A SPEAKER RECOGNITION SYSTEM USING COMBINATION METHOD BETWEEN VECTOR QUANTIZATION AND GAUSSIAN MIXTURE MODEL

NGUYEN TU HA[1,*], *NGO QUOC HUNG*[2,**]
[1]*Hue University of Education*
[2]*Hue College of Transportation*
[*]*Email: nguyentuha82@gmail.com*
[**]*Email: ngqhung.gtvthue@gmail.com*

Abstract: Speaker recognition is a biometric technique to recognize people's identity based on their voice signal. A recognition system has two main requirements, which are high accuracy recognition rate and short processing time under large amount of training data. This paper propose a method to solve the two above requirements by performing a combination of two advantages of each VQ and GMM model to provide a new model can be called a "Hybrid VQ/GMM-UBM model". This model not only takes the advantage of high accuracy in GMM method but also improve the accuracy rate and reduce the amount of computation of the system when combined with VQ model. The efficiency of the model is evaluated by computational time and accuracy rate compared to GMM models. Experimental results showed that the hybrid VQ/GMM-UBM model had better accuracy.

Keywords: Vietnamese Speaker recognition, Gausisian Mixture Model, Universal Background Model, Vector Quantization, Biometrics.

1. INTRODUCTION

Speaker recognition is a biometric technology derived from areas of speech processing. The speaker recognition field has over 50 years of research and development. The general idea of speaker identification tasks is to assume that the voice of human is unique to each individual, and it can be used as a distinguishing characteristics to identify the owner of that voice among other individuals.

Many researchers have proposed various speaker recognition techniques; and the two most popular methods are Vector Quantization (VQ) and Gaussian Mixture Model. Each method has its own advantage. VQ method can performe simply and has fast computation time. The major disadvantage of this method, however, is that its recognition accuracy rate is not high, especially with large data sets. Meanwhile, the GMM-UBM has greater accuracy rate than VQ. But still, for long processing time, this process does not always produce satisfying result in practice.

Speaker recognition system has two operation phases, *training phase* and *test phase.* In both phases, speech signal is pre-processed to improve the voice quality and reduce noise. It then was extracted characteristics to obtain the set of feature vector. In the training

Journal of Science, Hue University of Education
ISSN 1859-1612, No. 03(53)/2019: pp. 14-21
Received: 07/12/2018; Revised: 04/6/2019; Accepted: 05/7/2019

phase, the characteristic vector is used to train the speaker model. Many methods are used to train speaker model, from the simplest one which is used to build codebook model using vector quantization (VQ) (yet the accuracy of this method is not high) to complex methods such as Gaussian Mixture Model – Universal Background Model. The overall structure of speaker recognition system is depicted in Figure.1

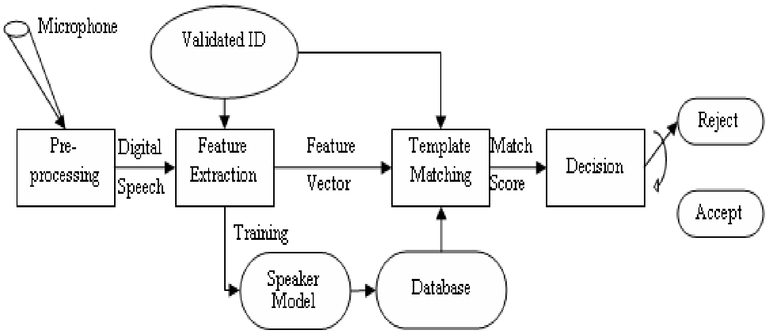

Figure 1. *General Speaker Recognition System*

2. PRE-PROCESSING AND FEATURE EXTRACTION

2.1. Amplitude Normalized

Voice data was obtained with the amplitude fluctuation. Even if the speaker says with a standard volume, the amplitude of obtained signal can still be unstable. This usually happens when the speaker slightly turns away or moves the microphone closer to his mouth or pulls it away more than a few centimeters. This fluctuation affects to the recognition results.

The normalization is necessary. But it is not required the signal amplitude to be good, not too small to lose its characteristics. Thus, we can simply implement by multiplying each point with an appropriate coefficient *k*.

$$k = \frac{(32767/2) - 100}{\max(|\mathrm{s(n)}|)}$$

2.2. Silence Removal

Speech signal usually contains many silence intervals at various points such as at the beginning points of the signal, between words of the sentence or at the end of the signal. If the signal contains silence intervals without treat-ment, it will occupy resources of system to process on these signal intervals. The silence intervals, however, do not have any contribution to the identification, even it can interfere to the processing. Hence the silence intervals must be treated and eliminated before implementing feature extraction. Nowadays, a number of met-hods can effectively solve this problem as voice activity detection – VAD [8], short time energy or spectral centroid.

2.3. Feature Extraction

The extraction of the best parametric repre-sentation of acoustic signals is an important task to produce a better recognition performance. The efficiency of this phase is important for the next phase since it affects its behavior. MFCC is based on human hearing perceptions which cannot perceive frequencies over 1Khz. In other words, in MFCC is based on known variation of the human ear's critical bandwidth with fre-quency is employed. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech. Logarithmic spacing is above 1000Hz.

The overall process of the MFCC is shown in Figure 2.

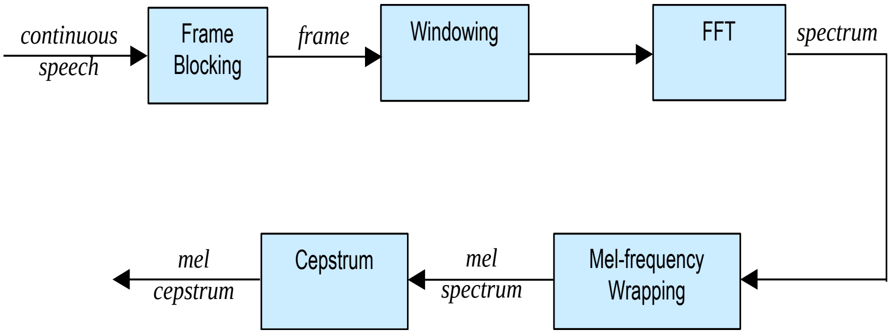

Figure 2. *Computing of mel-cepstrum*

3. MODEL TRAINING

3.1. Vector Quantization

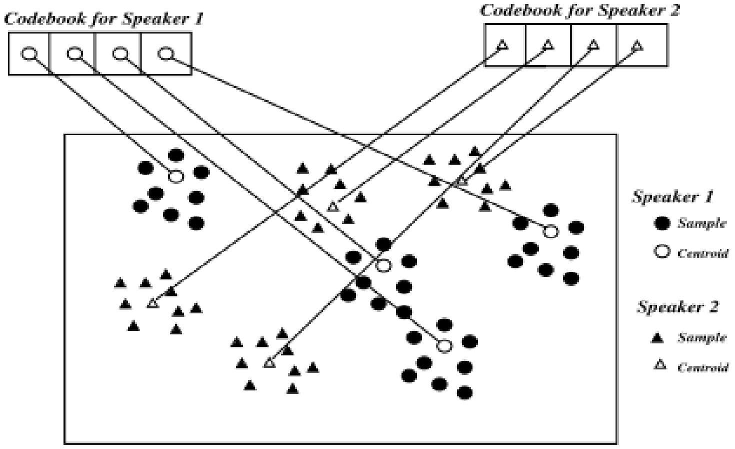

Figure 3. *Vector Quantization based Codebook of two speaker*

Vector quantization (VQ) is a process of mapping vectors from a vector space to a finite number of regions in that space. These regions are called clusters and are represented by their centroids. A set of centroids, which represents the whole vector space, is called a codebook. In speaker identification, VQ is applied on the set of feature vectors extracted from the speech sample and as a result, the speaker codebook is generated. Such codebook has a significantly smaller size than extracted vector set and is referred as a speaker model. This codebook is generated by many algorithm such K-mean, LBG…

During the matching, a matching score is computed between extracted feature vectors and every speaker codebook enrolled in the system. In this paper, match score is a Euclean distance between feature vectors and codebook of speaker as formula:

$$D\left(X,C\right)=\frac{1}{N}\sum_{i=1}^{N}\min_{j}\left\|x_{i}-c_{i}\right\|^{2}$$

where X is a set of N extracted feature vectors, C is a speaker codebook, x_i are feature vectors, c_i are codebook centroids.

3.2. Gaussian Mixture Model – Universal Background Model

Gaussian Mixture Model is a type of statistical model which was first introduced by Reynolds [6]. In this approach, UBM which is a large GMM trained to represent the speaker independent distribution of features is used. UBM can be gender independent/dependent model and use EM algorithm to training [6]. After UBM was trained, speaker dependent models are derived from the UBM by *maximum a posteriori (MAP)* adaptation [6]. To form a speaker dependent model, first, the log-likelihood of each gender dependent model given the input data is calculated. The gender is determined by selecting the gender-model with the higher score. The corresponding gender dependent UBM is used to adapt a speaker dependent model (Figure 4) [6]. Regarding speaker adaptation three EM-steps and a weighting factor of 0.6 for the adapted model and correspondingly 0.4 for the UBM are used to merge these models to final speaker dependent model [7].

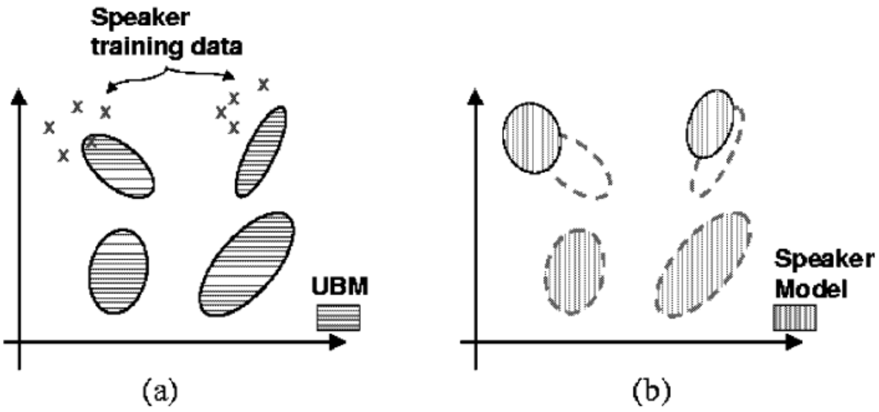

Figure 4. *Adaptation of speaker model from UBM* [6]

3.3. The combination of VQ and GMM-UBM (VQ/GMM-UBM)

As mentioned before, VQ based solution is less accurate than that of the GMM. In this paper, a method took the superiority of VQ, which is simplicity computation to distinguish between male and female speaker. After, we use of GMM merrits to identify the speaker identity in the smaller subgroup.

In this approach, a testing processing was built on three stages. In the first stage, feature vectors of testing speaker was compared with male codebook and female codebook using Euclean Distance to decide gender of testing speaker. Male codebook was trained from a large data of male speakers to represent the male speaker; the same procedure for female codebook. In the second stage, after knowing gender of testing speaker, feature vectors of testing speaker was compared to each VQ model of trained speaker in same gender group to define ten trained speaker which had the highest matching scores. In the third stage, ten trained speakers were computed the log-likelihood with feature vectors of testing speaker using GMM speaker to define a final speaker model who had the highest matching score. After, a threshold was applied to decide "accept" or "reject". Figure 5. represents speaker identification processing with com-bination of VQ/GMM-UBM.

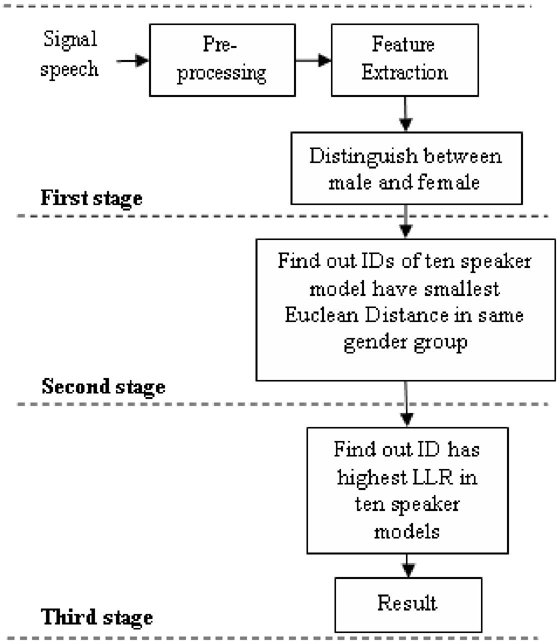
Signal speech
Pre-processing
Feature Extraction
Distinguish between male and female
First stage
Find out IDs of ten speaker model have smallest Euclean Distance in same gender group
Second stage
Find out ID has highest LLR in ten speaker models
Result
Third stage

Fig 5. *Speaker identification processing with combination of VQ/GMM-UBM*

Since the idea is using both models of VQ and GMM-UBM, in training phase, two speaker model groups were built for male speaker and female speaker as figure 6. Each group will contain VQ model and GMM-UBM model for each of training speaker.

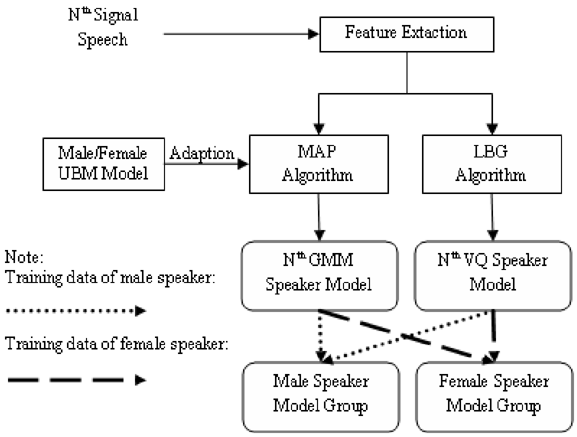

Fig 6. *Building of two speaker model groups*

4. EXPERIMENTAL SETUP AND RESULT

Speaker database was collected from 150 speakers (70 males, 80 females) who voices were recorded under the low-noise environment conditions. The audio files were recorded from Adobe Audition program, using PCM, sampling frequency was16000Hz, 16bit. The recording was done because of two purposes: preparing database for the training and identification processes.

- For the training process: in this research 100 people were recorded (50 males and 50 females), each one will 45 seconds.

- For the identifying process, testing database was taken from 150 people, including 100 people recorded in the training process who were identified to be the interested ones, the other 50.

In this paper, data was characteristically extracted with 39 characteristics per frame. For VQ, used size of codebook was 128. For GMM, model had 25 gaussian mixtures.

The time result of identifying process uses VQ, GMM-UBM and combination of VQ and GMM is shown in Figure.7. VQ based system had shortest calculating time when comparing with other models. This was the main advantage of model using VQ, but its accuracy is so low (Figure. 8). Although GMM model processed the computation for a long time, merrit of GMM model had higher accuracy. Therefore, with the idea of combining two merrits of the two previous models, time processing of VQ/GMM-UBM model was shortened (a reduction identification time up to 26% is reached) but system performance is still improved (Figure 7). The performance of VQ/GMM-UBM model was higher than GMM model due to the classifying way of testing speaker into male and female speaker. Thus, gender-dependent UBM model was used and accuracy rate of system was higher than gender-independent UBM model-based system (GMM-UBM model).

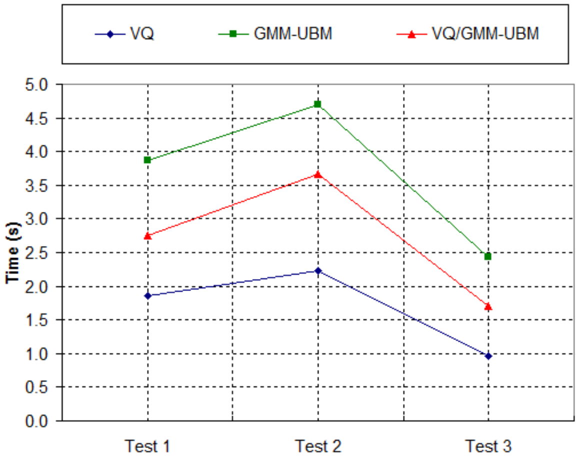

Fig 7. *Identifying time for each speaker with each testing database*

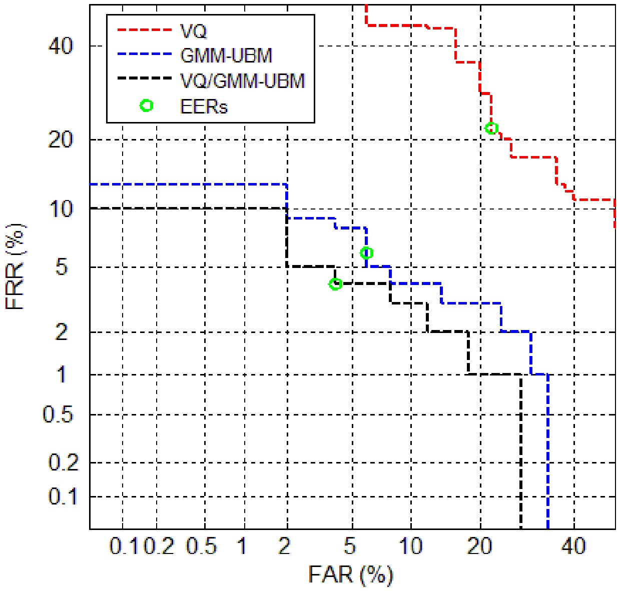

Fig 8. *DET curve of different modeling techniques*

5. CONCLUSIONS

In this paper, the combination of two techniques has been excuted. From obtained results, we observe that the combination approach between VQ and GMM is the good approach due to their different ways of classifying the data. With this combination, data was classified better in order to improve the calculating time as well as improve the system performance. Thus, the proposed model - VQ/GMM-UBM has been proven to be a powerful tool for text-independent speaker recognition system. It has successfully achieved the goal of this research which is solving the time consuming issue for GMM-UBM model.

REFERENCES

[1] Piyush Lotia, M.R. Khan (2011). Multistage VQ Based GMM For Text Independent Speaker identification System, *International Journal of Soft Computing and Engineering (IJSCE),* Vol. 1 (No. 2), pp 21-26.

[2] Joseph Campbell (1997). Speaker Recognition: A Tutorial, *Proceedings of IEEE,* Vol. 85 (No. 9), pp 1437-1462.

[3] Rafik Djemili, Mouldi Bedda, and Hocine Bourouba (2007). A Hybrid GMM/SVM System for Text Independent Speaker Identification, *World Academy of Science,* pp 448-454.

[4] Richard O. Duda, Peter E. Hart, David G. Stork (2001). Pattern Classification, *Willey Interscien,* 2nd.

[5] Douglas A. Reynolds, Richar C. Rose (1995). Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Model, *IEEE Transaction on speech and audio processing,* Vol. 3 (No 1), pp 72-83.

[6] Douglas A. Reynolds, Thomas F. Quatieri, Robert B. Dunn (2000). Speaker Verification Using Adapted Gaussian Mixture Models, *Digital Signal Processing,*Vol.10(1-3), pp.19-41.

[7] Tuan V. Pham, Michael Neffe, Gernot Kubin, Horst Hering (2007), "Speaker Segmentation for Air Traffic Control", *Speaker Classification II*, *LNAI 4441*, pp. 177-191.

[8] Tuan V. Pham, Michael Neffe, Gernot Kubin (2007), "Robust Voice Activity Detection For Narrow-Bandwidth Speaker Verification Under Adverse Environments", *Interspeech*, *ISSN: 1990-9772.*

[9] Tuan V. Pham (2008), *Wavelet Analysis for Robust Speech Processing and Applications,* VDM Verlag Dr. Muller Aktiengesellschaft & Co. KG, Dudweiler Landstr. *125 a.*