converter usually introduce undesired side effects. Because of the limited frequency response of analog telecommunications channels and the wide spread use of 8 kHz sampled speech in digital telephony, the most popular sample frequency for the speech signal in telecommunication is 8 kHz.

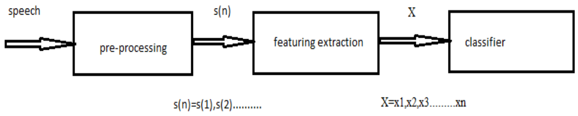

Fig. 2. block diagram of speech signal processing.

In non-telecommunications applications, sample frequencies of 12 and 16 kHz are used. The second step that is feature extraction, represents the process of converting sequences of pre-processed speech samples *s(n)* to observation vector **x** representing characteristics of the time varying speech signal. The kind of features extracted from speech signal and put together into feature vector **x** corresponds to the final aim of the speech processing. For each application, the most efficient features that is the features carrying best the mining information, should be used. The first two blocks represent straightforward problems in digital signal processing. The subsequent classification is then optimized to the final expected information.

### A. *Pre-Emphasis:*

The characteristics of the vocal tract define the current uttered phoneme. Such characteristics are evidenced in the frequency spectrum by the location of the formants that is local peaks given by resonances of the vocal tract. Although possessing relevant information, high frequency formants have smaller amplitude with respect to low frequency formants. To spectrally flatten the speech signal, a filtering is required. Usually, a one coefficient FIR filter ,known as a pre-emphasis filter, with transfer function in z-domain as shown in the " (1)," is used.

$$H(z) = 1 - \lambda z^{-1} \qquad (1)$$

In the time domain, the pre-emphasized signal is related to the input signal by the difference equation "(2).

$$s(n) = s(n) - \lambda\, s(n-1) \qquad (2)$$

A typical range of values for the pre-emphasis coefficient is λ belongs to 0.9 – 1. One possibility is to choose an adaptive pre-emphasis, in which λ changes with time according to the relation between the first two values of autocorrelation coefficients " (3)".

$$\lambda = R(1) / R(0) \qquad (3)$$

### B. *Frame Blocking:*

The most common approaches in speech signal processing are based on short-time analysis. The pre-emphasized signal is blocked into frames of *N* samples. Frame duration typically ranges between 10-30nmsec. Values in this range represent a trade-off between the rate of change of spectrum and system complexity. The proper frame duration is ultimately dependent on the velocity of the articulators in the speech production system.

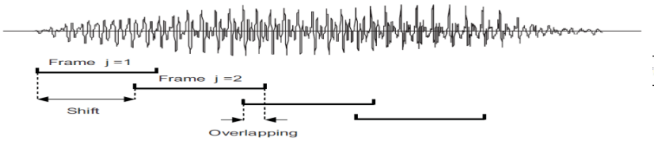

Fig. 3. frame blocking

### C. *Windowing*

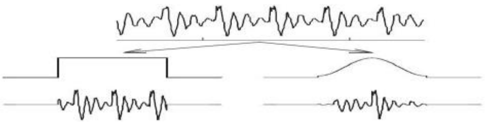

Fig. 4. Rectangle window(left) and hamming window(left)

A signal observed for a finite interval of time may have distorted spectral information in the Fourier transform due to the ringing of the sin*(f)/f* spectral peaks of the rectangular window. To avoid or minimize this distortion, a signal is multiplied by a window-weighing function before parameter extraction is performed. Window choice is crucial for separation of spectral components which are near one another in frequency or where one component is much smaller than another. In speech processing, the Hamming window is almost exclusively used. The hamming window is a specific case of the Hanning window.

### D. *Database*:

A data base is the collection of data .In our proposed work we have used speech samples for the database. In the database we find properties of the speech signals and then we store them into the database. The question comes that how we are going to store hundreds of files in the database. The procedure would be as follows. First of all we would fetch the properties of the voice samples. All those properties which are required would be computed and then it would be stored into an array. The array would move on as the files would move. We would fetch the features and would take the average by the end and then store them into the database for each category of the voice.

### E. *Voice Files*:

The voice files are the files which would be processed for the feature extraction.

### F. *Properties*:

When we would process the voice files their properties would be fetched .For the feature extraction there are several algorithms which can be used.

## III. SECTIONS OF THE RESEARCH WORK

There are several sections in our research work. The sections are explained as follows.

### A. *Training*:

The training section ensures that the database gets trained properly so that at the time of testing it produces extensive results. The features of the training are as follows.

### 1. *Linear Predictive Coding:*

Linear predictive coding (LPC) is a tool used mostly in audio signal processing and speech processing for representing the spectral envelope of a digital signal of speech in compressed form, using the information of a linear predictive model.It is one of the most powerful speech analysis techniques, and one of the most useful methods for encoding good quality speech at a low bit rate and provides extremely accurate estimates of speech parameters. LPC starts with the assumption that a speech signal is produced by a buzzer at the end of a tube, with occasional added hissing and popping sounds. Although apparently crude, this model is actually a close approximation of the reality of speech production. The glottis (the space between the vocal folds) produces the buzz, which is characterized by its intensity (loudness) and frequency (pitch). The vocal tract (the throat and mouth) forms the tube, which is characterized by its resonances, which give rise to formants, or enhanced frequency bands in the sound produced. Hisses and pops are generated by the action of the tongue, lips and throat during sibilants and plosives. LPC analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal after the subtraction of the filtered modeled signal is called the residue. The numbers which describe the intensity and frequency of the buzz, the formants, and the residue signal, can be stored or transmitted somewhere else.

### 2. *Cepstral Analysis*

A cepstrum is the result of taking the Inverse Fourier transform (IFT) of the logarithm of the estimated spectrum of a signal. There is a complex cepstrum, a real cepstrum, a power cepstrum, and phase cepstrum. The power cepstrum in particular finds applications in the analysis of human speech. The name "cepstrum" was derived by reversing the first four letters of "spectrum". Operations on cepstral are labeled frequency analysis, filtering, or cepstral analysis. The power cepstrum of a signal is defined as the squared magnitude of the inverse Fourier transform of the logarithm of the squared magnitude of the Fourier transform of a signal.

$$\text{Power capstrum} = \left|F^{-1}\left\{(\log\left|F(f(t)\right|^{2})\right\}\right|^{2} \qquad (4)$$

The real cepstrum is related to the power via the relationship (4 * real cepstrum) ^2 =power cepstrum, and is related to the complex cepstrum as real cepstrum = 0.5*(complex cepstrum + time reversal of complex cepstrum).. The complex cepstrum holds information about magnitude and phase of the initial spectrum, allowing the reconstruction of the signal. The real cepstrum uses only the information of the magnitude of the spectrum.

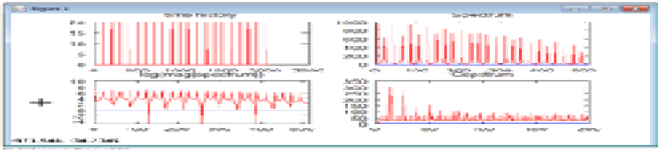

Fig. 5. Cepstrum signal analysis.

### 3. *PARCOR Coefficients*

In order to go from a predictor of order i to one of order i+1 is necessary to re-compute all the prediction coefficients the partial correlation coefficients (PARCOR) allow to obtain coefficients of order i from those of order i-1

$$a_j = a_j^{(i-1)} - K_i a_{i-j}^{(i-1)} \qquad (5)$$

Sometimes, it can turn out more convenient to work with the PARCOR. We passed from transversal structures to lattice structures

$$K_i = \frac{\sum_n e^{(i-1)}(n) b^{(i-1)}(n)}{\sqrt{\sum_n (e^{(i-1)}(n))^2 \sum_n (b^{(i-1)}(n))^2}} \qquad (6)$$

Where $K_i$ is the reflaction cofficient of the acostic tube model.

### 4. *Log Area Ratio*

Log area ratios (LAR) can be used to represent reflection coefficients (another form for linear prediction coefficients) for transmission over a channel. While not as efficient as line spectral pairs (LSPs), log area ratios are much simpler to compute. Let $\gamma_k$ be the *k*th reflection coefficient of a filter, the *k*th LAR is:

$$A_k = log\frac{1+\gamma_k}{1-\gamma_k} \qquad (7)$$

Where $A_k$ is kth LAR
$\gamma_k$ is be the *k*th reflection coefficient of a filter.

### 5. *Standard Deviation*

standard deviation (represented by the symbol sigma) shows how much variation or dispersion exists from the average (mean), or expected value. A low standard deviation indicates that the data points tend to be very close to the mean; high standard deviation indicates that the data points are spread out over a large range of values.

### 6. *Pitch*

It is the main feature of an audio file. Sounds may be generally characterized by pitch, loudness, and quality. The perceived pitch of a sound is just the ear's response to frequency, i.e., for most practical purposes the pitch is just the frequency. Pitch is equal to frequency of sound.

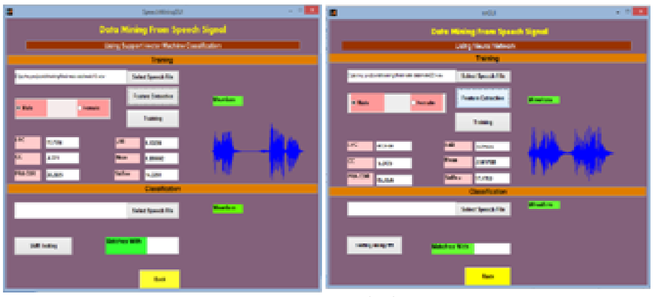
Data Mining From Speech Signal
Training
Classification
Data Mining From Speech Signal
Training
Classification

Fig. 6. Snap shot of features extracted.

### B. *Algorithm Helpful in the Training, Feature Extraction , Testing, Matching and Classification:*

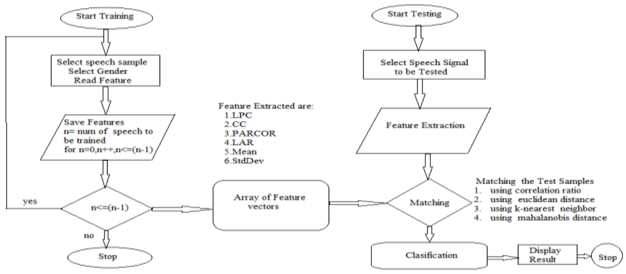

Fig. 8. Flow diagram.

The below algorithm in figure.8 is used for the training of the data set. It extracts the features of the voice sample and saves them to the database for the future use. The maximum frequency of a file is the value which we get at the peak on a frequency map. When so ever we put a voice sample over the time and frequency pattern, the maximum peak is called the maximum frequency of the voice sample. It is viewed as the counter part of the training and it is used to sample size the data for the further processing. In this approach we take each sample of data set as a unique item which has to be processed. The extraction of the feature and saving it to the data base can be classified with the following flow diagram.

### C. *Matching the Test Sample of Speech Signal*

#### 1. *Match Using Correlation Ratio:*

The correlation ratio is a measure of the relationship between the statistical dispersion within individual categories and the dispersion across the whole population or sample. The measure is defined as the *ratio* of two standard deviations representing there types of variation.

#### 2. *Match Using Euclidean Distance:*

Euclidean distance or Euclidean metric is the "ordinary" distance between two points that one would measure with a ruler, and is given by the Pythagorean formula. By using this formula as distance, Euclidean space (or even any inner product space) becomes a metric space. The associated norm is called the Euclidean norm. The formula used to calculate the Euclidean distance can be defined as following:
The Euclidean distance between two points P = (p1, p2…pn) and Q = (q1, q2...qn),

$$D = \sqrt{(p1-q1)^2+(p2-q2)^2+(pn-qn)^2} \qquad (8)$$

$$D = \sqrt{\sum_{i=1}^{n}(pi-qi)^{\boldsymbol{n}}} \qquad (9)$$

The speaker with the lowest distortion distance is chosen to be identified as the unknown person**.**

#### 3. *Match Using k-Nearest Neighbour:*

In pattern recognition, the *k*-nearest neighbor algorithm (*k*-NN) is a non-parametric method for classifying objects based on closest training examples in the feature space. *k*-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification.main object is classified by a majority vote of its neighbours, with the object being assigned to the class most common amongst its *k* nearestneighbours (*k* is a positive integer, typically small).

#### 4. *Match Using Mahalanobis Distance:*

The Mahalanobis distance is a descriptive statistic that provides a relative measure of a data point's distance (residual) from a common point. It is a unit less measure introduced by P. C .Mahalanobis in 1936. The Mahalanobis distance is used to identify and gauge *similarity* of an unknown sample set to a known one Mahalanobis distance (or "generalized squared inter point distance" for its squared value) can be defined as a dissimilarity measure between two random vectors x and y of the same distribution with the covariance matrix *S*:

$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\ \vec{y})^T s^{-1}(\vec{x}-\vec{y})} \qquad (10)$$

### *D. Performance Parameters*

Accuracy measuring the performance of an automatic speech recognition system which are described below:

#### 1. *TP Rate:*

The True Positive (TP) rate is the proportion of examples which were classified as class x, among all examples which truly have class x, i.e. how much part of the class was captured. It is equivalent to Recall.

#### 2. *FP Rate:*

The False Positive (FP) rate is the proportion of examples which were classified as class x, but belong to a different class, among all examples which are not of class x. TP Rate and FP Rate are calculated in accordance with mean and variance. For calculating TP Rate & FP Rate, Gain is calculated by processing all the rows of uploaded file. Value of TP is equal to gain calculated and value of FP is equal to value of gain subtracted from the whole database.

#### 3. *Receiver Operating Characteristic (ROC):*

ROC is used as a performance parameter for comparing various classification algorithms. ROC Curve is also called threshold curve. ROC Area of different classification algorithms is compared on the basis of ROC Area values using 10-Fold cross validation.

## IV. TESTING METHOD

The testing module of the speech processing includes the testing of the speech file on the basis of the trained data set. To perform a testing operation over the speech files different types of classifiers are used to analyze the services of the speech samples. Some of the classifiers are explained as follows .

### 1. *SVM (SUPP0RT VECTOR MACHINE)* **:**

SVM stands for support vector machine . It takes the entire data set as the binary input and classifiers for the same . The SVM classifier generates the FAR and FRR ratio successfully to determine the matching percentage . SVMs are linear classifiers (i.e. the classes are separated by hyper planes) but they can be used for non-linear classification by the so-called *kernel trick*. Instead of applying the SVM directly to the input space Run they are applied to a higher dimensional *feature space* F, which is nonlinearly related to the input space: _ : Run ! F. The kernel trick can be used since the algorithms of the SVM use the training vectors only in the form of Euclidean dot-products (x _ y). It is then only necessary to calculate the dot-product in feature space (_(x) __(y)), which is equal to the so-called *kernel function* k(x; y) if k(x; y) fulfils the Mercer's condition. Important kernel functions which fulfil these conditions are the polynomial kernel.

### 2. *NN(NEURAL NETWORK CLASSIFIERS) :*

The neural network classifier is one of the most advance classifiers which takes two inputs. the first input is the training set and the second input is the target set . The target is drawn on the basis of which the training set has been updated .[6] Neural nets are highly interconnected networks of relatively simple processing elements, or nodes, that operate in parallel. They are designed to mimic the function of neurobiological networks. Recent work on neural networks raises the possibility of new approaches to the speech recognition problem. Neural nets offer two potential advantages over existing approaches. First, their use of many processors operating in parallel may provide the computational power required for continuous-speech recognition. Second, new neural net algorithms, which could self-organize and build an internal speech model that maximizes performance, would perform even better than existing algorithms.These new algorithms could mimic the type of learning used by a child who is mastering new words and phrases.

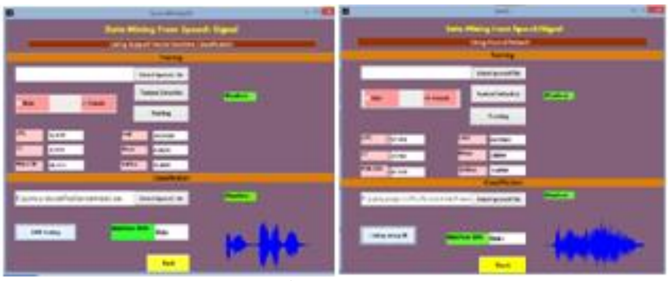

Fig. 9. Snap Shot of SVM and NN Testing phase.

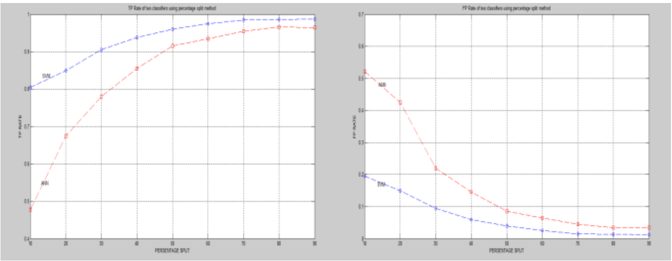

Fig.10. TP Rate and FP Rate of two classifiers using percentage split method

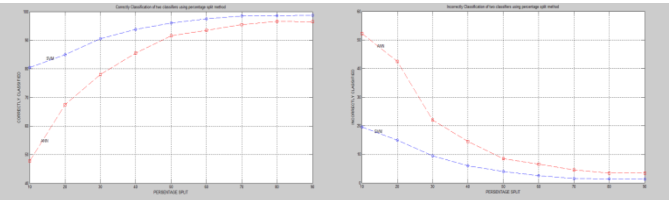

Fig.11. Correctly Classified and Incorrectly Classified Vs Percentage Split

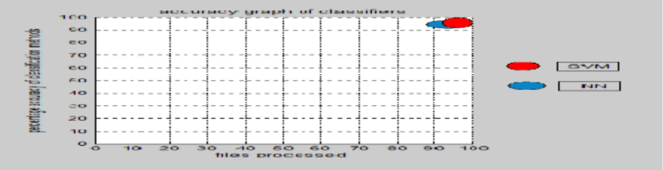

Fig. 12. percentage of accuracy.

## V. CONCLUSION

With the above text, it can be concluded that the speech recognition system is a process which requires two phases of data. The first phase is the training phase and the second phase is the testing phase. We have studied on speech recognition by means of ANN and SVM, and we believe that ANN makes significant impact on speech recognition as ANN proves to be a better training algorithm and SVM as better classification algorithm. We expect the accuracy to be increased in comparison with the ANN. SVM is expected to work in better manner because the training set created with the help of Matching techniques. Our result shows that our proposed SVM classifier gives accuracy of 98.5% whereas NN classifier gives an accuracy of 96.2%.

## REFERENCES

[1] Ekta Garg, Madhu Bahl,"Speech Emotion Recognition A Review and Related Terms" International Journal of Innovative Research in Computer and Communication Engineering*(An ISO 3297: 2007 CertifiedOrganization)* Vol. 2, Issu6, June 2014.

[2] Surendra Shetty, K.K Achar, "Audio Data Mining Using Multi-perceptron ANN" International Journal of Computer Science and Network Security, Vol. 8, No.10 October 2008.

[3] M.A. Anusuya, S.K. Katti, "Speech Recognition by Machine: A Review", International Journal of Computer Science and Information Security,Vol. 6, No.3,2009.

[4] Ying Shi; Weihua Song, "Speech emotion recognition based on data mining technology", Sixth International Conference on Natural Computation-Aug 2010.

[5] Dai Sheng-Hui, Lin Gang-Yong, Zhou Hua-Qing , "A Novel Method for Speech Data Mining", Journal of Software, Vol. 6, No. 1, January 2011.

[6] K.A.Senthildevi, Dr.E.Chandra," Data Mining Techniques and Applications in Speech Processing - A Review" iJARS/ Vol.I / Issue II /Sept-Nov, 2012/191 .

[7] Burhardt, F.; Paeschke, A.;Rolfes, M; sendlmeier,W. &Weiss, B. (2005). A Database of German emotional speech, *Proceedings of Eurospeech'05*, pp. 1517-1520, ISSN 1018-4074, Lisbon, September 2005, International Speech Communication Association, Grenoble.