Comparison of Vector Quantization and Gaussian Mixture Model using Effective MFCC Features for Text-independent Speaker Identification

S. B. Dhonde
Department of Electronics Engineering
A.I.S.S.M.S. Institute of Information Technology.
Pune, India

S. M. Jagade
Department of Electronics & Telecommunication Engineering
TPCT, College of Engineering
Osmanabad, India

ABSTRACT
In this paper, the performance of speaker modeling schemes such as vector quantization (VQ) and Gaussian mixture model (GMM) is compared for speaker identification. Along with the effective size of feature set, model based approaches are typically used as a solution for robustness issues of speaker recognition systems. Gaussian Mixture Model (GMM) is versatile parameter estimation approach whereas; Vector Quantization (VQ) is based on template modeling. Here, first, MFCC features are used to extract speaker specific speech features for text-independent speaker identification. MFCC features are then modeled using Vector Quantization (VQ) and Gaussian mixture model (GMM) and their performance is compared in the context of speaker identification. The average recognition rate achieved for MFCC with GMM is 99.2% and for MFCC with VQ is 98.4% on TIMIT database consisting of 64 speakers.

General Terms
Speech Signal Processing

Keywords
Speaker Identification, Text-independence, Feature Extraction, Vector Quantization, Gaussian Mixture Model

1. INTRODUCTION
Speaker identification is to identify person from known set of voices. The process of speaker identification is divided into two main phases, i.e., the enrolment phase and the identification phase. During the enrolment phase which is also known as training phase, speech samples are collected from the speakers, and are used to train their models. The collection of enrolled models is also called a speaker database. In the identification phase, a test sample from an unknown speaker is compared against the speaker models stored in the speaker database. Both the phases involve a common step, i.e., feature extraction, where the speaker dependent features are extracted from the speech sample. The main purpose of this step is to reduce the amount of test data while retaining the speaker discriminative information.

The recent studies in speaker recognition has mainly focused on robustness issues by providing typical solutions categorized as feature based and model based approaches [1]. There has been many attempts to enhance the robustness of MFCC scheme for speaker identification either by combining complementary features with MFCC or even replacing Mel filter bank [2] – [8]. However, concatenating additional features with MFCC increases feature vector size. The increased feature vector size requires more computational time and storage space [8] [9].

In model based approaches, parameters of speaker model are modified instead of feature vectors [10]. Template models and stochastic models are two speaker models for classification in speaker identification. These models are also known as nonparametric and parametric models respectively [10]. Template model such as vector quantization directly compares training and testing features. It assumes that either one is the replica of other [10]. The degree of similarity is represented by the amount of distortion between training and testing features. The vector quantization (VQ) and dynamic time warping (DTW) are template models used for text-independent and text-dependent speaker identification respectively. A probabilistic model of the speech signal can be built as an alternative approach to template models. Such models are termed as stochastic models. The parameters of the probability density function from a training sample are estimated in training phase. The likelihood of the test utterance with respect to model is evaluated for the matching of training and testing features in testing phase. The Gaussian mixture model (GMM) and hidden Morkov model (HMM) are used for text-independent and text-dependent speaker identification respectively. The artificial neural networks (ANNs) and support vector machines (SVMs) are other approaches for speaker identification. These approaches model the boundary between speakers and hence termed as discriminative models [10]. Whereas, vector quantization and Gaussian mixture model are generative models which estimates feature distribution within each speaker. Gaussian mixture model is widely used for speaker modelling in speaker identification system [1] - [3]. Use of neural networks for speaker identification system has been studied in [5] – [6].

In this paper, the performance of vector quantization and Gaussian mixture models is compared for text-independent speaker identification system by modeling MFCC features. This paper is organized as follows. Speaker modeling techniques i.e. vector quantization and Gaussian mixture model are described in section 2. Experimental set-up is presented in section 3 followed by results and discussion in section 4. The conclusion is presented in section 5.

2. SPEAKER MODELLING TECHNIQUES
The classification of speaker is a decision process based on previously learned or stored information for validating the speaker. First, the speech features emphasizing on speaker specific properties are needs to be computed from input speech signal. These features should be robust suppressing statistical redundancies as the quality of sub-sequent steps mainly depends on features [13]. MFCC feature extraction

scheme is widely used in speaker recognition system [13]. These features are used to create speaker model in speaker recognition system. In training phase, the speaker model is trained and stored in the database using the features computed in the feature extraction step. A measure of similarity of the features extracted from unknown speech sample in testing phase and the speaker model stored in database (training phase) is computed using matching score. Template model such as vector quantization and stochastic model such as Gaussian Mixture Model are two speaker models for classification in text-independent speaker identification.

2.1 Vector Quantization

Vector quantization is the process of data compression in which a small set of feature vectors is produced from the large set of feature vectors of a particular speaker. This small set represents the centroids of the distribution. VQ is a process of mapping feature vectors from a vector space to a finite number of regions in that space. These regions are called clusters and represented by their central vectors or centroids. A set of centroids, which represents the whole vector space, is called a codebook. Code book of a speaker is generated by applying VQ on the set of feature vectors extracted from the speech sample. For clustering of feature vectors into a set of codebook, LBG algorithm is used in this work. In testing phase, Euclidean distance between features of an unknown speaker and speaker models stored in the database is calculated. The speaker is identified on the basis of minimum distance. Speaker model which is having minimum distance with the features of an unknown voice is selected as an identity of unknown speaker

2.2 Gaussian Mixture Model

Gaussian mixture model can be considered as an extension of the Vector quantization model, in which the clusters are overlapping. That is, a feature vector is not assigned to the nearest cluster as in, but it has a nonzero probability of originating from each cluster [10]. Representation for each speaker with his/her GMM, which is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters of GMM are estimated using the expectation maximization algorithm. These parameters of GMM are computed in training phase to create a speaker model. In testing phase, the speaker model having highest a posteriori probability for the features of an unknown voice is selected as identity of that unknown speaker.

3. EXPERIMENTAL SETUP

The performance of vector quantization and Gaussian mixture model for speaker identification is evaluated on TIMIT [12] database consisting of 64 speakers. Four male and four female speakers from 8 dialect regions of TIMIT database were used as suggested in [1]. For training of speaker model, eight sentences consisting of five SX and three SI (approximately 24 seconds) were used. For testing purpose, two remaining SA sentences (duration of 3 seconds each) were used. All the experiments have been performed using HP Pavilion g6 laptop with CPU speed of 2.50 GHz, 4 GB RAM and MATLAB 8.1 signal processing tool. Twelve cepstral coefficients out of 20 MFCC filters were selected. For MFCC feature extraction, speech signal is pre-processed with pre-emphasis coefficient 0.95, frame length of 256 samples per frames with 50 % overlap followed by the hamming window. The speaker model is generated for each speaker from the MFCCs using vector quantization (LBG algorithm) and Gaussian mixture model (GMM). The number of clusters of vector quantization was 32. For GMM, the number of mixtures was 32.

4. RESULT AND DISCUSSION

First, MFCC features are extracted from speech signal and speaker models are created. Here, speaker modeling schemes vector quantization and Gaussian mixture model are used and their performance is compared by calculating average recognition rate.

Table 1. Evaluation of MFCC with VQ and GMM on 64 speakers of TIMIT

Sr. No.	Approach	Average Recognition rate
1	MFCC with VQ	98.43
2	MFCC with GMM	99.22

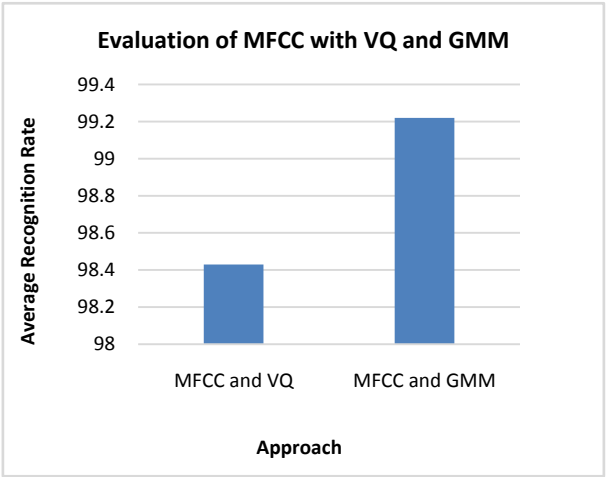

Fig 1: Evaluation of MFCC with VQ and GMM

The average recognition rate achieved for MFCC and Gaussian mixture model (GMM) is 99.2%. It is found that MFCC features excluding temporal coefficients have achieved same recognition rate as discussed in [1]. Excluding temporal derivatives reduces the size of the speaker model which is stored in speaker database. This will reduce the storage space required to save these models. Also, low dimensional feature vector size reduces computational time required for training and testing phase. This is because classifier using high dimensional features requires more parameters to characterize speaker model. The average recognition rate achieved using MFCC with Gaussian mixture model (GMM) approach is better than MFCC with vector quantization (VQ).This is because in case of GMM, a feature vector is not assigned to the nearest cluster as in, but it has a nonzero probability of originating from each cluster.

5. CONCLUSION

In this paper, the performance of vector quantization and Gaussian mixture model is compared for text-independent speaker identification system. It is demonstrated that MFCC features modelled using Gaussian mixture model are superior to vector quantization. Also, effective feature size is important because if high dimensional features are considered then GMM requires more parameters to characterize speaker model. This increases the computational time. Twelve MFCC features are modelled using Vector Quantization (VQ) and Gaussian mixture model (GMM) .The average recognition

rate achieved for MFCC with GMM is 99.2% and for MFCC with VQ is 98.4% on TIMIT database consisting of 64 speakers.

6. REFERENCES

[1] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues," IEEE Circuits Syst. Mag., vol. 11, no. 2, pp. 23–61, Second Quarter 2011.

[2] R.Shantha Selva Kumari, S. Selva Nidhyananthan, Anand.G, "Fused Mel Feature sets based Text-Independent Speaker Identification using Gaussian Mixture Model", International Conference on Communication Technology and System Design 2011, Journal on Procedia Engineering, Elsevier, vol. 30, pp. 319–326, 2012.

[3] Mangesh S. Deshpande, Raghunath S. Holambe, "New Filter Structure based Admissible Wavelet Packet Transform for Text-Independent Speaker Identification", International Journal of Recent Trends in Engineering, vol. 2, no. 5, pp. 121-125, 2009.

[4] Sumithra Manimegalai Govindan, Prakash Duraisamy, Xiaohui Yuan, "Adaptive wavelet shrinkage for noise robust speaker recognition", Journal on Digital Signal Processing, Elsevier, vol. 33, pp. 180-190, 2014.

[5] Noor Almaadeed, Amar Aggoun, Abbes Amira, "Speaker identification using multimodal neural networks and wavelet analysis", IET Journals and Magazines, vol. 4, no. 1, pp. 18-28, 2015

[6] Khaled Daqrouq, Tarek A. Tutunji, "Speaker identification using vowels features through a combinedmethod of formants, wavelets, and neural network classifiers", Journal on Applied Soft Computing, Elsevier, vol. 27, pp. 231-239, 2015.

[7] M.Hassan Shirali-Shahreza, Sajad Shirali-Shahreza, "Effect of MFCC Normalization on Vector Quantization Based Speaker Identification", Signal Processing and Information Technology (ISSPIT), 2010, pp.250-253, December 2010.

[8] Pawan K. Ajmera, Dattatray V. Jadhav, Ragunath S. Holambe, "Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram", Journal on Pattern Recognition, Elsevier, Vol.44, Issue 10-11, Pages 2749-2759, 2011.

[9] Amol A. Chaudhari, S. B. Dhonde, "Effect of Varying MFCC Filters for Speaker Recognition", International Journal of Computer Applications, vol. 128, no. 14, pp. 7-9, October 2015.

[10] Tomi Kinnunen, Haizhou Li, "An overview of text-independent speaker recognition: From features to supervectors", Journal on Speech Communication, Elsevier, vol. 52, no. 1, pp. 12–40, 2010.

[11] Holambe, Raghunath S., Deshpande, Mangesh S., "Advances in Non-Linear Modeling for Speech Processing", SpringerBriefs in Speech Technology, Section 2, Section 6, pp. 11-15, 77-82, ISBN 978-1-4614-1505-3, 2012.

[12] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren, V. Zue, TIMIT acoustic-phonetic continuous speech corpus, http://catalog.ldc.upenn.edu/ldc93s1, 1993.

[13] Md Jahangir Alam , Tomi Kinnunen , Patrick Kenny , Pierre Ouellet, Douglas O'Shaughnessy, "Multitaper MFCC and PLP features for speaker verification using i-vectors", Journal on Speech Communication, Elsevier, vol. 55, no. 2, pp. 237-251, 2013.