Using Vector Quantization in Automatic Speaker Verification

DJELLALI Hayet
Department of Computer Science, LRS Laboratory
Badji Mokhtar University
Annaba, Algeria

LASKRI Mohamed Tayeb
Department of Computer Science, LRI Laboratory
Badji Mokhtar University
Annaba, Algeria

Abstract— This article investigates several technique based on vector quantization (VQ) and maximum a posteriori adaptation (MAP) in Automatic Speaker Verification ASV. We propose to create multiple codebooks of Universal Background Model UBM by Vector Quantization and compare them with traditional approach in VQ, MAP adaptation and Gaussian Mixture Models.

Keywords-component; Automatic Speaker Recognition;Vector Quantization; Linde Buzo Gray Algorithm; False Rejection; False Acceptance; Gaussian Mixture Models; Impostor Models; Speaker Verification.

I. INTRODUCTION

Recognize a person by machine from their voice is the aim of Automatic Speaker Recognition. It has been proved that the variation factors like speakers identity, utterance length, gender, session, transmission channel affect the system performance[1][2][3].

In most of speaker recognition system an input speech utterance is compared to enrolled target speaker model, resulting in a similarity score. The target model is obtained from a set of training speech utterances from a known speaker [2].

In the verification mode, the system has to decide whether the identity of the speaker is the same as a claimed one or not. This output decision is generated by performing a trial with a test speech utterance and a speaker model representing the claimed identity. The score is then compared to a threshold to obtain the final decision accepted or rejected.

The state of the art of text independent speaker recognition is Gaussian mixture model. Speaker dependent GMM are derived from the speaker independent model called universal speaker model (UBM) and Maximum a posteriori adaptation MAP using target speaker speech data. It was proved that GMM-UBM is better for short utterance where VQ-UBM outperforms GMM-UBM in situation where test data and the length of training increase [4].

Vector Quantization (VQ) model was introduced in 1980, Its roots are originally in data compression[5]. VQ is one of the simplest text independent speaker model, and often used for computational technique. It also provides competitive accuracy when combined with background model adaptation [1, 4,5, 6].

VQ-based speaker recognition is a conventional and successful method. The VQ speaker model is often used as a baseline when studying other methods. In VQ based speaker recognition, each speaker is characterized with several code vectors, and the set of code vectors for each speaker is referred to as that speaker's codebook[4]. Normally, a speaker's codebook is trained to minimize the quantization error for the training data from that speaker. The most commonly used training algorithm is the Linde-Buzo-Gray (LBG) algorithm [5].

If the speaker speech data becomes huge, it faces the time consuming problem. Gurmeet et al replaced the EM algorithm with LBG algorithm. Experimentally, they found that the complexity of calculation can be reduced by 50% compared to the EM algorithm. The reason is the LBG algorithm utilize apart of feature vectors for classification. Gurmeet et al have successfully proved that LBG provided comparable performance to the EM algorithm and significantly decreased computational complexity [7],[8].

We applied Vector Quantization in Automatic Speaker Verification; Usually, each target speaker had his own codebook, when usually the speaker independent models had two gender dependent codebook originate from impostor speakers (male, female).

Our approach aim to select the best universal background model UBM, we try another way to model VQ UBM with set of sub UBM. We divide the features vectors extracted from processing step (Mel Cepstral Coefficients: MFCC) in a equal size and applied for each of them the LBG algorithm to obtain its codebook (cd1,cd2..,cdK).

The aim is to get the best sub model with LBG algorithm for impostors (UBM) and then compute the distortion error from optimal Sub UBM.

We aim to reduce EER in the presence of small training data of each client and select the best sub UBM. In addition to that, we propose reduced VQ-UBM which provides another UBM codebook by choosing the best one from different impostor session and finally compare their performances.

We organized paper as follows, modeling speakers based on vector quantization and MAP adaptation is introduced in Section 2, and the ASV architecture proposed in Section 3

978-1-4673-1166-3/12/$31.00 ©2012 IEEE

followed experiments in Section 4, discussion in section 5 and conclusion in 6.

II. VECTOR QUANTIZATION AND MAP ADAPTATION

A. *Vector Quantization*

Vector Quantization (VQ) is a pattern classification technique applied to speech data to form a representative set of speaker features. VQ is a process of mapping vectors from a large data space to a finite number of regions called a cluster and represented by its codeword(center). The codebook is a set of all code words.

It was introduced to speaker recognition by Soong et al (1985). In speaker verification, Vector quantization (VQ) model were applied in Soong and Rosenberg[1], It is one of the simplest text-independent speaker models and usually used for computational speed-up techniques, it also provides competitive accuracy when combined with background model adaptation [1][6][9][10].

In the training phase, a speaker-specific VQ codebook is generated for each known speaker by clustering his training acoustic vectors. The distance from a vector to the closet codeword of a codebook is called a VQ distortion [4][5].

In the Test phase, an input utterance of a known voice is "vector-quantized" using trained codebook from proclaimed identity and the speaker independent model codebook (Universal Background Model). The total VQ distortion is computed.

In principle, when we get a large amount of training vectors representing speaker in the training vectors. We should reduce it by vector quantization. Suppose there are N vectors, to be quantized, the average quantization error is given by:

$$E = \frac{1}{N}\sum_{t=1}^{N} e(x_t) \qquad (1)$$

The task of designing a codebook is to find a set of code vectors so that E is minimized. However, the commonly used method is the LBG algorithm [7].

The LBG trained codebook is optimal in the sense that the quantization error is minimized. In speaker verification, the codebook is used for classification and minimizing the quantization error.

We selected LBG algorithm defined as the iterative improvement algorithm or the generalized Lloyd algorithm. Given a set of N training feature vectors, {t1, t2,.tn} characterizing the variability of a speaker, we search a partitioning of the feature vector space, {S1, S2,..., SM}, for that particular speaker where S, the whole feature space, is represented as S = S1 U S2 U...U SM.

The performance of a quantizer is designed by an average distortion between the input vectors and the final vectors, where E represents the expectation operator (equation1).

B. *Gaussian Mixture Models and MAP Adaptation*

- GMM-UBM-Maximum Likelihood Modeling: this approach is based on training UBM male model with Gaussian mixture model and the other female UBM (from female speech). The model parameters (mean, covariance and weight of the Gaussian) are trained with the EM algorithm (Expectation-Maximization).
- Maximum a Posteriori approach MAP resolve the problem of maximum likelihood ML(can't generalize well to unseen speech data in low training data). MAP use prior knowledge of the distribution of the model parameters and insert it in modeling process[12][13]. The Maximum A Posteriori MAP approach is to use the world model and client training data to estimate the client model on the basis of his data and MAP Adaptation [12][13] [14][15].
- Two models are created, the client model based on his data and the acoustic model of the world UBM whose acoustic vectors are derived from a large population of speakers other than our target speakers and the GMM-UBM ML model, based on the estimation of Maximum Likelihood ML (Maximum Likelihood)[2][3][4].
- The client model is derived from the world model by adapting the GMM parameters (mean, covariance, weights) are estimated. However, experimentally, only the averages of GMM are adapted [13]. The maximization of average parameter is expressed as follows: for a Gaussian i of GMM, expressed as:

$$\mu_i = \alpha_i * \mu_i^{\prime} + (1 - \alpha_i) * \mu_i^{W} \qquad (2)$$

$\mu_i^{\prime}$ *is client mean* ; μ_i^{W} *is UBM mean* , α : is a weight that allows you to assign more or less weight, the parameters a priori by the parameters estimated on the training data. It is defined by:

$$\alpha_i = \frac{n_i}{n_i + \tau} \qquad (3)$$

ni number of frames assigned to a Gaussian i.

τ : The relevance factor, it controls the degree of adaptation of each Gaussian in terms of frames allocated.

III. SPEAKER VERIFICATION ARCHITECTURE BASED ON VECTOR QUANTIZATION

This new modeling approach is specified as Reduced UBM vector quantization VQ used as speaker independent model. It was modeled by a set of codebook generated by clustering.

We achieve data reduction by dropping out no significant speaker data in UBM. For computational reasons, the numbers of vectors are reduced.

We divide UBM speech data in N subsets instead of one global UBM, in figure 1, after features extraction, the MFCC

vectors are used as input for L.B.G algorithm which provide K codebooks.

We aim to optimize the score by improving the background model for that purpose we applied our algorithm2 in training stage.

A. *Training Phase*

We proposed two VQ-UBM, the first one is the baseline system, the second is Reduced VQ-UBM. We describe our new modeling UBM:

1) *Codebook*

There are several different approaches to finding an optimal codebook. The idea is to begin with a vector quantizer and a codebook and improve upon the initial codebook by iterating until the best codebook is found[16]. We aim to reduce redundancy in UBM data by clustering, to do that, we implement this algorithm:

2) *Algorithm 1:*

✓ *Training Phase*

1. Input : MFCC vectors
2. Output: Codebook CDU(1..M).
3. We divide MFCC vector in equal sub matrix and applied LBG algorithm for each of them.
   Input[C]= MFCC vector(Feature Extraction).

   Split C in M equal sub matrix Ci;
4. Train UBM of each Ci for different size of codebook(k=16,32,64,128,256);
   Result= CDU (i=1..M).

✓ *Test Phase*

5. In recognition phase, we compute Euclidean distance and evaluate quantization error from each codebook and test vector,
6. We choose the best codebook with minimal quantization error.

The quantization error MSE

$$MSE(X,Y) = \frac{1}{|x|}\sum_{xi} \min_{yk} ||Xi - yk||^2 \qquad (4)$$

Where $yk \in Y$; Xi : vector data; yK : centroid

3) *Algorithm 2: Reduced VQ UBM*

Input: MFCC vectors of each impostor (each session)
Output: NC codebook

For the same impostor belongs to UBM, we had different M sessions: 1..M and T enrollment for each session.

Impostor Speaker I: Codebook (I, j) where j=1..M

1. Compute codebook with LBG algorithm for every session enrollment(T)
2. Get couple of codebook constraint to minimal distance between sessions:

$$C_i^j = \min\{ d(C_i^j, C_i^k)\},\ j,k : \text{session} \}$$

3. Delete $C_i^k = \min\{ d(C_i^j, C_i^k)\}$, j, k session
4. NC= NC-1; NC: Number of codebook
5. Repeat 1 to 4 until condition;

Condition:➔ iteration = M div 2.

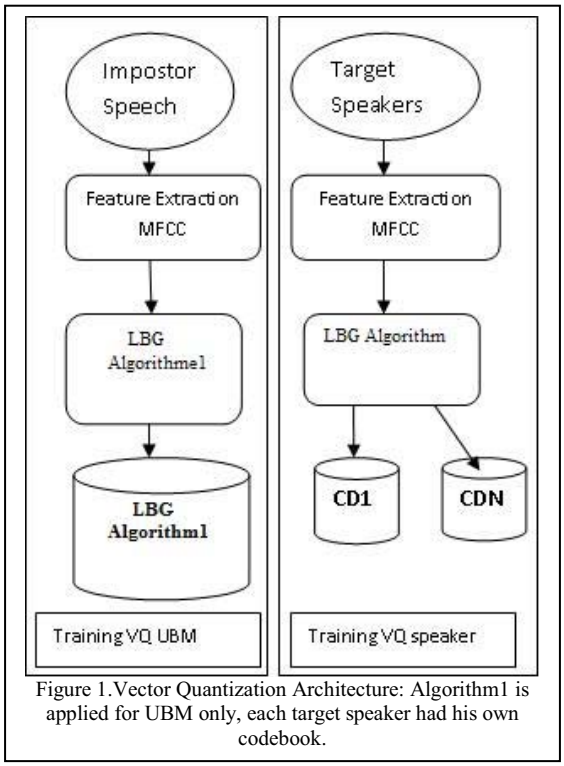

Figure 1.Vector Quantization Architecture: Algorithm1 is applied for UBM only, each target speaker had his own codebook.

B. *Test Phase*

We compute the threshold (CDT) from 8 male and 8 females speakers others than UBM speakers and trained by LBG algorithm.

Test algorithm

CDU: UBM codebook;
CDS: Speaker codebook;
VQdist : VQ distorsion
Input : X=speaker speech ,claimed identity
MFCC = Feature Extraction(X)

For i=1 to M
VQcdu=VQdistorsion(X,CDUi)
End

CDUoptimal=CDU best codebook UBM
where Argmin(VQdistorsion(X,CDUi))

VQdist(speaker)=VQdist(X,CDS)– Vqdist(X,CDUoptimal)

If VQdist(speaker)>VQdist(CDT) then client acces
Else reject

IV. PROTOCOL EXPERIMENT

In this section, we describe a set of experiments designed to evaluate the performance of the proposed system under a variety of condition and compare it to baseline system GMM MAP and standard VQ UBM.

A. Database and Baseline System

The Arabic database is recorded in Goldwave frequency 16KHz for a period of 60s for each speaker when training and 30s in the testing phase. The UBM population is 15 men's and 15 women. Four sessions are recorded for each speaker at an interval of 1 month. Ten clients are registered in the database (5 men and 5 women). The ASV reference system GMM-UBM-IG is independent gender obtained by merging the two models male and female speakers [1, 17].

B. Baseline VQ-UBM Model

We extract MFCC vector for all acoustics data allowed to UBM training and applied LBG algorithm for it. We obtain one centroid (NxT), where we try different value of k=16, 32, 64, 128, 256. In recognition phase, we compute Euclidean distance and evaluate quantization error (4) from centroid and test vector, we did the same thing for target speaker and finally evaluate the score.

TABLE I. BASELINE VQ UBM PERFORMANCES

CodeBook Size	FA(%)	FR(%)
CD32	14,29	73,03
CD64	**12,86**	**4,89**

C. VQ UBM Model (ALGORITHM 1)

We built UBM models from 30 arabic speakers; UBM male with 15 male speakers and UBM female from 15 female speakers. We evaluate LBG algorithm for k=16, 32, 64, 128.The global threshold is computed from other database: 8 male and 8 female speakers.

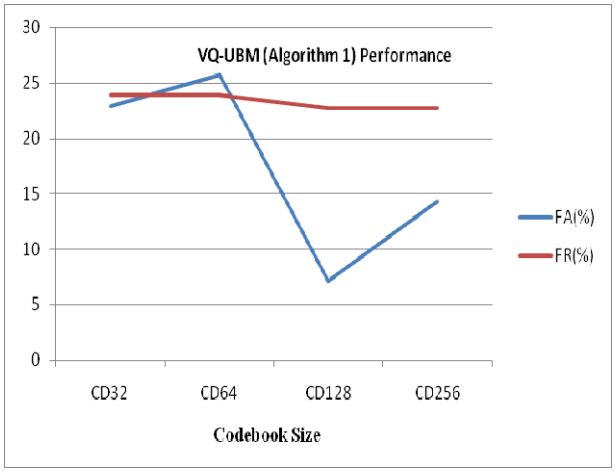

Figure 2. VQ UBM(Algorithm 1) Performance: 128 codebook size is the best

TABLE II. VQ-UBM (ALGORITHM 1) PERFORMANCES

CodeBook Size	FA(%)	FR(%)
CD32	22,86	23,86
CD64	25,71	23,86
CD128	**7,14**	**22,73**
CD256	14,29	22,73

D. Baseline GMM MAP system

We train universal background model UBM gender dependent(male, female) under expectation maximization algorithm EM and create each target speaker model with GMM MAP approach, we try different sizes of GMM (8, 16, and 32) and evaluate the value of false acceptance and false rejection[17].

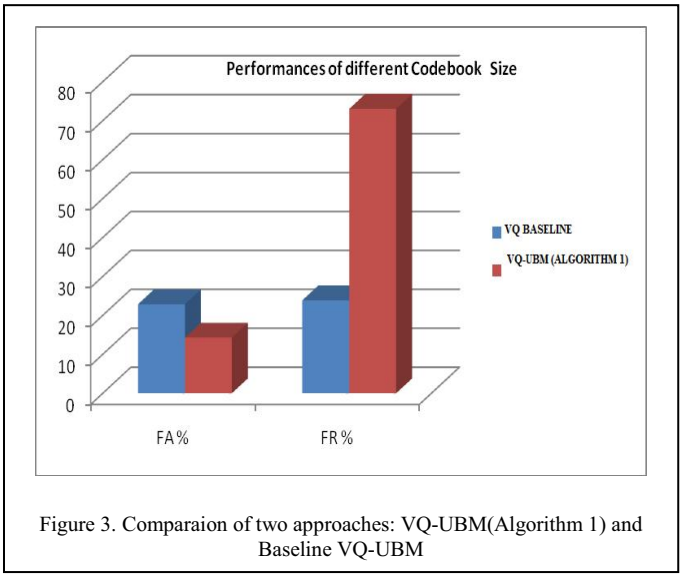

Figure 3. Comparaion of two approaches: VQ-UBM(Algorithm 1) and Baseline VQ-UBM

TABLE III. GMM MAP SYSTEM RESULT

Methods	EER%	#Mixtures
GMM MAP	**19,16**	**8**
	36,12	16
	35,20	32

V. DISCUSSION

We compare different modeling speaker techniques: VQ-UBM (Algorithm 1), baseline VQ-UBM and GMM MAP their performances were evaluated using the same data and front end processing.

Table I shows the value of false acceptance and false rejection for different codebook size (32,..256) in VQ-UBM (Algorithm 1) approach and observe that the best value is designed for 128 codebook size. The result in table 1 provide more accuracy recognition than table II, we observe that the size of codebook influence the performance and the multiple UBM provide better result.

Figure 2 demonstrates the performance of VQ-UBM (Algorithm 1) had the minimal error for 128 codebook size (7.14% and 22.73%).

In Baseline GMM MAP system, Equal error rate is 22.50%

for 32 mixtures and is not enough for well speakers modeling second we didn't apply normalization technique like Tnorm.

VI. CONCLUSION

VQ-UBM(Algorithm 1) achieved (FA=7.14% and FR=22.73%) for 128 codebook size and improved the performance of vector quantization applied in speaker verification compared to baseline vector quantization. The size of speech data should be increased in order to validate our experiments in large database.

REFERENCES

[1] A. E. Rosenberg and F. K. Soong, "Evaluation of a vector quantization talker recognition system in text independent and text dependent modes," Comput. Speech Lang., vol. 22, (1987). pp. 143–157.

[2] A. Preti,."Thesis 'Surveillance de réseaux professionnels de communication par la reconnaissance du locuteur. Académie d'Aix Marseille ,Laboratoire d'informatique d'Avignon(2008).

[3] F Bimbot, JF Bonastre, C Fredouille, G Gravier, , M. Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Petrovska-Delacretaz, D., Reynolds, D.A.. A tutorial on text-independent speaker verification. J. Appl. Signal Process. 4, (2005). 430−451.

[4] T.Kinnunen,J Saastamoinen,V Hautomaki, M Vinni,Pasi Franti, "Comparing Maximum a Posteriori Vector Quantization and Gaussian Mixture Models in Speaker Verification",Pattern recognition letters,(2008).

[5] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," IEEE Trans. Commun., vol. 20, (1980).pp.84–95.

[6] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech Audio Processing, vol. 3, (1995). pp. 72–83.

[7] J Campbell, Speaker recognition: a tutorial. Proc. IEEE 85 (9), (1997). 1437-1462.

[8] S Gurmeet ,S Panda,Bhattachryya S.Srikanthan S, "Vector Quantization technique for GMM Based Speaker Verification.IEEE International conference on acoustics,peech and signal processing,USA. (2003). 65-68.

[9] J S Pan, Thesis "Improved Algorithms for VQ Codeword Search, Codebook Design and Codebook Index Assignment",University of Edeinbourgh (1996).

[10] J He, Li Liu, and G¨unther Palm, "A Discriminative Training Algorithm for VQ-Based Speaker Identification".IEEE Transactions on Audio and Signal Processing,vol 7,(1999).

[11] D,A Reynolds, T.F Quatieri, Dunn, R.B. "Speaker verification using adapted Gaussian Mixture Models". Digital Signal Process. 10, (2000). 19−41.

[12] S Furui ,"Speaker-dependent-feature extraction, recognition and processing techniques" ,NTT Human interface Laboratories,Japan Speech Communication , Elsevier Science Publishers North-Holland (1991). 505-520.

[13] D Sturim, D A Reynolds, "Speaker adaptive cohort selection for Tnorm in text-independent speaker verification". In: Proc. of ICASSP, pp. (2005). 741−744.

[14] F K Soong ,A E Rosenberg ,L R Rabiner ,Juang B.H 1985,"Avector Quantization Approach to speaker recognition", IEEE International Conference on Acoustics ,speech and signal Processing,(1985). 387-390.

[15] B Vesnicer,F Mihelic.'The likelihood ratio decision criterion for nuisance attribute projection in GMM speaker Verification', Hindawi Publishing Corporation Eurasip Journal On advances in Signal Processing volume(2008).

[16] W C Chen, C T Hsieh, C-Hsu Hsu, "Robust Speaker Identification System Based on Two-Stage Vector Quantization", Tamkang Journal of science and engineering, Tamkang Journal of Science and Engineering, Vol. 11, No. 4, (2008). pp. 357_366.

[17] G Doddington, W Liggett, Martin, A., Przybocki, M., Reynolds, D.A.. Sheeps, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In: Proc. of ICSLP(1998).