Dynamic Segmental Vector Quantization in Isolated-word Speech Recognition

Vo Dinh Minh Nhat & Sungyoung Lee
Department of Computer Engineering, Kyung Hee University
Giheung-Eup, Yongin-Si, Gyeonggi-Do, 449-701, South Korea
vdmnhat@oslab.khu.ac.kr

Abstract *– The standard Vector Quantization (VQ) approach that uses a single vector quantizer for each entire duration of the utterance of each class suffers from the following two limitations: 1) high computational cost for large codebook sizes and 2) lack of explicit characterization of the sequential behavior. Both of two these disadvantages can be remedied by treating each utterance class as a concatenation of several information sub-sources, each of which is represented by a VQ codebook. With this approach, segmentation schemes obviously need to be investigated. And we call this VQ approach Dynamic Segmental Vector Quantization (DSVQ). This paper shows how to design DSVQ with some effective segmentation schemes. Better performances could be seen when applying this approach itself or mixed with Hidden Markov Model (HMM) in isolated-word speech recognition.*

Index terms *– Dynamic segmental vector quantization, segmentation scheme, speech recognition .*

I. INTRODUCTION

VQ is one of the very efficient source-coding techniques and has a lot of applications in many fields such as data compression, computer vision, speech recognition, and so on . In general, the key advantages of the VQ are reduced storage for feature analysis information, reduced computation for determining similarity, and discrete representation of speech sounds. However, it suffers from some disadvantages such as an inherent distortion in representing the actual analysis vector and the storage required for codebook vectors is often nontrivial [1].

In the field of speech recognition, VQ based recognizer is one of the most promising of the low cost recognizers , originally proposed by Shore and Burton [2], and modified by Burton et al. [4]. The basic idea in this recognition system is to design a separate VQ codebook for each word in the vocabulary, based on a training sequence of several tokens of each word by one or more talkers. In the original Shore and Burton implementation [2], the recognizer chose the word in the vocabulary whose average quantization distortion (according to its particular codebook) was minimum.

This word-based VQ recognizer worked very well for small vocabularies; however as the vocabulary size and/or complexity grew, the ability of the VQ processor to resolve among similar sounding words decreased dramatically , and the effectiveness of the recognizer similarly decreased.

The major problem with the VQ-based approach, for large vocabularies, was its inability to use temporal information; i.e. to integrate information about the times of occurrence of the speech sounds with the fact that the sounds occurred within the word [2]. Another thing is that the complexity of clustering algorithm increases rapidly when the size of codebook becomes higher and higher. One approach proposed in this paper to remedy these problems is incorporating this type of temporal information by using DSVQ with variety of segmentation schemes. In this approach, gross temporal information was incorporated into the recognizer by subdividing each input word into N_S, non-overlapping, regions, and using a separate codebook for each region. In this manner each word class was characterized by N_S codebooks, obtained from a training procedure in which a similar subdivision of each training class was made.

For convenience, we review the VQ approach in terms of mathematical symbols for later use. Some elements and symbols used in VQ are shown as follows :

A large set of analysis vectors $\{x_t\}_{t=1}^{T}$, which forms a training set. Each analysis vector is a k-dimensional vector.

Let $£ = \{y_i\}_{i=1}^{N}$ be a set of reproduction vectors (code words) and $d(x_t, y_i)$ be a prescribed distortion measure between the input x_t and the code word y_i. Here we mention Euclidean distance.

The codebook $£$ is designed to minimize $D = \frac{1}{T}\sum_{t=1}^{T} d(x_t, x_t^{\wedge})$ where $x_t^{\wedge} = \arg\min_{y_i \in £} d(x_t, y_i)$.

Some algorithms are used to create the codebook $£$ such as the generalized Lloyd algorithm or the K-means clustering algorithm, LBG algorithm, and so on.

The concept of vector quantization can be easily applied to speech-recognizer designs. Suppose there are M utterance classes (e.g., words, phrases) to be recognized. Each utterance class can be considered an information source. We thus collect M sets of training data $\{x_t^{(i)}\}$, where i = 1, 2, …, M is the class index. Each training set should contain a number of utterances of the same class. M codebooks $\{£^{(i)}\}_{i=1}^{M}$ are then designed for the M information sources, respectively. And we can see that each codebook represents a characterization of the information source (class).

During the recognition operation, the M codebooks are used to implement M distinct vector quantizers as shown in *Fig. 1*. An unknown utterance $\{x_t\}_{t=1}^{T_u}$ is vector-quantized by all M

[1] This research was partially supported by ITRC project of Sunmoon University

quantizers, resulting in M average distortion score $D(£^{(i)})$, i = 1, 2, …, M, where

$$D(£^{(i)}) = \frac{1}{T_u}\sum_{t=1}^{T_u} d(x_t, x_t^{\wedge(i)}) \quad (1)$$

with $x_t^{\wedge(i)} \in £^{(i)}$ satisfying

$$x_t^{\wedge(i)} = \arg\min_{y_j \in £^{(i)}} d(x_t, y_j^{(i)}) \quad (2)$$

The utterance is recognized as class k if

$$D(£^{(k)}) = \min_i D(£^{(i)}) \quad (3)$$

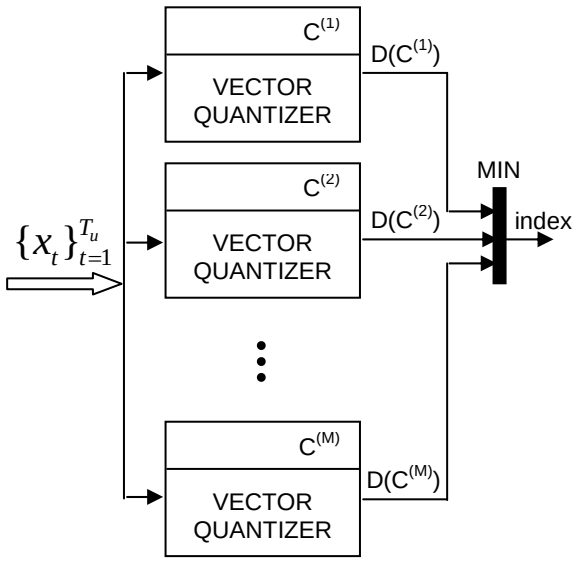

Fig. 1 : A vector-quantizer-based speech recognition system

This standard VQ approach that uses a single vector quantizer for the entire duration of the utterance for each class is not designed to preserve the sequential characteristics of the utterance class. And this will degrade the performance of recognition system because the characteristic of speech signal is sequential. This lack of explicit characterization of the sequential behavior can be remedied by using DSVQ. With DSVQ we can keep track with the temporal order from the sequential codebooks of each utterance class because they correspond to different portions of the utterance class. Therefore the performance will be better in case of large vocabulary.

Another thing we notice is the matter of complexity of the standard VQ in case of large codebook sizes. As we know, the K-means and LBG are very popular algorithms and ones of the best for implementing the clustering process. These algorithms have a time complexity that is dominated by the product of the number of training patterns (T), the number of clusters (L), and the number of iterations (I). Intuitively, we can see that the complexity of clustering algorithm is O(TLI). So when the number of training patterns and codebook size become large the computational cost of the standard VQ is also very high.

By using DSVQ , we give an effective way to decompose each utterance into a concatenation of N_s sub-sources. Each sub-source class will be represented by a VQ codebook. Based on this method we can reduce the complexity of clustering algorithms. Suppose we divide each utterance into N_s segments, so the complexity of clustering algorithm in DSVQ is $N_S xO(\frac{T}{N_S}\frac{L}{N_S}I)$. We can see that the DSVQ's complexity is reduced by N_S comparing to the standard VQ's. The rest of the paper is organized as follows: In Section II, we describe DSVQ techniques with some segmental schemes. An implementation and simulation results are provided in Section III. Finally, Section IV provides a summary of conclusions.

II. DYNAMIC SEGMENTAL VECTOR QUANTIZATION

This part comes to the detailed design of the DSVQ approach. For an utterance $\{x_t\}_{t=1}^{T_u}$, the simplest (but not the best) way to decompose it into a concatenation of N_S information sub-sources is to equally divide the utterance into N_S segments $\{x_t\}_{t=1}^{T_u/N_S}, \{x_t\}_{t=T_u/N_S+1}^{2T_u/N_S}, \ldots$ and so on. Other, more sophisticated, segmentation schemes obviously are possible and need to be studied. So we start by giving some definitions of basic elements of a segmentation scheme. A segmentation scheme is characterized by the following:

- N_s be the number of segments .
- α_i : the portion of the number of analysis vectors in segment i^{th} . Here we only need to define α_i with i = 1,2, …, N_s – 1, because the final portion can be inferred easily.
- k_i : the end index of segment i^{th}. We define $k_0 = 0$, $k_{N_S} = T_u$ and $k_i = \lfloor \alpha_i T_u \rfloor$ with i = 1,2, …, N_s – 1.
- Given an unknown utterance $U = \{x_t\}_{t=1}^{T_u}$, we define $S_i(U) = \{x_t\}_{k_{(i-1)}+1}^{k_i}$ as the sub-set of analysis vectors in segment i^{th} of utterance U.
- w_i : the weighted factor of each segment. These factors are used when we calculate the average distortion scores $D(£^{(i)})$.

The Table 1 shows some proposed segmentation schemes.

It can be seen from above discussion that a complete specification of an segmentation scheme requires specifications of N_S , (N_S – 1) element set $\{\alpha_i\}$, and N_S element set $\{w_i\}$.

Next, we will show how the segmentation scheme works with DSVQ.

Suppose we need to classify M utterance classes. Each class has N utterances which include U_1, U_2,…, U_N and each utterance is represented by an analysis vector set $U_i = \{x_t\}_{t=1}^{T_i}$.

Parameters	N_S	α_i	w_i
Scheme 1	3	{1/3, 1/3}	{1/3, 1/3, 1/3}
Scheme 2	3	{1/5, 3/5}	{1/3, 1/3, 1/3}
Scheme 3	3	{1/5, 3/5}	{1/6, 2/3, 1/6}
Scheme 4	5	{1/5, 1/5, 1/5, 1/5}	{1/5, 1/5, 1/5, 1/5, 1/5}
Scheme 5	5	{1/5, 1/5, 1/5, 1/5}	{1/9, 2/9, 1/3, 2/9, 1/9}

Table 1 : some proposed segmentation schemes

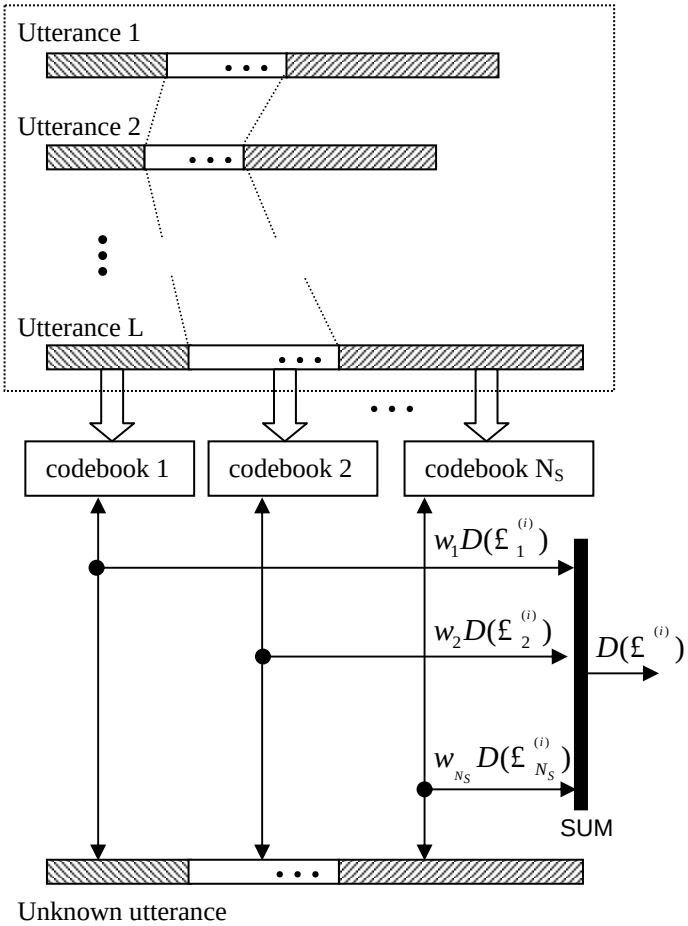

Fig. 2 : Calculate $D(£^{(i)})$ in DSVQ

For each class i^{th} , we have following things :

- Training set of segment j^{th} , j = 1, 2,…, N_S :

$$Tr_j^{(i)} = \prod_{k=1}^{N} S_j(U_k) \quad (4)$$

- Codebook of segment j^{th} is symbolled $£_j^{(i)}$. And from a class i^{th} we have a set of segmental codebooks $£^{(i)} = \{£_1^{(i)}, £_2^{(i)}, \ldots, £_{N_S}^{(i)}\}$

In recognition stage, an unknown utterance $U_u = \{x_t\}_{t=1}^{T_u}$ is vector-quantized by all M sets of quantizers (each set has N_s sub-quantizers), resulting in M average distortion score $D(£^{(i)})$, i = 1, 2, …, M, where

$$D(£^{(i)}) = \sum_{j=1}^{N_S} w_j D(£_j^{(i)}) \quad (5)$$

with $D(£_j^{(i)})$ is the average distortion score between $S_j(U_u)$ and sub-codebook $£_j^{(i)}$ and w_j is the weighted factors in segmentation scheme. We can see *Fig. 2* for an intuitive view of the calculation of $D(£^{(i)})$.

Finally, the utterance is recognized as class k if $D(£^{(k)}) = \min_i D(£^{(i)})$.

III. EXPERIMENTAL RESULTS

This section presents an evaluation of the proposed DSVQ approach with some segmentation schemes based on several experiments involving two speech databases : 1) our own database and 2) Alphadigit Corpus database of Corpora group at CSLU.

- Our own database is 50-word Vietnamese vocabulary database, each word is spoken 100 times by 5 Vietnamese people.

- The Alphadigit Corpus is a collection of 78,044 examples from 3,025 speakers saying six digit strings of letters and digits over the telephone.

In the feature extracting stage , we use Mel-frequency cepstral coefficients and human factor cepstral coefficients for testing. The test protocol was the same for all experiments. The parameters used are : number of cepstrum coefficients : 26 (12 order HFCC, 1 energy and 13 delta cepstral coefficients), number of filters : 20. Standard VQ approach, DSVQ approach, VQ used with Hidden Markov Model (HMM) approach and DSVQ/HMM approach are tested with two above databases. Below are some resulting tables.

Experimental results with the Vietnamese database

Approach	*% Correct*	
	MFCC	***HFCC***
VQ and DSVQ approach		
VQ	75.8	76.3
DSVQ Scheme 1	76.1	76.5
DSVQ Scheme 2	78	78.3
DSVQ Scheme 3	78.6	79.6
DSVQ Scheme 4	79.4	79.9
DSVQ Scheme 5	79.3	79.8
VQ/HMM and DSVQ/HMM approach		
VQ/HMM	63.2	65.4
DSVQ/HMM Scheme 1	67.5	68.1
DSVQ/HMM Scheme 2	67.9	69
DSVQ/HMM Scheme 3	68.7	69.5
DSVQ/HMM Scheme 4	70.1	72.3
DSVQ/HMM Scheme 5	69.9	71.5

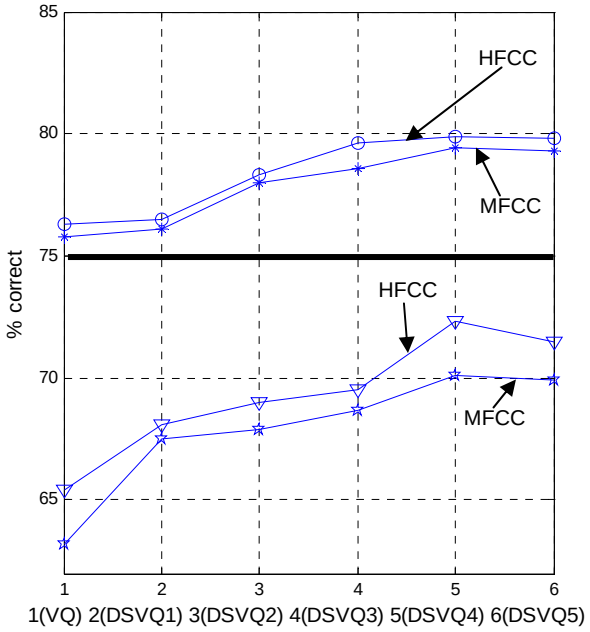

Fig. 3 : Recognition error rate versus VQ approach and DSVQ approach (with segmentation scheme 1, 2, 3, 4 and 5) in 50 word vocabulary vietnamese database

We can see the results clearly from the graphs in *Fig. 3*. The upper part is "*VQ and DSVQ approach"* with MFCC and HFCC. And the lower part comes with "*VQ/HMM and DSVQ/HMM approach*" . From those graphs , we can see the better results when applying DSVQ and by changing the segmentation scheme we can choose the best scheme for our specific speech database.

The same discussion when we take a look at the graphs in *Fig. 4* which shows the results in case of the Alphadigit Corpus database. And through the experimental results we can see the domination of HFCC comparing to MFCC.

Experimental results with the Alphadigit Corpus database

Approach	*% Correct*	
	MFCC	***HFCC***
VQ and DSVQ approach		
VQ	84.4	85.6
DSVQ Scheme 1	87.5	87.9
DSVQ Scheme 2	89.2	90.2
DSVQ Scheme 3	89.4	90.9
DSVQ Scheme 4	90.2	92
DSVQ Scheme 5	90	91.5
VQ/HMM and DSVQ/HMM approach		
VQ/HMM	65.2	66
DSVQ/HMM Scheme 1	68.1	68.4
DSVQ/HMM Scheme 2	68.9	69.7
DSVQ/HMM Scheme 3	69.6	70.8
DSVQ/HMM Scheme 4	71.3	72
DSVQ/HMM Scheme 5	71	71.2

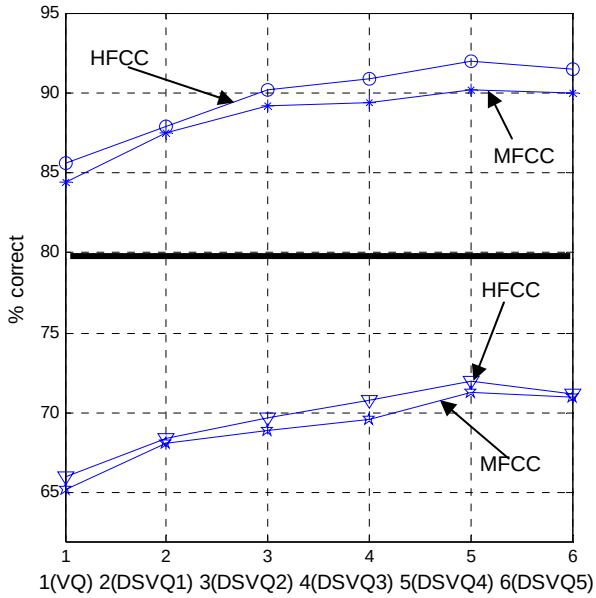

Fig. 4 : Recognition error rate versus VQ approach and DSVQ approach (with segmentation scheme 1, 2, 3, 4 and 5) in the Alphadigit Corpus database

From the above experiment results, we can see the improvement on performance of recognition system when applying DSVQ.

IV. CONCLUSION

We have introduced the DSVQ for designing an isolated word recognition system. By creating the segmentation scheme as an independent design parameter, DSVQ allows one to increase recognition system performance by choosing the suitable segmentation scheme. The flexibility of DSVQ helps us to deal with many kind of applications . Some other

advantages of DSVQ are the reducing of computational cost and getting better performance in case of large vocabulary.

REFERENCES

[1] Lawrence Rabiner, Biing-Hwang Juang "Fundamentals of speech recognition" 1993 - Prentice-Hall, Inc.

[2] J. E. Shore and D. K. Burton, "Discrete Utterance Speech Recognition without Time Alignment," IEEE Trans. On Information Theory, Vol. IT-29, No. 4, pp. 473-491, July 1983.

[3] L. R. Rabiner, S. F. Levinson, and M. M. Sondhi, "On the Application of Vector Quantization and Hidden Markov Models to Speaker-Independent, Isolated Word Recognition," Bell Syst. Tech. J., Vol. 62, No. 4, pp. 1075-1105, April 1983.

[4] D. K. Burton, J. T. Buck, and F. Shore, "Parameter Selection for Isolated Word Recognition Using Vector Quantization," Proc. ICASSP 84, San Diego, CA, pp. 9.4.1- 9.4.4, March 1984.

[5] M. D. Skowronski and J. G. Harris, "Human factor cepstral coefficients," IEEE Trans. Speech and Audio Processing, Submitted July 2002.

[6] Mark D. Skowronski, John G. Harris, "Improving The Filter Bank Of A Classic Speech Feature Extraction Algorithm," in the IEEE International Symposium on Circuits and Systems, Bangkok, Thailand, vol IV, pp 281-284, ISBN: 0-7803-7761-3, May 25 - 28, 2003.

[7] Wei-Wen Hung and Hsiao-Chuan Wang, 2001 March, "On the use of weighted filter bank analysis for the derivation of robust MFCCs," in IEEE Signal Processing Letters (SPL). vol. 8, no. 3. (SCI).

[8] Minh N. Do - Digital Signal Processing Mini-Project "An Automatic Speaker Recognition System".

[9] Hubert Wassner, Gerard Chollet "New Cepstral Representation Using Wavelet Analysis And Spectral Transformation For Robust Speech Recognition" Proc. ICSLP '96