Emotion based Speaker Recognition with Vector Quantization

Shraddha Bhandavle, Rasika Inamdar, Aarti Bakshi
KCCEMSR, Thane (E), India

ABSTRACT:

Speech is a most popular biometrics nowadays used for human interaction. An emotion is a mental and a physiological state of a person. Emotion Based Speaker Recognition has attracted many researchers. Emotions are associated with the variety of feelings and thoughts. An emotion based speaker recognition system, recognizes the person's emotionsbased on pitch, speaking style, intensity, sampling frequency. Mel frequency CepstralCoefficient is the first step in a speaker recognition system. In this paper, we are implementing the gender - based modified MFCC approach to differentiate the individuals. For the classification purpose we have used the K-means algorithm.

KEYWORDS:

EmotionRecognition from Speech,Fourier Transform, Traditional MFCC, Modern MFCCApproach,Nearest NeighborAlgorithm, K-means, Vector Quantization.

I. INTRODUCTION

A person's feelings and thoughts are expressed by his/her emotions. Recognizing a person's emotions is a very challenging task. [2]

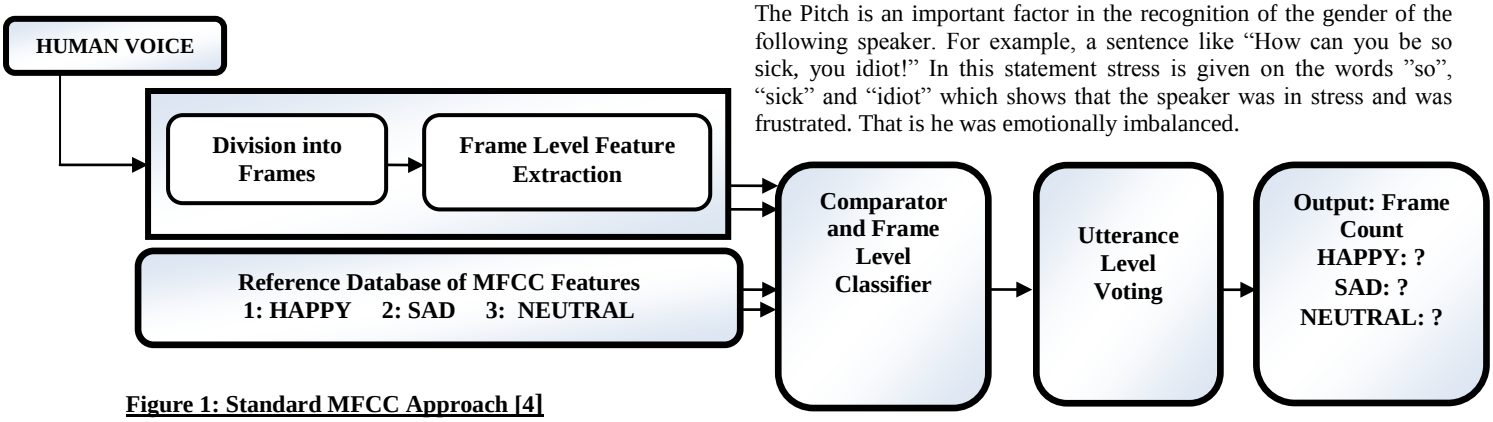


Figure 1: Standard MFCC Approach [4]

These types of systems are also used in telecommunication industries where they are put to use by way of improving the telephone-based speech recognition performance. [5] It is also very widely used in call-centers to build good relationships with the customer by settling their disputes by way of knowing their Section II describes the standard MFCC approach. Section III illustrates the Proposed Modified MFCC Approach. Section IV deals with the Algorithm of Modified MFCC Approach. Section V discusses the Experimental Setup and the Results. Section VI concludes the paper and Section VII focuses on the Future Scope of the paper.

emotional state. [5]

The acoustic features play a major role in identification of emotions of a person. There are two types of Speaker Recognition: Text-Dependent and Text-Independent. In Text-Dependent recognition, the user is authenticated with the help of his spoken phrase. On the other hand, in Text-Independent recognition, the system does not have to remember the phrase spoken given by a user. [5] This paper focuses more on the Text- Independent recognition. The emotional utterance of a person is considered for this purpose. For this, the appropriate acoustic features need to be selected for better results.

Emotions can be characterized using discrete theory and dimensional theories. The various emotions like angry, sad, neutral, happy and fearwork on the principles of discrete theory.[2] The database used in the experiment can be either a natural database where people will record their voice phrases or the database can also be downloaded. Facial expressions also play a vital role in the detection of the emotional state of a person. When facial expressions are clubbed together with the voice utterance of a person, the accuracy rate of the recognition of emotion increases. [4] The centroid values are considered and the probability is calculated to determine the emotions of a person. Apart from spoken words, the voice of a person also gives the information about his age, gender and identity [6].

The Pitch is an important factor in the recognition of the gender of the following speaker. For example, a sentence like "How can you be so sick, you idiot!" In this statement stress is given on the words "so", "sick" and "idiot" which shows that the speaker was in stress and was frustrated. That is he was emotionally imbalanced.

I. STANDARD MFCC APPROACH

A. FrameBlocking

The voice samples are recorded using a Microphone.The voice sample is segmented into frames of 25 milliseconds each.

B. FeatureExtractionfor Recognition Setup

Acoustic features are extracted by applying hamming window, FFT, MEL-filter bank to each speech frame.

C. Comparison and Feature Matching

After the features are extracted, the centroid value for each frame is calculated and it is matched with centroid values of the frames saved in the reference database.
Now these calculated centroids of each frame are clustered together after matching their values and compared with the thresholds set which will help in finding out Happy, Sad or Neutral moods of the Speaker.

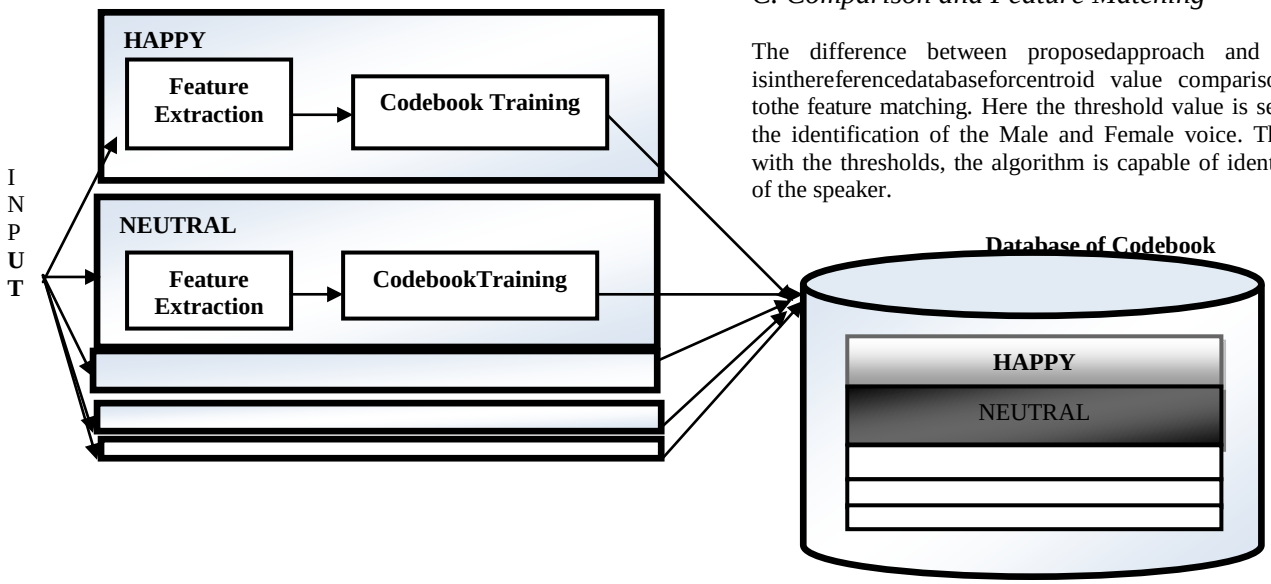


Figure 2: Codebook Training Flowchart [4]

II. PROPOSED MODIFIED MFCC APPROACH

The problem with the standard approach was that it considered the data in a homogenous manner without segregating it into male or female voice. But in the modified approach, we will first separate the voice samples into male samples and female samples using the gender reference databases. This will only help in the increase in the accuracy of detecting the emotions of a person. [4]

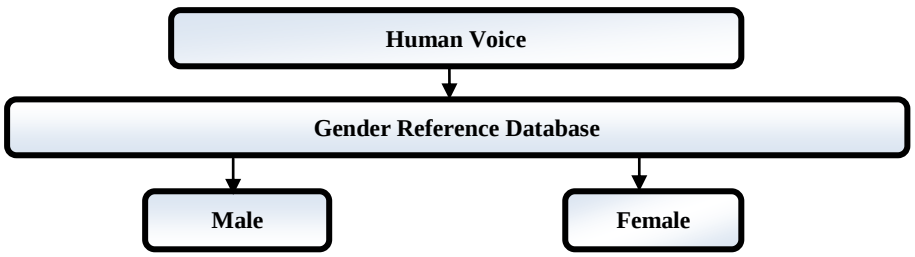
Differencesinproposedalgorithm

Before framing, speech samples are recorded using microphone. These samples are compared with gender based database. For gender recognition, a lower and upper bound of pitch for male and female is taken as feature and then these voice samples are segmented into frame size of 25 milliseconds.

Steps A and B will remain same as in Standard Approach

C. Comparison and Feature Matching

The difference between proposedapproach and standardapproach isinthereferencedatabaseforcentroid value comparison when itcomes tothe feature matching. Here the threshold value is set for the pitch for the identification of the Male and Female voice. Then by comparing with the thresholds, the algorithm is capable of identifying the Gender of the speaker.



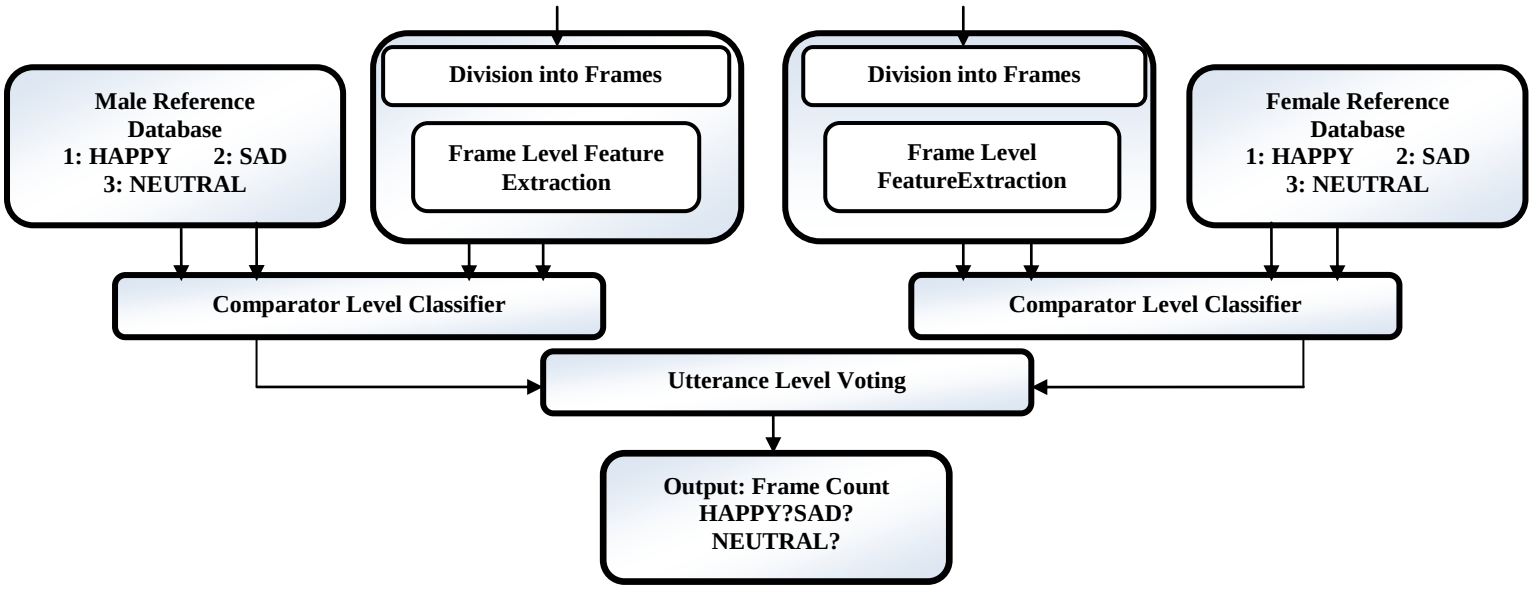


Figure 3: Proposed MFCC Approach [4]

IV. ALGORITHM

1. Capture audio as input.
2. Get frames data from the audio using MATLAB function audioread().
3. Form N groups from M frames. M= no. of frames we retrieve from the audio.
4. Find centroid of each group of the N groups. Now we have a group X having N centroid values.
5. Perform K-means clustering on X. Input : cluster = X, no. of cluster to be formed. = 3
6. Here, we get three cluster centroid values each denoting an emotion.
7. Find mean of three cluster centroids which will be used for gender reference.
8. Compare three cluster centroids with available gender/ emotion reference values.
9. The cluster centroid value having minimum distance from the reference emotion centroid value will be the final EMOTION.

V. EXPERIMENTAL SETUP AND RESULTS

Experimental Setup:

Voice samples are recorded by using Sound Forge 5.0 at a sampling frequency of 8000 Hz with the help of a Microphone. Real Time voices are recorded. This work records each individual's voice sample in Angry, Happy or Neutral mode. Hence, three samples of each mode are recorded.

Below table gives the details of basic audio recognition:

Table 1. Basic Audio Recognition Details

Item	Description
Data source	Microphone
Speaker gender	Male & Female
Speaking language	English, Hindi, Marathi
Audio file format	WAVE (.wav)
Sample resolution	16 bit
Length of sample audio by speaker	10 seconds
Average sentences per speaker	3
Testing	5 seconds
Sample Resolution	8000 Hz

Eventually the number of samples increases that is from 10 to 50 samples. Accuracy performance is used for evaluation purpose. When VQ with K-means algorithm, the accuracy starts from 80% but as the number of samples increases, the performance gradually declines and it comes down to 60% for large number of samples.

But when VQ with LBG algorithm for the same number of samples, we get a low accuracy rate of 68% but here as the number of samples gradually increases, the performance accuracy remains consistently remain the same for large number of samples. Hence the LBG algorithm is gives better performance for large database as compared to K-means algorithm.

V. CONCLUSION

This paper has implemented Vector Quantization algorithm using K-Means. Vector Quantization using K-means works wellenough for lowDatabase showing 80%accuracy approximately. ButVectorQuantization using LBG algorithm shows low accuracy of around 68% but its performance remains consistently constant for evenlargeDatabases. Hence the Recognition accuracy is better andit is high if Vector Quantization with K-means is used.

Results:

Table 2. Gender recognition using reference database performance results

No. of Samples	Total Recognized	Correct Recognition of Gender	Incorrect Recognition of Gender	Efficiency %	Avg Time Sec
10	10	8	2	80	0.2
20	18	15	3	75	0.5
30	25	20	5	68	1.10
40	32	25	7	63	1.40
50	38	30	8	60	2.20

Table 3. Comparison of Vector Quantization using K-means and Vector Quantization using LBG

VQ Method	No. of Samples	Overall Accuracy %	Average Time Sec
Vector Quantization using LBG algorithm	10	68	0.20
	25	63	1.01
	50	59	2.10
Vector Quantization using K-Means algorithm	10	80	0.22
	25	70	1.10
	50	60	2.15

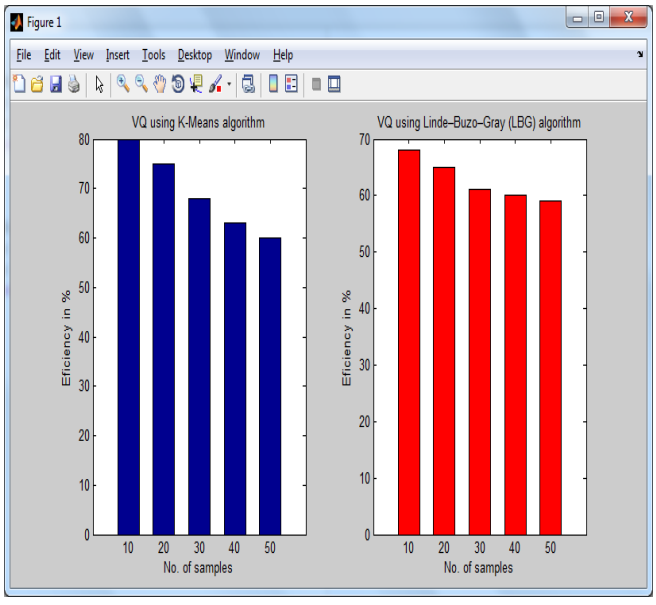


Figure 4: Comparison of VQ using K-means and VQ using LBG algorithm

VI. FUTURE SCOPE

We can build a fear-type emotion recognition system in surveillance systems in the future by considering the public safety. It takes into account the fear experienced by a person. The disadvantage faced by this approach is that, if the emotions are wrongly recognized along with the genders, it can lead to more frustration in a person.

VII. REFERENCES

[1] DimitriosVerveridis and Constantine Kotropoulos, "Emotional speech recognition: Resources, features, and methods", Artificial Intelligence and Information Analysis Laboratory, Department of Informatics, Aristotle University of Thessaloniki, Box 451, Thessaloniki 54124, Greece.

[2] BjörnSchuller and Gerhard Rigoll, "Timing Levels in Segment-Based Speech Emotion Recognition", INTERSPEECH 2006 – ICSLP.

[3] AnkurSapra, Nikhil Panwar, SohanPanwar, "Emotion Recognition From Speech", International Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 2, Feb 2013).

[4] Ismail shahin, "Speaker Identification in Emotional Environments", Iranian Journal of Electrical and Computer Engineering, Vol.8, No.1, WINTER-SPRING 2007.

[5] Nobuo Sato and YasunariObuchi, "Emotion Recognition using MFCC's", Information and Media Technologies 2(3):835-848 92007) reprinted form: Journal of Natural Language Processing 14(4): 83-96 (2007)

[6] Shupeng Xu, Yan Liu and Xiping Liu, "Speaker Recognition and Speech Emotion Recognition Baesd on GMM", Computer Science and Engineering Department Changchun University of Technology Changchun, China.

[7] Shashidhar G. Koolagudi, Kritika Sharma and K. Srenivasa Rao, "Speaker Reciognition in Emotional Environment", School of Computing, Graphic Era University, Dehradun-248002,Uttarakhand, India, School of Information Technology, Indian Institute of Technology Kharagour,Kharagpur-721302,West Bengal, India.

[8] Ling Feng and Lars Kai Hansen, "A New Database For Speaker Recognition", Informatics and Mathematical Modelling,TechnicalUniversity of Denmark Richard PetersensPlads ,Building 321,DK-2800 KongensLyngby, Denmark.

[9] Bjorn Schuller, Gerhard Rigoli and Manfred Lang, "Speech Emotion Recognition Combining Acoustic Features and Linguistic Information IN A Hybrid Support Vector Machine- Belief Network Architecture",Institute for human-computer communicationtechnischeuniversitatMunchen

[10] C. Clavel I. Vasilescu L. Devillers G. Richard T. Ehrette" Fear type emotion recognition for future audio-based surveillance systems" Thales Research and Technology France, RD 128, 91767 PalaiseauCedex, France, LIMSI-CNRS, BP 133, 91403 OrsayCedex, France, TELECOM Paris Tech, 37 rue Dareau, 75014 Paris, France.

[11] SanaulHaq and Philip J.B. Jackson, "Speaker-Dependent Audio-Visual Emotion Recognition", Center for Vision, Speech and Signal Processing (CVSSP), University ofSurrey, UK.