Speaker Authentication Using Vector Quantization

Bushra Q. Al-Abudi* ***Mohammed S. Mahdi*****

Date of acceptance 15/4 / 2009

Abstract

In this paper, the role of the vector quantization in the speaker authentication system was studied. Vector quantization based speaker authentication system was considered in two phases; training and testing. The training phase concerned with enrolling the speaker models to build the codebook. The codebook generated from a set of feature vectors belong to each sample of speaker's voice. The testing phase includes matching the unknown input speaker with the models. The matching is performed by evaluating the similarity measure between the unknown speech sample and the models in the speaker database to authenticate the input speaker. A weighted similarity measure was introduced; it takes into regard the correlations between the known models in the database. Larger weights are assigned to vectors that have high discriminating power between the speakers and vice versa. The proposed system gave an encourage results; the authentication rate was about 86.6% during a time 4 *s*.

Key words: Speech processing, Speaker authentication, Vector quantization

Introduction:

Speaker recognition is a generic term used for two related problems; speaker identification and verification [1]. In the identification task the goal is to recognize the unknown speaker from a set of *N* known speakers. In verification, an identity claim (e.g., a username) is given to the recognizer and the goal is to accept or reject the given identity claim. In this work, the authentication concept concentrates on the identification task. The input of a speaker authentication system is a sampled speech data, and the output is the index of the identified speaker.

There are three important components in a speaker authentication system; the feature extraction component, the speaker models, and the matching algorithm. Feature extractor derives a set of speaker specific vectors from the input signal. Speaker model is then generated from these vectors for each speaker. The matching procedure performs the comparison of the speaker models. It is expected that the feature extraction is the most critical component of the system but it is also much more difficult part to be designed than the matching procedure [2]. Various features as well as speaker models have been proposed for speaker authentication. Classical features include the *cepstrum* with many variants [3, 4], and *line spectral frequencies* [5]. Recently, *subband processing* has also become a popular technique [6, 7, 8, 9, 10, 11]. Some of the various modeling techniques include *vector quantization* (VQ) [12, 13, 14, 15], *Gaussian mixture models (GMM)* [16, 17], *covariances models* [18], and *neural networks* [19]. In this paper, we will focus on the vector quantization (VQ) approach because it is fast to implement and computationally efficient. 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 represented by their central vectors or *centroids*. A set of

*Baghdad University/College of Science/Astronomy Department.
**Nahrain University/College of Science/Computers Department.

centroids, which represents the whole vector space, called a *codebook*. In VQ approach, the computation of matching score is based on similarity measure between the unknown speaker features vector and the model.

VQ-Based Speaker Authentication System:

In this work, we study the role of the vector quantization in a VQ-based speaker authentication system. The proposed system based on VQ contains offline training and online matching sub-systems. The training sub-system is used to produce VQ codebooks, while the testing sub-system is used to make authentication decision. In VQ-based system [20, 21, 22, 23], vector quantization is employed to determine the features set in both training and matching branches. In the training, the speaker models are constructed by clustering the feature vectors in K separate clusters. Each cluster is represented by a *code vector* $\mathbf{c}_i$, which is the centroid of the cluster. The resulting set of code vectors is called a *codebook*, and notated by $C(j) = \{\boldsymbol{c}_1(j), \boldsymbol{c}_2(j), ..., \boldsymbol{c}_K(j)\}$, where (j) denotes to speaker index, and K referes to the voice sample. In the codebook, each vector represents a single acoustic unit typical for the particular speaker. Thus, the distribution of the feature vectors is represented by a smaller set of sample vectors with similar distribution than the full set of feature vectors of the speaker model. In the matching or testing, the *distortion measure* $D(X, C(i))$ is computed to estimate the authentication rate between an unknown speaker feature vectors $X = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_T\}$ and all codebooks models $\{C(1), C(2), ..., C(N)\}$ [23].

Vector Quantization of the Feature Vectors:

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 represented by their central vectors or *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 referred as a speaker model. Mathematically a VQ task is defined as follows: given a set of feature vectors, find a partitioning of the feature vector space into the predefined number of regions, which do not overlap with each other and added together form the whole feature vector space. Every vector inside such region is represented by the corresponding centroid.

There are two important design issues in VQ: the method for generating the codebook and codebook size. In this work, we consider the following *modified* algorithm for codebook generation:

1-Given a set of feature vectors $X = \{\boldsymbol{x}_i \mid i = 1,...,L\}$,

2-partition the data set into $K << L$ clusters such that similar vectors are grouped together and vectors with different features belong to different groups. The codebook $C = \{\boldsymbol{c}_1, ..., \boldsymbol{c}_k\}$ can then be constructed from the cluster representatives, which are the vector averages of each cluster

3- Form an initial codebook by choosing the first N-input vectors as reproduction vectors

4-Compare each input vector with all N-reproduction vectors. Best match is achieved when the minimum mean square error (MSE) between the reproduction and the input vectors is

within a pre-specified threshold. In this case the input matched vectors should be given the same index of the reproduction vector. MSE formula is represented in the following:

$$MSE(X,C)=\frac{1}{N}\sum_{i=1}^{N}\min_{j}(d(x_i,c_j))^2 \qquad (1)$$

where X is a set of N extracted feature vectors, C is a speaker codebook, xi are feature vectors, c_j are codebook centroids and d is any of distance functions. The final identification decision is made based on the matching score: speaker who has a model with the smallest matching score is selected as an author of the test speech sample.

5- For each index, find the centriod of all input vectors. The centriods are the new codebook.

6- Sort the codebook vectors in descending order from high count to low count.

7- Eliminate the last reproduction vector, which has very low count and split the first reproduction vector (i.e., high count) into two vectors by multiplying the vector contents by enlargement/reduction factors (say, 1.1/0.9) to reproduce two new vectors.

8- The procedure repeats until the process converges to solution, which is a minimum of the total reproduction error.

In this work, set of speech file records contains a specific utterance pronounced by 30 speakers was used. Since the database has continuous speech with large vocabulary of recorded words, a common utterance that occurred frequently in the file was chosen for the speakers. This common voiced utterance corresponding to each speaker was extracted to be an input data into proposed speaker authentication system. The general structure of the proposed system was shown in Fig.1. In the following, more explanation about each stage will be discussed in details.

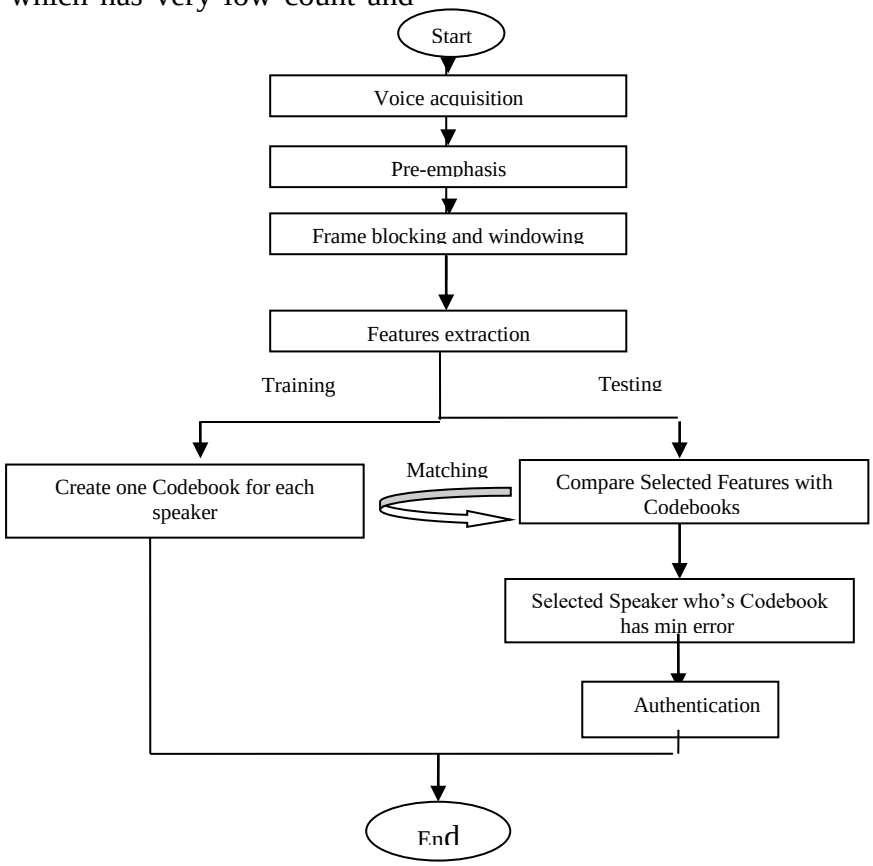

Fig. 1: The proposed speaker authentication system

A- Pre-Emphasis

The voice signal in this step is passed through a low-order FIR filter to spectrally flatten the signal and to make it less susceptible to finite precision effect. The transfer function of this filter is

$$H(t) = 1 - at^{-1} \qquad (2)$$

The value for *a* usually ranges from 0.9 to 1.0, at present work $a = 0.9375$.

B- Frame blocking and windowing

The signal is then put into frames (each 256 sample long). This corresponds to about 23 *ms* of speech per frame. Each frame put through a Hamming window. Windowing is used to minimize the discontinuities at the beginning and end of each frame. The Hamming window has the form:

$$W(m) = 0.54 - 0.46\cos\left(\frac{2\pi m}{M}\right); 0 < m < M\text{-}1 \qquad (3)$$

Where M is the number of samples per frame (i.e. M=256).

C- Features extraction

For purpose of authenticate speakers, VQ is used to give a set of feature vectors extracted from the voice sample. These features generate the speaker codebook in the training phase. Such codebook has significantly smaller size than extracted vector set and referred as speaker models.

D- Matching and Recognition

During the matching, the score of the matching is computed between extracted feature vector and every speaker codebook that enrolled in the system. The process of matching two speakers by VQ is represented in Figure 2.

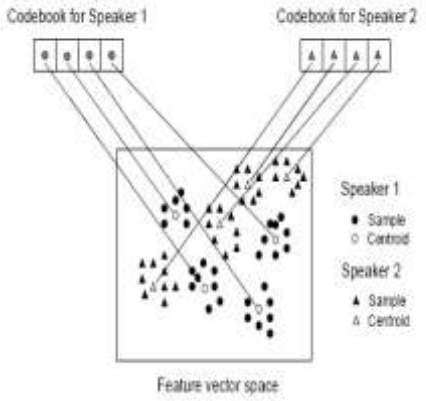

Fig. 2: Vector quantization of two speakers.

Experimental Results and Analysis:

The results extracted by this work reflected the perfect picture for auto authentication systems when regarding the accuracy and speed of the performance. The inputs of the system were speech wave files recorded for different speakers pronounce same utterance. Some of those speakers have models in the database, while others don't have. The test was beginning by input spoken records of 8 *s* long, from the implementation, the system show ability to perform the authentication task well and fast. The authentication rate was about 86.6% during spent time of 4 *s* when comparing with a codebook of 30 models.

The study of affecting the codebook size variation on the authentication rate indicated that the behavior of results was continues increasing till reaching a saturation stage. In the saturation, the authentication rate increased with increasing the codebook size as shown in fig 4. This indicates that the error rate (*MSE*) decreased slightly with increasing the codebook size. The *MSE* doesn't reaches zero value, but it

remains at the normal value licensed by the nature, at which the error was about 13.4% from the authentication rate as shown in Fig. 5. Also, it was very important to notice that how the frame size variation affecting the authentication rate. It was found the authentication rate is directly proportional to the frame size, since larger frame contains more information and lead to give a chance to detect the most similar model. Fig. 6 shows how the authentication rate increases by increasing the frame size. It very important to mentioned that, both the codebook size and frame size were significantly affecting the run times, which in turn determine the time of implementing the authentication task. It is noticeable the run times behavior start to be raised by increasing the codebook size as shown in Fig. 7. Then it conserves its progressing with monotonic behavior, since the differences between successive estimated clusters became small. Whereas the behavior of the run times with frame size variation was increasingly, but it takes instable behavior at higher frame size region as shown in Fig. 8. In general, the behavior of the run times was slightly increasing except when the frame size is greater than 128, there is a noticeable difference in the behavior of the run times. It was arise by a rate greater than the normal, which lead to doublet the time needed to complete the authentication task. In general, the proposed speaker authentication system was performed its task successfully, which ensure that the VQ method is appropriated in describing the speakers characteristics and verifying them.

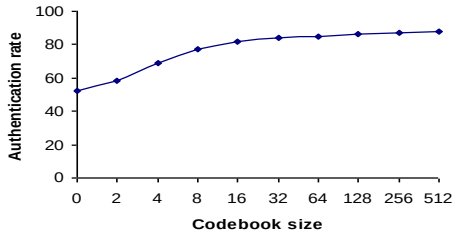

Fig. 4: Authentication rate as a function of codebook size.

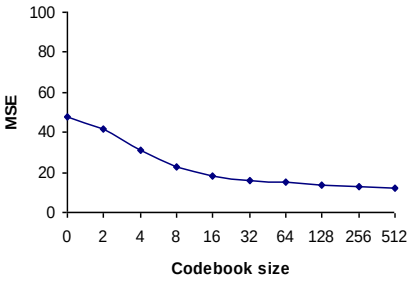

Fig. 5: Quality of the codebook as a function of codebook size.

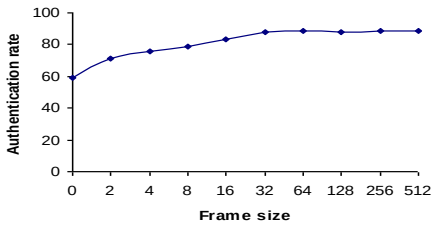

Fig. 6 Authentication rate as a function of frame size.

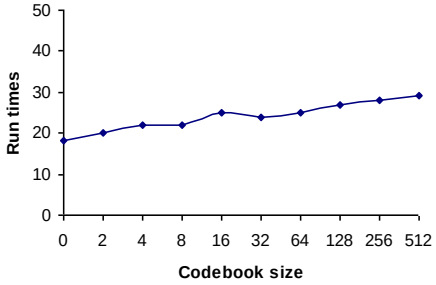

Fig. 7 Run times versus codebook size.

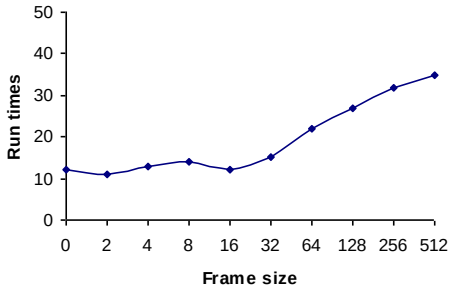

Fig. 8 Run times versus frame size.

Conclusion:

We evaluated the performance of VQ based speaker authentication**.** The easiest way for improving the authentication accuracy was to increase the codebook size high enough. No side-effect was observed due to the increase, except the increase in the running. Codebook size is a trade-off between running time and authentication accuracy. With large size, authentication accuracy is high but this cost more running time and vice versa. The codebook size must be selected carefully, typical speaker codebook size is around 64-512 vectors depending on the selected features.

References:

1- Furui S., **1997,** *Recent advances in speaker recognition.* Pattern Recognition Letters, 18: 859-872,

2- Rubust Q,,J., **2007,** *Speaker Recognition*,Thesis Submitted to School of Computer Science,Carenegie Mellon University

3- Atal B. **, 1974** *Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification*. Journal of the Acoustic Society of America, 55(6):1304–1312

4- Furui, S. **, 1981** "*Cepstral analysis technique for automatic speaker verification*," IEEE Transactions on Acoustics, Speech and Signal Processing, 29(2): 254-272

5- Campbell J. **, 1997** *Speaker recognition: a tutorial.* Proceedings of the IEEE, 85(9):1437–1462

6- Besacier L. and Bonastre J. **, 2000** *Subband architecture for automatic speaker recognition.* Signal Processing, 80:1245–1259

7- Besacier L. , Bonastre J., and Fredouille C.**, 2000,** *Localization and selection of speaker-specific information with statistical modeling*. Speech Communications, 31:89–106.

8- Kinnunen T.**, 2002,** *Designing a speaker-discriminative filter bank for speaker recognition.* In Proc. Int. Conf. on Spoken Language Processing (ICSLP 2002), pages 2325–2328, Denver, Colorado, USA.

9- Sivakumaran P., Ariyaeeinia A., and Loomes M.**, 2003** *Sub-band based text-dependent speaker verification.* Speech Communications, 41:485–509.

10- Damper R. and Higgins J.**, 2003,** *Improving speaker identification in noise by subband processing and decision fusion.* Pattern Recognition Letters, 24:2167–2173.

11- Ming J., Stewart D., Hanna P., Corr P., Smith J., and Vaseghi S.**,2003,** *Robust speaker identification using posterior union models*. In Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003), pages 2645–2648, Geneva, Switzerland.

12- Soong F.K., Rosenberg A.E., Juang B.-H., and Rabiner L.R., **1987**. A vector quantization approach to speaker recognition. *AT & T Technical Journal*, 66:14–26.

13- Kinnunen T., Kilpel¨ainen T., and Fr¨anti P., **2000** Comparison of clustering algorithms in speaker identification. In Proc. IASTED Int. Conf. Signal Processing and

Communications (SPC 2000), pages 222–227, Marbella, Spain.
14- Kinnunen T. and Fr¨anti P. **, 2001** Speaker discriminative weighting method for VQ-based speaker identification. In Proc. Audio- and Video-Based Biometric Authentication (AVBPA 2001), pages 150–156, Halmstad, Sweden**.**
15- Fan N. and Rosca J. **, 2003** Enhanced VQ-based algorithms for speech independent speaker identification.In Proc. Audio- and Video-Based Biometric Authentication (AVBPA 2003), pages 470–477, Guildford, UK.
16- Reynolds D.A. and Rose R.C.. **1995.** Robust text-independent speaker identification using gaussian mixture speaker models. IEEE Trans. on Speech and Audio Processing,3: 72–83.
17- Reynolds D.A., Quatieri T.F., and Dunn R.B. **, 2000**. Speaker verification using adapted gaussian mixture models.Digital Signal Processing, 10(1):19–41.
18- Bimbot F. and Mathan L. **, 1993** Text-free speaker recognition using an arithmetic-harmonic sphericity mesure. In Proc. 3th European Conference on Speech Communication and Technology (Eurospeech 1993), pages 169–172, Berlin, Germany
19- Farrell K.R., Mammone R.J., and Assaleh K. T.**, 1994**.Speaker recognition using neural networks and conventional classifiers. IEEE Trans. on Speech and Audio Processing, 2(1):194–205.
20- He, J., Liu, L., and Palm, G. **, 1999** "A discriminative training algorithm for VQ-based speaker identification," IEEE Transactions on Speech and Audio Processing, 7(3): 353-356.
21- Jin, Q., **2000**, A.: "A naive de-lambing method for speaker identification," Proc. ICSLP 2002, Beijing, China.
22- Kinnunen T., Kilpeläinen T., Fränti P.**, 2000**:"Comparison of clustering algorithms in speaker identification," Proc. IASTED Int. Conf. Signal Processing and Communications (SPC): 222-227, Marbella, Spain
23- Soong F.K., Rosenberg A.E., Juang B-H., and Rabiner, L.R., **1987** "A vector quantization approach to speaker recognition," *AT&T Technical Journal*, 66: 14-26,.

التحقق من هوية المتكلم باستخدام التكميم الاتجاهي

بشرى قاسم العبودي* ***محمد صاحب مهدي*****

*قسم الفلك/كلية العلوم/جامعة بغداد
**قسم الحاسبات/كلية العلوم/جامعة النهرين

الخلاصة:
في هذا البحث تم دراسة دور التكميم الاتجاهي في نظام التحقق من هوية المتكلم .تم دراسة التكميم الاتجاهي للتحقق من هوية المتكلم في طورين التجريبي والاختباري ,اهتم طور التجريب بتجميع موديلات المتكلم لبناء كتاب التشفيرالذي تم توليده من عدد من خواص المتجهات العائدة لكل نموذج من صوت المتكلم. اما طور الاختبار فيتضمن مطابقة المتكلم الغير معروف مع الموديلات.لقد انجزت عملية المطابقة من خلال تقييم مقياس التشابه بين نموذج المتكلم الغير معروف والموديلات في قاعدة بيانات المتكلم للتحقق منه.لقد تم استخدم مقياس وزن التشابه والذي ياخذ في نظرالاعتبار التطابق بين الموديلات الغير المعروفة في قاعدة البيانات حيث خصصت اوزان كبيرة للمتجهات التي لها قدرة تمييز عالية بين المتكلمين وبالعكس. اعطى النظام المقترح نتائج مشجعة حيث كان معدل التحقق حوالي 86% خلال زمن 4 ثواني.