VOICE SIGNAL SYNTHESIS USING NON NEGATIVE MATRIX FACTORIZATION

Chaitra Burshi [1], Dr. Bhuvaneshwari C. Melinamath [2]

[1]*Student at BLDEA's VP Dr. PG Halakatti College of Engineering and Technology, Vijaypur*
[2] *Professor at Sanjay Ghodawat Institution.*

---***---

ABSTRACT - *Speech recognition has various applications like, medical field, military field, education field, criminal field, telephonic communication, etc. Speech signal processing has wide range of applications in criminal case may be in the form of telephonic communication. Therefore, work proposes a system that is used to recognize person's gender and emotion state for the speech signals. System is able to recognize two gender states (male or female) and three emotion states (angry, happy and sad). The system is implemented using the Non negative matrix factorization and Mel Frequency Cepstrum. The wave surfer tool is used for recording speech signals. The results regarding gender recognition are encouraging, the model is tested on 100 data samples and for emotion recognition the model is tested on 60 data samples. The accuracy obtained from this system is 87%*.

1. INTRODUCTION

Speech is the communication between the two people. Speech not only carries the information to communicate but also contains the information regarding the particular speaker. Features like identifying Gender (male or female), emotion state (angry, sad, happy) etc. The special features like social factors, affective factors and the properties of the physical voice production apparatus for which human being are able to recognize whether the speaker is a male or a female easily, during telephone conversation or any hidden condition of the speaker.

Speaker identification is used in the many applications like in medical field, kidnapping, education, military etc. But this is mainly used in the criminal detection methods. Consider for example from a speech signal we get the information about the gender, emotion, age, etc.

So, the speaker identification is used in many cases such as kidnapping, false calls, etc. So if we do the reorganization process, identifying the unknown person becomes easy.

In this paper, we performed a gender and emotion detection approach. To develop this method, we gone through two phases they are, training and testing. In training phase, first we read the recorded speech, those speech are break into frames, generally similar speech is taken as a single frame. So, these are broken down into respective frames spaced M samples apart. These single frames are next taken for windowing, where the smoothing of the signal is done, then for each frame we find MFCC coefficient using kmeans. Finally we get trained data. Next phase is the testing phase, in which a single speech signal is taken for the testing. That signal is gone through a process of framing, windowing, MFCC coefficient calculation using k-means like training phase. Finally we compare the speech signal with the trained signals using Euclidean distance, where the smallest calculated distance is taken as an output.

2. RELATED WORK

In [1] author proposed a system for gender recognizing, they used the Mel frequency Coefficients classifiers along with Gaussian Mixture Model for the gender recognition. The system uses the different database for the recognition, in one set they used the database of 760 sentences of 76 speakers and they obtained the 100% success, and in other set they used the database of 6100 sentences of 610 speakers and they obtained the result of 97.76%.

In [2] author has explained about the automatic gender recognition using the Fast Fourier Transform technique, they say that speech is widely used in many technical applications like telephonic communication, in which detecting the gender recognition plays a very important role. The paper explains a comparative investigation of speech signals to produce automatic gender detection by using various features of Fast Fourier Transform technique. The system uses the databases of 100 speakers and obtained the success of 91%.

In [3] author describes about the gender detection and age estimation, based on hybrid architecture of Weighted Supervised Non-Negative Matrix Factorization and General Regression Neural Network. In the experiments the database, for the 222 speaker is used and the experiment deals with 96% success.

In [4] author explains about the different emotions of the people are detected. The voice signals are used to detect the different emotion states. Here Support vector machine technique is used to identify the speaker emotion state. The main advantage of this method is, it made easy way to interact between the human and the computer. The database is recorded in the two ways that is, first by using the 50 different voices of male and 50 different voices of the female.

The accuracy obtained from this method is about the 88%. It is also used to distinguish the single emotion verses other all emotions.

3. PROPOSED SYSTEM

This section deals with the methodology used to develop the recognition of the speech signal. We used two steps to do this process that is training phase and testing phase, which is explained in the following section:

3.1 Training Phase

The first phase is the training phase, first we read the recorded speech, those speech are break into frames, generally similar speech is taken as a single frame. So, these are broken down into respective frames spaced M samples apart. These single frames are next taken for windowing, where the smoothing of the signal is done, then for each frame we find MFCC coefficient using k-means. Finally we get trained data

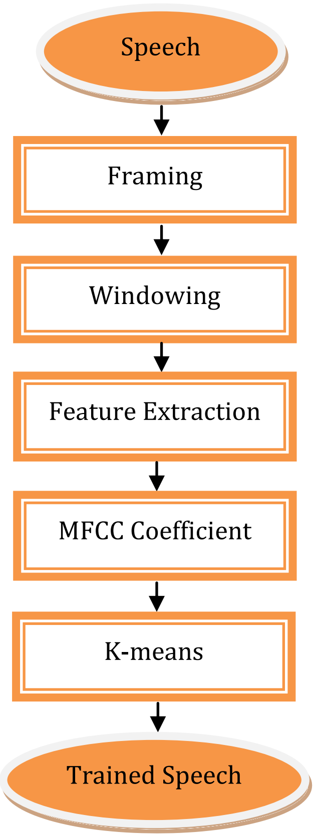

Fig -3.1: Block diagram for training phase

3.1.1 Framing

The recorded speech signal is taken and broken into single frame. These single frames are spaced M samples. These are processed as frame by frame basis. These are passed through filtering. In filtering, signals are partitioned into N sample frames as per the equation shown below:

$$P_s[q] = \begin{cases} p\,[q+sQ], 0<=q<=Q-1, \\ 0, \text{otherwise} \end{cases} \quad \text{... (3.1)}$$

Where p[q] is the input signal, defined for 1<q<1, each frame is indexed by the variable s, with s = 0, 1, 2. . . And the s frame is denoted by sr[q]. Thus, each frame consists of Q samples, with q = 0 to Q - 1 addressing samples inside the frame.

3.1.2 Windowing

Next phase is windowing, this process is carried out to minimize the distortion and to obtain smoother truncation of frames. Where frames are multiplied by Q sampling windows i.e. W (q). To carry out the windowing process, that is to chopping of Q sample sections and to do smoothing process hamming window is used. While breaking the N sample signals, it may give a bad effect on the signal parameters. To carry out the process of minimizing this effect windowing is done using the formula below:

$$S(a)=T(a)*U(a), 0<=a<=A-1 \quad \text{... (3.2)}$$

$$U(a)=(0.54–0.46*Cos(2*pie*a/A)) \quad \text{... (3.3)}$$

Where 0<=a<=A-1

3.1.3 Feature Extraction

Next step is the feature extraction, which deals with reducing the information while taking the information from the speaker. During the speech production, the large amount of data is generated, but only essential characteristics are taken from the speech. So feature extraction is the process of analyzing the speech signals. There are two types in feature extraction that is temporal analysis and spectral analysis technique. In temporal analysis the speech waveform itself is used for analysis. In spectral analysis spectral representation of speech signal is used for analysis.

3.1.4 MFCC

Mel-Frequency Cepstrum is the one of the good technique for extraction of features from the speech signals. Speech is formed by the person's vocal tract. So Mel- Frequency Cepstrum helps in finding what has been said by the human's vocal tract. This is represented using short term power spectrum. This process is carried out using the log power of spectrum. And it is derived from the non linear spectrum of a spectrum representation of the audio clip.

3.1.5 K-means

K-means is the one of the clustering method, which is very well know to solve the problems based on clustering. The main process is to classify the given number of data set into certain number of clusters. These are called as centroid. This centroid is placed far away from each other. So those are taken and find the nearest neighboring using this centroid. We will continues this process till the all values is assigned to all centroids. So from this a matrix to be minimized is calculated.

3.2 Testing Phase

Next phase is the testing phase, in which a single speech signal is taken for the testing. That signal is gone through a process of framing, windowing, MFCC coefficient calculation using k-means like training phase. Finally we compare the speech signal with the trained signals using Euclidean distance, where the smallest calculated distance is taken as an output. This is illustrated as shown in fig 3.2

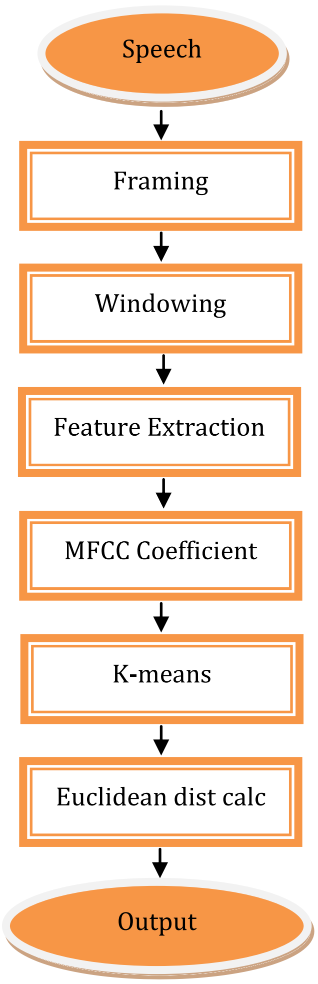

Fig -3.2: Block diagram for testing phase

In this phase, a single speech signal is taken for the testing. That signal is gone through a process of framing, windowing, MFCC coefficient calculation using kmeans as in training phase. Finally we compare the speech signal with the trained signals using Euclidean distance, where the smallest calculated distance is taken as an output. Euclidean distance is calculated as shown in equation 3.4:

The Euclidean distance between two points S= (s1, s2…..sn) and Q= (r1, r2…rn),

$$=\sqrt{(s1-r1)^2 + (s2-r2)^2 + \ldots\ldots\ldots + (sn-rn)^2}$$
$$=\sqrt{\sum_{i=1}^{n}(s_i-r_i)^2} \quad \ldots (3.4)$$

From this the lowest distance is taken as a estimated output.

4. RESULTS

To evaluate this method, the model is tested on 100 data samples for gender recognition and for emotion recognition the model is tested on 60 data samples. Table 4.1 shows the evaluated result for the gender recognition using the different training set. Table 4.2 shows the evaluated result for the emotion recognition for 10 trained data, Table 4.3 shows the evaluated result for the emotion recognition for 20 trained data, and Table 4.3 shows the evaluated result for the emotion recognition for 25 training data.

Table -4.1: Accuracy for gender recognition

Trained data	Tested data	True Positive	False Positive	True Negative	False Negative
10	80	87.5%	15%	85%	12.5%
25	50	92%	12%	88%	8%
40	20	80%	10%	90%	20%

Table -4.2: Accuracy for emotion recognition of 10 training data and 40 testing data

Emotion State	True Positive	False Positive	True Negative	False Negative
Angry	60%	15%	57.5%	40%
Happy	60%	25%	57.5%	40%
Sad	55%	22.5%	60%	45%

Table -4.3: Accuracy for emotion recognition of 20 training data and 30 testing data

Emotion State	True Positive	False Positive	True Negative	False Negative
Angry	60%	10%	35%	40%
Happy	65%	5%	37.5%	35%
Sad	65%	5%	37.5%	35%

Table -4.3: Accuracy for emotion recognition of 25 training data and 15 testing data

Emotion State	True Positive	False Positive	True Negative	False Negative
Angry	80%	5%	17.5%	20%
Happy	80%	5%	17.5%	20%
Sad	85%	2.5%	20%	15%

5. CONCLUSION

In this work, we used Non Negative Matrix Factorization along with MFCC is used to detect the gender and emotion information of the speaker, the process is the testing data is compared with the trained data to get the estimated output. The feature extraction is done by using MFCC and K-means and stored those trained speech in database. Next in testing phase a single speech signal is chosen and compared with the trained data using the Euclidean distance in which the minimum distance is taken as an estimated output. The accuracy obtained from this method is 87%. We can enhance the performance by increasing training data and also using other methods such as Gaussian Mixture Method, Super Vector Method, etc.

REFERENCES

[1] Ergun Yucesoy1, Vasif V.Nabiyev2, "Gender Identification Of A Speech Signal Using GMM and MFCC", Proc. of IEEE TENCON'97, pp. 145-148, Dec. 2014.
[2] M.Prabha, P.Viveka G.Bharatha Sreeja "Advanced Gender Recognition System Using Speech Signal" IJCSET(www.ijcset.net) |April 2016 | Vol 6, Issue 4, 118-120.
[3] S.Sravan Kumar, T.RangaBabu, " Emotion and Gender Recognition of Speech Signals Using SVM" International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 4, Issue 3, May 2015.
[4]Shivaji Chaudhari1, Ramesh Kagalkar " Automatic Speaker Age Classification,Recognition and Identifying Speaker Emotion Using Voice Signal" International Journal of Science and Research (IJSR)2014.
[5] Tobias Bocklet1, Andreas Maier1, Josef G. Bauer2, Felix Burkhardt3, Elmar Noth, "Age and gender recognition for telephone applications based on gmm supervectors and support vector machines", Institute of Pattern Recognition, University of Erlangen-Nuremberg, Germany.
[6] Mohamad Hasan Bahari, Hugo Van hamme, "Speaker Age Estimation and Gender Detection Based on Supervised Non-Negative Matrix Factorization" Centre for Processing Speech and Images Katholieke University Leuven, Belgium, Jan.2010.