Comparison of Various Techniques for Speaker Recognition

Ajay Kumar
Department of Electronics and Communication Engineering
Ambedkar Institute of Advanced Communication Technologies and Research, Delhi, India
kumarstephen007@gmail.com

Ravindra Singh
Department of Electronics and Communication Engineering
Ambedkar Institute of Advanced Communication Technologies and Research, Delhi, India
rravisingh1990@gmail.com

Kavita
Department of Electronics and Communication Engineering
Ambedkar Institute of Advanced Communication Technologies and Research, Delhi, India
kavitajangra.1014@gmail.com

Shravan Kumar Sehgal
Department of Electronics and Communication Engineering
Ambedkar Institute of Advanced Communication Technologies and Research, Delhi, India
sehgalshravan733@gmail.com

***Abstract*- In this paper, a comparison on different speaker recognition techniques is presented. The techniques we are describing here are Vector Quantization (VQ) by using Linde Bozo Gray (LBG), Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) using the iterative Expectation-Maximization (EM) algorithm. VQ adds the method of considering a large group of feature vectors of a known user and generating a smaller group of feature vectors that signifies the centroid of spreading, i.e. points set apart, so as to reduce the distance between the points. The GMM can be represented in the form of a summation of the VQ model where the clusters are overlying. HMM is a finite group of states, each of which is united with a probability distribution. These techniques are used with their respective algorithm to compute the accuracy rate of speaker recognition. Based on the theoretical analysis, the GMM is more precise as compared to the other techniques.**

Keywords- Speaker Identification; MFCC; Gaussian Mixture Model; Vector Quantization; LBG; Hidden Markov Model.

I. INTRODUCTION

Speaker recognition is that technique of distinctive the speaker from the databank within the speech waveform. Mostly speaker recognition systems comprise two sections. In the initial section, feature extraction is completed. The exceptional feature from the speech signal take-out that is used later for distinctive the speaker. Another section is feature matching therein we tend to equate the extracted speech feature with the databank of speakers [1].

All speakers recognition systems contain three main modules: (a) Acoustic Processing (b) Feature Extraction (c) Feature Matching.

In acoustic processing, an analog signal is received from a speaker and regenerates it into a digital signal for additional digital processing. The signal is then fed into the spectral instrument for feature extraction. Feature extraction in which a quantity of information is extracted from the voice analog signal that may later be used to characterize every speaker.

Feature matching includes the particular method to spot the unknown speaker by analysis the extracted features from voice input with those from a group of well-known speakers. Feature matching techniques such as Dynamic Time Wrapping (DTW), VQ, GMM [2] and HMM are employed in speaker recognition.

VQ is a technique in which a vectors are plotted from a huge vector space to a restricted variety of areas in that space. Every region is referred to as a cluster and can be considered as its center referred to as code word. The cluster of all code words is termed a codebook. VQ is employed for lossy information compression technique based on the concept of block coding that's a fixed-to-fixed length algorithm [3] whereas pdf considered as a weighted sum of Gaussian component densities. GMM parameters are expected from training information using EM algorithm approximation from well-trained prior model.

The problems that were considered are (a) GMM may be efficiently used to complete the aforementioned task, (b) Accuracy of the GMM as constant quality modelling, (c) Performance examine of the system, (d) Presentation of the GMM as a constant quantity modelling technique [4]. The weighted vectors of those exponential models are constrained to lie in a subspace shared by all the Gaussians [5]. The principle behind the speaker recognition supported GMM, consistent with the parameters were extracted from speakers' feature set to create GMM. Parameters of this model were determined using spatial distribution of the voice feature parameters [6].

HMM is a probabilistic model consists of variable expressive observations, variables which are hidden, the initial state distribution matrix, transition matrix, and the constraints for all observation distributions [7].

II. VECTOR QUANTIZATION (VQ)

In VQ technique, a vector of a large space is mapped to a fixed number of regions in that space. These types of regions are called clusters and these are represented by its center that

978-93-80544-34-2$31.00 ©2019 IEEE

is also known as centroid .To form a codebook, all the centroids are collected.

For the process of speech coding the step of quantization is mandatory for reducing the bits number that have been used to characterize the samples of a signal. The memory necessities for storage and computational complexity in the recognition space finally become excessively high. To wrapping the unique data into a small set of desirable points, we used a VQ codebook as an effective means of representing speaker specific features.

Recognition involves choosing a codebook that quantizes an unknown word with minimum distortion. The most widespread codebook vogue formula is LBG [8] . The LBG iteratively performs two steps, first one is the partition and second is generation of new codebook, supported the criterion of minimizing the quantization error. LBG formula is variety of a K-means clump formula that takes a bunch of input vectors as input and generates a representative set of vectors as output in step with the similarity live.

A speaker associated to VQ codebook is made for every speaker training vectors. LBG algorithmic rule is employed for clump a group of L coaching vectors into a group of M codebook vectors is being employed. The LBG algorithmic is used to design the codebook as we have discussed the source input is grouped into vector then we find the closest value of these vectors with the help of codebook.

The algorithm is employed by the following procedure (Figure 1):

1. Find the centroid that is the entire set of training vector of the codebook to design the first vector codebook.
2. Splitting a centroid, means doubling the size of present codebook Y_n. It can be represented as per rule:
$$Y_n^+ = Y_n(1+\beta)\,, Y_n^- = Y_n(1-\beta) \qquad (1)$$
where the range of n is from one to the size of the present codebook, and β is known as splitting parameter.
3. Assign each training vector to a cluster related with the nearest codeword through nearest neighbor search procedure (clustering of vector).
4. Update each centroid of a codebook.
5. At last the distortion (distances of all training vectors) is measured by repeating third and fourth step until the distance lies below threshold value.
6. Repetition of second, third and fourth steps until a codebook size of M is calculated.

The speaker associated to the codebook with smallest total distortion is recognized as the given speaker. When X(n) is quantized ,then distortion is measured that is an average distortion is minimalized over the full training set of the distortion between the vectors a_i and b_j is represented as d (a_i, b_j).). I is denoted as a given set of training.

$$D = \frac{1}{I}\sum_{i=1}^{I} \min d\ (a_i, b_j) \qquad (2)$$

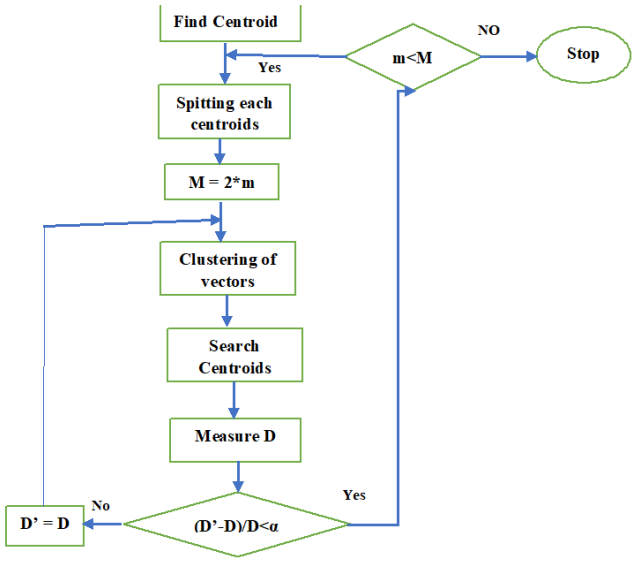

Fig. 1. Flow chart of LBG algo

III. Gaussian Mixture Model (GMM)

A GMM is understood as a combination of gaussian and mixture model (Figure 2). Gaussian may be a features symmetric 'bell curve' form that quickly in a GMM constant quantity likelihood falls off to zero describe a mixture model as a 'probabilistic model that accepts the first information that's associated to a mixture distribution'. The common mixture distribution is that the Gaussian density function where every of the mixture elements are Gaussian distributions, each with their mean and variance parameters

$$P(x) = w1N(x|\mu_1\sum_1)\ \ldots + w1N(x|\mu_2\sum_2)\ldots + w1N(x|\mu_n\sum_n) \qquad (3)$$

μ_i's are mean and $\sum_i$'s are covariance matrix of individual components.

A GMM is described by its PDF, as given in Equation (4).

$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \qquad (4)$$

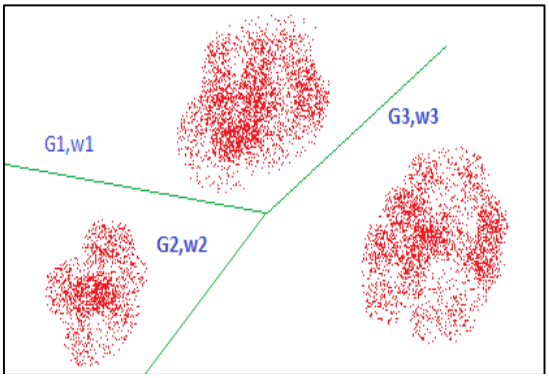

Fig. 2. Gaussian mixture model

The whole GMM is parameterized through the mean vectors, mixture weights and covariance matrices from all constituent densities. GMMs are usually employed in biometric systems due to their ability of representative of an oversized class of sample distributions. A block diagram of GMM is presented in Figure 3.

The belief of speaker recognition using GMM with the speech parameters which is taken from speakers feature set to create GMM, parameters were classified using spatial distribution of the speech feature parameters. Known speaker is recognized by comparison as a result of completely different speaker has different voice model parameters. GMM was primarily a linear weighted mixture of a multi-dimensional PDF, it is given in Equation (5):

$$P\left(\frac{x_t}{\lambda}\right) = \sum_{i=1}^{M} w_i \, f_i(x) \qquad (5)$$

Here, w_i is a weight, it shows the scale of the Gaussian distribution, therefore $\sum_{i=1}^{M} w_i = 1$. $f_i(x)$ is a joint Gaussian probability distribution with D-dimensional, expressed as Equation (6),

$$f_i(x) = \frac{1}{|\sum i|^{\frac{1}{2}} 2\pi^{\frac{D}{2}}} e^{[-\frac{1}{2}(X-\mu_i)^t \sum_i^{-1}(X-\mu_i)]} \qquad (6)$$

Here, μ_i is a mean, $\sum_i$ shows covariance matrix. GMM is expressed by the weights, mean and covariance matrix $\lambda = (w_i, \mu_i, \sum_i)$.

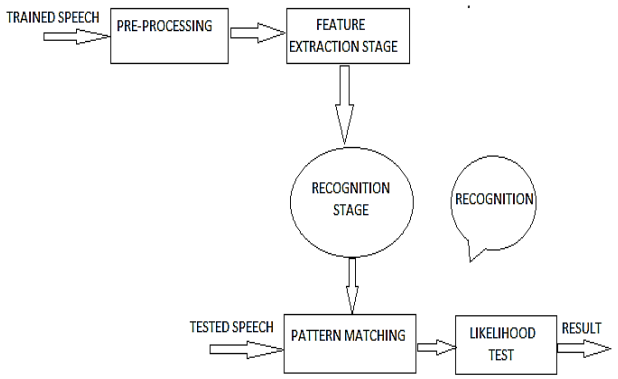

Fig. 3. Block Diagram of GMM

Initially it takes the trained speech data which is pre-processed in the feature extraction module. The Gaussian Mixture Model was designed as the fundamental acoustic model. The rule matches the fundamental quantity speech input by the user then implements the actual technique of categorizing the speaker by equating the extracted data from the period speech input with a information of given speakers. A series of acoustic choices area measure extracted from the speech signal, then recognition algorithms are used. Thus, the choice of acoustic choices is crucial for the system performance.

A. Expectation-Maximization (EM) Algorithm

EM algorithm is an iterative optimization technique which is operated locally.

We know the basic Gaussian function, as given in Equation (7),

$$g(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \qquad (7)$$

The Expectation Step: In this step we calculate all of the responsibities using Equation (8).

$$r_{ic} = \frac{w_c N(x_i|\mu_c,\sigma_c)}{\sum_{j\epsilon[0,k]} N(x_i|\mu_j,\sigma_j) w_j} \qquad (8)$$

The Maximization Step: We have the responsibilities for every dataset with respect to each Gaussian curve, we use this to improve our guess of each curve's mean, standard deviation and the weighting factor w_c using Equations (9), (10) and (11).

$$\text{Mixture Weights } (w_c) = \frac{1}{T}\sum_{t=1}^{T} P(i \parallel x_t) \qquad (9)$$

$$\text{Mean } (\mu i) = \frac{\sum_{t=1}^{T} P(i\|x_t,\lambda) x_t}{\sum_{t=1}^{T} P(i\|x_t,\lambda)} \qquad (10)$$

$$\text{Diagonal Co-Variance } (\sigma i^2) = \frac{\sum_{t=1}^{T} P(i\|x_t,\lambda) x_t{}^2}{\sum_{t=1}^{T} P(i\|x_t,\lambda)} - \mu i^2 \qquad (11)$$

IV. HIDDEN MARKOV MODEL (HMM)

This model basically refers to a model where system that is modelled by using Markov process in which the states always be unobserved. It is characterized as the easiest dynamic Bayesian process.

Markov model refers to model in which observer can view the state directly, and thus the state transition probabilities are the only featured parameters, the HMM, differs in the fact that the observer cannot directly view the state, but the output is visible in the form of token or data and this output which is depends on the state. Each and every state has probability distribution over the possible output tokens. Therefore, the series of tokens produced by an Hmm model gives some information about the series of states. This Model is also referred to as statistical finite state machine, in which the system being demonstrated is predictable to be in the Markov process. It comprises of states to model the sequence of observation data. Here the states are not visible but the output depending on the states is known to us. In this model, the Bayesian dynamic network with Markov rule is used. A block diagram of HMM is given in Figure 4.

A. Hidden Markov Model Process

- Set of states: $\{A_1, A_2, A_3....A_N\}$.
- Process changes from one state to another producing a sequence $\{A_{i1},A_{i2},A_{i3}.....A_{ik}\}$.
- Markov chain property is when the probability of each following state depends only on the previous state that is represented as :
- $P(A_{ik}|A_{i1},A_{i2},A_{i3}.....A_{ik-1}) = P(a_{ik}|A_{ik-1})$ (12)
- States are not visible, but each states are randomly produced M observation {V1,V2,….., Vn}.
- The subsequent chances got to be specified: matrix of transition chances B = (b_{ij}), $b_{ij} = P(a_i \mid a_j)$, matrix of observation chances C = (c_i (v_m)), $c_i(v_m) = P(v_m \mid a_i)$

and a vector of initial chances $\pi = (\pi_i)$, $\pi_i = P(a_i)$. Model is categorized by M=(B, C, π).

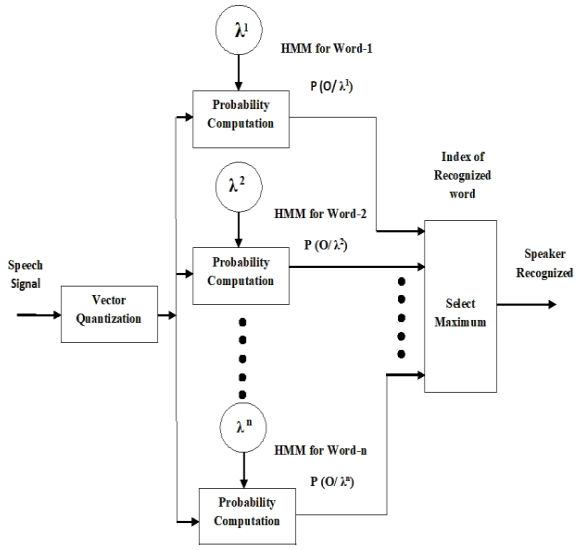

Fig. 4. Block diagram of HMM

B. *Algorithm for Hidden Markov Model for recognition*

1. Consider every spoken sentence or word to be characterised with a series of speech vectors or observations λ, expressed by Equation (13):
$\lambda_t = \lambda_1, \lambda_2, \lambda_3, \lambda_{4.....} \lambda_N.$ (13)
where the λ_t spoken word observed vector at time t.
2. The likelihood of the word with the whole sentence victimization Bayes' Theorem using Equation (14):
$P(Wi|\lambda) = Wi)\ P(Wi) / P(\lambda).$ (14)
3. For the observed vector sequence, the direct calculation of the joint conditional probability using Equation (15):
$P(\lambda_t = \lambda_1, \lambda_2, \lambda_3, \lambda_{4,....,} \lambda_N| W_i)$ (15)
from examples of spoken words is not attainable. So, arithmetical finite machine like HMM is used.
4. The chance that λ is formed by the model H moving through the state series Y is calculated only because the product of the transition chances of states and therefore the output chances of the states. Thus , the state series as represented as :
$YP(\lambda, Y|H) = a_{12}\ b_2(\lambda_1)\ a_{22}\ b_2(\lambda_2)\ a_{32}\ b_2(\lambda_3)$ (16)
The parameter which is known to us is the observation sequence λ and only this parameter is visible where the hidden parameter is underlying state sequence which is denoted by Y'. Hence it also known as HMM.
5. An unknown Y, the likelihood is estimated by combining over all the possible state series Y=y(1), y(2), y(3), ……., y(T), that is

$$P(\lambda|H) = \sum_Y a\ y(0)y(1)\ \prod_{t=1}^{T} by(t)\ (\lambda t)\mathrm{ay(t)}y(t+1) \quad (17)$$

Y(0) is constrained to be the model entry state y(T+1).

$$P\hat{}(\lambda|H) = \max\ \{\sum_Y a\ y(0)y(1)\ \prod_{t=1}^{T} by(t)\ (\lambda t)\mathrm{ay(t)}y(t+1)\} \quad (18)$$

6. For a given set of models H with respect to words Wi, arg max{ $P(W_i|\lambda)$} can be solved using Equation (19).
$P(W_i|\lambda) = P(\lambda|H_i)$ (19)

Hence, above condition is satisfied and the speech is recognized.

C. *Advantages of HMM formulation*

- HMM has a modest & uniform structure. This makes it so easy to apply the recognizer. E.g. all tri-phones can have the same left-to-right 3 state topology differing only in separate parameter values rather than in HMM structure.
- HMM covers both temporal & spectral changeability flexible manner.
- Not only an allophone, phoneme, syllable, or a word can be characterised by a HMM, but an entire sentence can be represented by 1 large composite HMM.
- Highly effectual algorithms also exist for defining HMM parameters straight from speech data-like Viterbi algorithm.
- HMM formulation can be applied not only to English but equally well to the other languages.

V. EXPERIMENT AND RESULTS

To evaluate the performance of VQ-LBG training technique by running experiments, the information was taken from the Arabic digit corpus collected at the University of Badji Mokhtar in urban center, Algeria [9]. From these data, the codebooks are created for each speakers with their updated centroid using LBG algorithm. The accuracy rate of VQ is recognized for each speaker associated to the codebook with smallest total distortion. To find the performance of proposed speaker identification system signal of 29 speakers are recorded in a laboratory environment. In the experiment of text dependent identification, speaker has uttered the word 'Hello' [10]. The speaker model was generated for respectively speaker with the help of MFCCs using GMM [11]. EM algorithm is used to calculate all of the responsibities for every dataset with respect to each Gaussian curve. For accuracy, the following factors are improved that is mean, standard deviation and the weighting factor w_c of curve. Like others, HMM is also used on given speakers with algorithm to find the parameters of model with the help of finding the various required probability.

TABLE I. PERFORMANCE OF DIFFERENT TECHNIQUES

Acoustic Techniques with MFCC	Highest recognition rate with accuracies (%)

Vector Quantization	91.43
Gaussian mixture model	99.22
Hidden Markov model	97.26

VI. CONCLUSION

The performance of VQ using LBG to measure and create the codebook, HMM with suitable algorithm to find the all possible probability of given condition and GMM with EM algorithm to find the finest value of responsibities (mean, standard deviation and the weighting factor of curve) are compared for speaker recognition. It is confirmed the MFCCs feature model using GMM is superior to VQ and HMM. To calculate peak of dimensional features, actual feature size is important, then GMM needs more parameters to describe speakers. The average recognition rate attained for MFCCs with GMM is 99.22%, for HMM is 97.26% and for VQ with LBG is 91.43%. These given data are taken based on theoretical knowledge of speaker recognition from various proposed papers.

REFERENCES

[1] Kamale, H. E, "Vector quantization approach for speaker recognition" , International Journal of Computer Technology and Electronics Engineering, vol. 3, 2013.

[2] Soni, M. K., "Speaker recognition using MFCC and vector quantisation", International Journal on Recent Trends in Engineering & Technology, vol. 11, no. 1, 2014.

[3] Ishak, K. A., "Speaker verification using vector quantization and hidden Markov model", in Proc. of the 5th IEEE Student Conference on Research and Development (SCOReD 2007), 2007, pp. 1-5.

[4] Vyas, M, "A Gaussian mixture model based speech recognition system using Matlab", Signal & Image Processing, vol.4, no.4, 2013.

[5] Visweswariah, K, "Subspace constrained Gaussian mixture models for speech recognition", IEEE Transactions on speech and audio processing, vol 13, no. 6, 2005.

[6] Juan, Z. C, "The research of speaker recognition based on GMM and SVM", in Proc. of the International Conference on System Science and Engineering (ICSSE), 2012.

[7] Wu, X, "Acceleration of the LBG algorithm", IEEE Transactions on Communication.

[8] Charles, S. P, "A non-homogeneous hidden Markov model for precipitation occurrence", Journal of the Royal Statistical Society.

[9] Ma, D. D, "An improved VQ based algorithm for recognizing speaker-independent isolated words". in Proc. of the IEEE International Conference on Machine Learning and Cybernetics (ICMLC), 2012.

[10] S. M ,"Comparison of Vector Quantization and Gaussian Mixture Model using Effective MFCC Features for Text-independent Speaker Identification", International Journal of Computer Applications, vol. 134, no.15, 2016.

[11] M. K. I, "Speaker identification using Cepstral based features and discrete Hidden Markov Model", in Proc. of the International Conference on Information and Communication Technology (ICICT 2007), 2007.