Binary Quantization of Feature Vectors for Robust Text-Independent Speaker Identification

Zhong-Xuan Yuan, *Associate Member, IEEE,* Bo-Ling Xu, and Chong-Zhi Yu

***Abstract*— In this paper, we present a novel approach to vector quantization in which a feature vector is represented by a binary vector. It is called *binary quantization* (BQ). The performance criterion of vector quantization, distortion (distance) measure, was employed for investigating the effectiveness of BQ. At 12 b/analysis frame, the average distortion caused by BQ is even lower than the intraspeaker average distance between two repetitions of the same word (after DTW alignment). Since the output of BQ is a binary sequence, it is possible to combine it with forward Hamming net classifier. In terms of the idea of hierarchical model for describing a speaker individual characteristics, a text-independent speaker identification system was set up. Experimental results show that the performance of this system is very good. Not only are the small memory space and little computation required, in the speaker identification system, but, more importantly, it shows strong robustness in additive Gaussian white noise.**

***Index Terms*—Feature vector, Hamming net, speaker identification, vector quantization.**

I. Introduction

VECTOR quantization (VQ) is a very efficient compression technique which permits low rate speech coding. Over the past decades, it has been widely exploited. The well-known technique of VQ is called the Linde–Buzo–Grey (LBG) algorithm [1], which is based on two necessary conditions for optimality: the centroid and the nearest neighbor conditions. The clustering algorithm is used to obtain a finite reproduction alphabet (codebook), $\hat{\boldsymbol{x}} = \{\boldsymbol{x}_m; m = 1, \cdots, \boldsymbol{M}_{\mathrm{VQ}}\}$, representative of a long training sequence of speech spectrum. The codebook is designed to minimize the average distortion that results from representing this training sequence. VQ has been widely applied to both low bits vocoders [2], [3] and speech (including speaker) recognition systems [4]–[9], [19], [20]. By using VQ, the amount of data to be processed is reduced.

To improve the robustness of a speaker recognition system and further decrease computational complexity and storage, we propose a new approach to quantize a feature vector, which is represented by a binary vector. We call this kind of vector quantization *binary quantization* (BQ). Combined BQ with forward Hamming net, a new model of speaker recognition is proposed. The model is partially motivated by the success of VQ-based speech recognition [9] and speaker recognition [7]. In the former case, Rabiner *et al.* used a vector quantizer to transfer the continuous set of the feature vectors (LPC) into a finite observation set (the indexes of codebook entries). The discrete and finite set of observations were used as the input data of the hidden Markov model (HMM) recognizer. In the later case, Buck *et al.* used multisection VQ codebooks to replace their previous VQ codebooks [8] for isolated word recognition (IWR) and text-dependent speaker recognition [4], [5]. In this paper, BQ is used to find a partitioning of the multidimensional feature vector space $\{\boldsymbol{S}_0, \boldsymbol{S}_1, \cdots, \boldsymbol{S}_{\boldsymbol{M}_{\mathrm{BQ}-1}}\}$, the whole vector space is represented as $\boldsymbol{S} = \boldsymbol{S}_0 \cup \boldsymbol{S}_1 \cup \cdots \cup \boldsymbol{S}_{\boldsymbol{M}_{\mathrm{BQ}-1}}$. Each partition $\boldsymbol{S}_m$ forms a nonoverlapping hypercube and every vector inside $\boldsymbol{S}_m$ is represented by a binary vector $\boldsymbol{b}_m$. Not only does a binary vector $\boldsymbol{b}_m$ represent every vector inside $\boldsymbol{S}_m$, but also the decimal value of binary sequence of $\boldsymbol{b}_m$ equals the index m of partition $\boldsymbol{S}_m$. The method of BQ integrates the binary representation of every vector inside the partition $\boldsymbol{S}_m$ with the index of the partition and separates the representation of the vectors inside the partition from the centroid $\hat{\boldsymbol{x}}_m$ of the partition. The special property of BQ makes it possible to be used as the input of forward Hamming net classifier. The centroid vector $\hat{\boldsymbol{x}}_m$ of $\boldsymbol{S}_m$ is only used as the reproduction vector, a substitute for each vector inside. The set of centroid vectors $\boldsymbol{x}_m$, $m = 0, 1, \cdots, \boldsymbol{M}_{\mathrm{BQ}}$, will be used to investigate the average distortion caused by BQ. We know that the characteristics of the speaker's individual feature reside in his/her word voice. In small text-fixed vocabulary, such as digits, a speaker's template is described by the acoustical vectors derived from every word, each word is represented by a time-ordered sequence of binary vectors. A hierarchical model is employed to implement text-independent mode speaker identification in digital vocabulary.

Three kinds of feature parameters, cepstral features derived from linear predictive coefficients (LPCC), mel-frequency cepstrum coefficients (MFCC), and line spectrum pair (LSP), are chosen for investigating the effectiveness of BQ and BQ-forward Hamming net (BQHN) based speaker identifier. LPCC is chosen because of its effectiveness in speaker recognition [10] and its widespread uses in speech recognition. MFCC is chosen because it is calculated using a filterbank approach in which the set of filters are equal bandwidth with respect to the mel-scale of frequencies. This is because human perception of the frequency content of sounds does not follow a linear scale. MFCC is also a robust feature parameter [13]. LSP frequencies have been extensively used as an equivalent

Manuscript received April 25, 1996; revised January 27, 1998. This work was supported by the National Natural Science Foundation of China under Grant 69872014. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Joseph Picone.

The authors are with the Institute of Acoustics, State Key Laboratory of Modern Acoustics, Nanjing University, Nanjing 210093, China (e-mail: zxyuan@nju.edu.cn).

Publisher Item Identifier S 1063-6676(99)00176-5.

representation of the speech spectrum, which is modeled as the LPC technique in low bit rate speech coding systems because of their favorable properties in terms of stability, distribution, and spectral sensitivity. From normalized LSP frequency histograms [11], [12], it is clear that the distribution range varies from one LSP frequency histogram to another. It has also been observed by other authors that the distribution ranges vary with the speaker's individual characteristics [14]. In our previous study [15], the effectiveness of the absolute values of LSP frequencies used as speaker's individual features has been investigated by using F ratio formula [21]. Based on that work, we proposed an approach to a text-independent speaker identification system in which fuzzy mathematical algorithm was combined with functional-link networks [17].

The rest of this paper is organized as follows. In the first part of Section II, the principle of VQ is briefly summarized, then an explanation of the method of BQ is given. In Section III, the average distortions caused by BQ are calculated for investigating its performance. To establish baselines for comparison, the average distortions caused by VQ and the intraspeaker average distances of repetition utterances of same word are calculated. In Section IV, we describe the speaker identification scheme in which the method of BQ is combined with the forward Hamming net. The classifier is called the BQHN-based speaker identifier. In Section V, a group of experiments is performed to investigate the performance of the BQHN-based speaker identifier. In Section VI, the computation and storage required in the BQHN-based speaker identifier are evaluated roughly. In Section VII, we summarize and conclude this paper.

II. Binary Quantization

Before the explanation of binary quantization is given, we briefly review the principle of vector quantization.

A. Theory of Vector Quantization

Assume we have a training set of feature vectors, $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_1\}$, which could be LPCC, MFCC, or LSP from front-end analysis. The data in training set are a good representation of the types of feature vectors that occur when the words in the vocabulary are uttered by a wide range of speakers. The main idea behind vector quantization is to determine the optimum set of codebook entries $\hat{\boldsymbol{x}}_m$ (centroid vector), $m = 1, 2, \cdots, \boldsymbol{M}_{\mathrm{VQ}}$, such that for a given $\boldsymbol{M}_{\mathrm{VQ}}$, the average distortion in replacing each of the training set vectors $\boldsymbol{x}_i$ by the nearest codebook entry $\hat{\boldsymbol{x}}_m$ is minimum.

More formally stated, we define $d(\boldsymbol{x}_m, \boldsymbol{x}_i)$ as the local distance between two vectors, $\boldsymbol{x}_m$ and $\boldsymbol{x}_i$. In this paper, the local distance is the Euclidean distance

$$d(\boldsymbol{x}_m, \boldsymbol{x}_i) = \sqrt{\sum_{j=1}^{J} (\boldsymbol{x}_{mj} - \boldsymbol{x}_{ij})^2} \tag{1}$$

where J is the number of components in a feature vector. In training phase, the training algorithm is used to find a set of partitions of the feature vector space, $\{\boldsymbol{S}_1, \boldsymbol{S}_2, \cdots, \boldsymbol{S}_{\boldsymbol{M}_{\mathrm{VQ}}}\}$, S is the whole feature vector space which is represented as $\boldsymbol{S} = \boldsymbol{S}_1 \cup \boldsymbol{S}_2 \cup \cdots \cup \boldsymbol{S}_{\boldsymbol{M}_{\mathrm{VQ}}}$. Each partition $\boldsymbol{S}_m$ forms a convex, nonoverlapping region, and every vector inside $\boldsymbol{S}_m$ is represented by the centroid vector $\hat{\boldsymbol{x}}_m$ of $\boldsymbol{S}_m$. The set of partitions is done in such a way that the average distortion

$$\|D_M\| = \min_{\hat{\boldsymbol{x}}_m} \left\{ \frac{1}{I} \sum_{i=1}^{I} \min_{1 \le m \le M} [d(\hat{\boldsymbol{x}}_m, \boldsymbol{x}_i)] \right\} \tag{2}$$

is satisfied over the whole training set. The quantity $\|D_M\|$ is the average distortion (distance) of the vector quantization over the training set.

One way in which (2) is solved, for a given value of $\boldsymbol{M}_{\mathrm{VQ}}$, could be found in papers [1], [18]. The splitting algorithm starts with a one-level ($\boldsymbol{M}_{\mathrm{VQ}} = 1$) quantizer consisting of the centroid vector of training sequence. The centroid vector is then split into two vectors and the two-level ($\boldsymbol{M}_{\mathrm{VQ}} = 2$) quantizer algorithm is run on this pair to obtain an optimum two-level quantizer. Each of these two vectors is then split, and the algorithm is run to produce an optimum four-level ($\boldsymbol{M}_{\mathrm{VQ}} = 4$) quantizer and so on. This procedure iterates until $\boldsymbol{M}_{\mathrm{VQ}}$ is as large as desired.

B. Method of Binary Quantization

Given a training set of feature vectors $\boldsymbol{x}_i$, $i = 1, 2, \cdots, I$, in the original multidimensional Cartesian vector space, the centroid vector $\boldsymbol{x}_c$ is derived from the training set

$$\boldsymbol{x}_{cj} = \frac{1}{I} \sum_{i=1}^{I} \boldsymbol{x}_{ij} \quad 1 \le j \le J, \tag{3}$$

where J is the number of components in a feature vector. $\boldsymbol{x}_c j$ is the jth component of $\boldsymbol{x}_c$ and I is the number of vectors in the training set. Then the origin of the coordinates is moved to $\boldsymbol{x}_c$. The new coordinate system, center-of-mass coordinates, is produced. There are J coordinate planes in the new Cartesian vector space. The J coordinate planes, together with the maximum value and minimum value of each component of feature vectors form a set of partitions in the new Cartesian vector space, $\{\boldsymbol{S}_0, \boldsymbol{S}_1, \cdots, \boldsymbol{S}_{\boldsymbol{M}_{\mathrm{BQ}-1}}\}$. It is obvious that $\boldsymbol{M}_{\mathrm{BQ}}$ is determined by J and $\boldsymbol{M}_{\mathrm{BQ}} = 2^J$. Each partition $\boldsymbol{S}_m$ forms a hypercube, nonoverlapping region. Every feature vector inside $\boldsymbol{S}_m$ is represented by a binary vector $\boldsymbol{b}_m$. The decimal value of the binary sequence of $\boldsymbol{b}_m$ equals the index m of partition $\boldsymbol{S}_m$. In speaker identification system, a feature vector from front-end analysis is preprocessed and represented by a binary vector. The binary vector is used as the input of forward Hamming net identifier. In order to achieve a suitable binary representation $\boldsymbol{b}_i$ for a vector $\boldsymbol{x}_i$, each component of the vector, $\boldsymbol{x}_{ij}$, is compared with the corresponding component of $\boldsymbol{x}_c$, $\boldsymbol{x}_{cj}$

$$\boldsymbol{b}_{ij} = \begin{cases} 1 & \boldsymbol{x}_{iij} - \boldsymbol{x}_{cj} \ge 0 \\ 0 & \boldsymbol{x}_{ij} - \boldsymbol{x}_{cj} < 0 \end{cases} \qquad 1 \le j \le J. \tag{4}$$

When $\boldsymbol{x}_{ij} - \boldsymbol{x}_{cj} \ge 0$, the jth component $\boldsymbol{x}_{ij}$, of $\boldsymbol{x}_i$, is represented by binary one, namely $\boldsymbol{b}_{ij} = 1$. When $\boldsymbol{x}_{ij} - \boldsymbol{x}_{cj} < 0$, the jth component $\boldsymbol{x}_{ij}$, of $\boldsymbol{x}_i$, is represented by binary zero, namely $\boldsymbol{b}_{ij} = 0$. So the vector $\boldsymbol{x}_{ij}$ can be represented by binary vector $\{\boldsymbol{b}_{ij}\}$. $\{\boldsymbol{b}_{ij}\}$ is from $\{\underbrace{00\cdots0}\}_{\boldsymbol{J}}$ to $\{\underbrace{11\cdots1}\}_{\boldsymbol{J}}$.

The decimal value of binary vector $\{\boldsymbol{b}_{ij}\}$ is satisfied for the following relation.

$$0 \le \{b_{ij}\} \text{ value } \le 2^J - 1. \tag{5}$$

The binary vector $\{\boldsymbol{b}_{ij}\}$ is just corresponding to the index of hypercube $\boldsymbol{S}_m$ to which the vector, $\boldsymbol{x}_i$, belongs. Based on the discussion above, the binary vector $\{\boldsymbol{b}_{ij}\}$ is used as the output of binary quantizer. It is also clear that the decimal value of the binary vector, $\boldsymbol{b}_i$, is just equal to the index of the hypercube $\boldsymbol{S}_m$ in which the $\boldsymbol{x}_i$ resides.

Averaging the vectors which are from the training set and fall into hypercube m, the centroid vectors $\hat{\boldsymbol{x}}_m$, $m = 0, 1, \cdots, \boldsymbol{M}_{\mathrm{BQ}} - 1$, of the region are obtained.

$$\hat{\boldsymbol{x}}_{mj} = \frac{1}{\boldsymbol{N}_m} \sum_{i'=1}^{\boldsymbol{N}_m} \boldsymbol{x}_{i'j} \quad 1 \le j \le J \tag{6}$$

where $\boldsymbol{N}_m$ is the number of vectors, $\boldsymbol{x}_{i'}$, inside the hypercube m. Every vector inside $\boldsymbol{S}_m$ is reproduced by $\hat{\boldsymbol{x}}_m$. The set of $\hat{\boldsymbol{x}}_m$ will be used to investigate the average distortion caused by BQ in Section III.

Comparing BQ with VQ, we can see both quantizers are trained to find partitions of the feature vector space. The vector quantizer is trained to get the optimum set of codebook vectors $\hat{\boldsymbol{x}}_m$, $m = 1, 2, \cdots, \boldsymbol{M}_{\mathrm{VQ}}$. As the quantization level $\boldsymbol{M}_{\mathrm{VQ}}$ increases, the computation for training the optimal vector quantizer increases rapidly. Since the small decrease in average distortion from $\boldsymbol{M}_{\mathrm{VQ}} = 128$ to $\boldsymbol{M}_{\mathrm{VQ}} = 256$ did not justify the increased computation owing to the large codebook. Many VQ-based speech (including speaker) recognition systems decided to implement vector quantizer by using a $\boldsymbol{M}_{\mathrm{VQ}} \le 128$ [6], [9]. A binary quantizer is trained to find the origin of the new coordinate system. The computation of finding $\boldsymbol{x}_c$ in (3) only equals that of training one-level ($\boldsymbol{M}_{\mathrm{VQ}} = 1$) vector quantizer. The centroid $\hat{\boldsymbol{x}}_m$ of each partition $\boldsymbol{S}_m$ is obtained by averaging the vectors inside the region in terms of (6). Apparently, the calculation procedure is very simple. The splitting algorithm and optimizing procedures performed to produce an optimum higher level ($\boldsymbol{M}_{\mathrm{VQ}} > 1$) vector quantizer, which caused a large number of computations, are omitted in training a binary quantizer. In order to obtain good quantization performance, the vector space is divided into a large number of regions—much more than that of VQ, namely, $\boldsymbol{M}_{\mathrm{BQ}} \gg \boldsymbol{M}_{\mathrm{VQ}}$. If BQ is applied to a speech coding system, the weakness of BQ is that the number of centroid vectors (codewords) in BQ is much more than that in VQ. Fortunately, the centroid vectors are not necessary in BQHN-based speaker identification system. The centroid vectors are used as the reproduction vectors when we investigate the distortion caused by BQ because the performance criterion of vector quantization is the distortion (distance) measure [1]. In following section, this average distortion criterion is used to investigate the performance of BQ.

III. Investigation of the Average Distortion

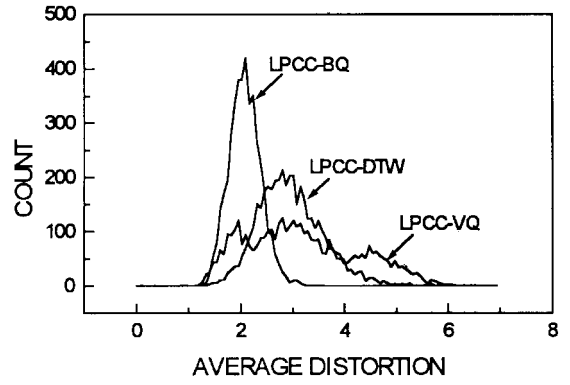
To have a better understanding of the average distortions caused by a binary quantizer, the average distortions caused by a vector quantizer and the intraspeaker average distances between two repetitions of a word voice with dynamic time warping (DTW) alignment will be investigated as well.

Before performing the distortion investigation, we introduce the speech data base and front-end analysis methods used in this paper. The data base consists of ten repetitions of each of ten Chinese digits ($0 \sim 9$) arranged in different sequences uttered by each of 42 (20 male, 22 female) talkers. So the total number of utterances is 4200, namely 420 repetitions for each digit, in the data base. The input speech signal, recorded with a universal microphone and cassette tape recorder in common room, is lowpass-filtered with 4 kHz cut-off frequency, and sampled at a 8 kHz sampling rate, and digitized with a 12-b A/D converter. The speech samples are preemphasized by a first-order filter with transfer function H(z) = $1 - 0.95\ z^{-1}$. The preemphasized speech data is blocked into analysis frames with a 30 ms (240 samples) Hamming window with each consecutive frame spaced 10 ms (80 samples) apart. Three kinds of feature parameters are derived from each analysis frame, respectively. A 12th-order LPC analysis is performed on each frame using the autocorrelation method, then i) 12th-order LSP frequencies and ii) 12th-order LPCC's are derived from the LPC analysis. The quefrency lifting [13] weights each individual LPCC components by its component order index because the higher-order coefficients contain more speaker individual characteristics [15] and cepstrum is a decaying sequence [16]. Each analysis frame is padded with 16 samples valued zero at the end of the frame; then iii) the 12th-order MFCC is derived from it.

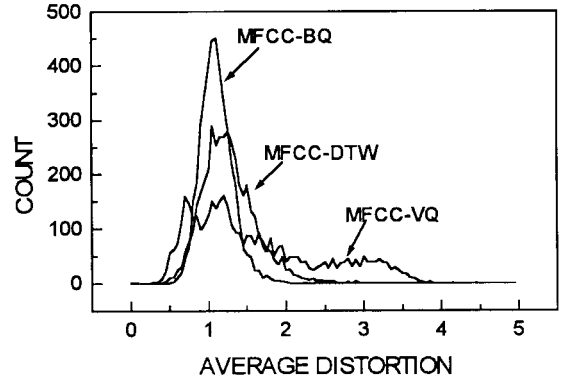
For the distortion investigation, the data base is divided into two sets: a training set and a testing set. The training set consists of 840 utterances (33 612 vectors), two repetitions of each digit uttered by each of 42 speakers. Three binary quantizers and three vector quantizers are trained with LSP, LPCC, and MFCC, respectively. The testing set consists of 3360 utterances, eight repetitions of each of ten digits uttered by each of 42 speakers. We assume the distortion caused by reproducing an input vector $\boldsymbol{x}_i$ by a reproduction vector $\hat{\boldsymbol{x}}_m$ is given by Euclidean distance measure defined in (1). For each utterance, the average distortion caused by the binary quantizer is defined as

$$\|D_{\mathrm{BQ}}\| = \frac{1}{I} \sum_{i=1}^{I} d(\hat{\boldsymbol{x}}_m, \boldsymbol{x}_i) \tag{7}$$

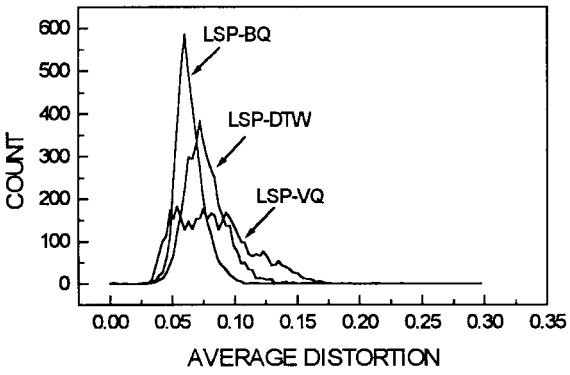
where I is the number of feature vectors in an utterance and m is the decimal value of binary vector $\boldsymbol{b}_m$ denoting the hypercube m, in which $\boldsymbol{x}_i$ resides. Since the number of components in a feature vector in an analysis frame is 12, namely $J = 12$ in (3)–(6), the binary value of vector $\boldsymbol{b}_m$ is from 0000 0000 0000 to 1111 1111 1111 corresponding to $2^{12} = 4096$ hypercubes. $\hat{\boldsymbol{x}}_m$ is the reproduction (centroid) vector of hypercube m, which is determined by (6). If hypercube m is blank, namely $\boldsymbol{N}_m = 0$ in (6), the $\hat{\boldsymbol{x}}_m$ is replaced by $\boldsymbol{x}_c$ derived from (3). d denotes the distortion when a feature vector $\boldsymbol{x}_i$ inside hypercube m is replaced by the reproduction vector $\hat{\boldsymbol{x}}_m$ of the hypercube.



(a)



(b)



(c)

Fig. 1. Histograms of the average distortions of VQ, BQ, and the intraspeaker repetition distances with DTW alignment in a set of digital utterances: (a) LPCC parameters, (b) MFCC parameters, and (c) LSP parameters.

For each test utterance, the average distortion caused by vector quantizer is defined as

$$||D_{\mathrm{VQ}}|| = \frac{1}{I}\sum_{i=1}^{I} \min_{1\le m\le \boldsymbol{M}_{\mathrm{VQ}}} [d(\hat{\boldsymbol{x}}_m, \boldsymbol{x}_i)] \tag{8}$$

where $\boldsymbol{M}_{\mathrm{VQ}}$ is quantization level of VQ. Considering the computation complexity and the small decrease in average distortion from $\boldsymbol{M}_{\mathrm{VQ}} = 128$ to $\boldsymbol{M}_{\mathrm{VQ}} = 256$, we let $\boldsymbol{M}_{\mathrm{VQ}} = 128$. In fact, $\boldsymbol{M}_{\mathrm{VQ}}$ is less than or equal to 128 in most VQ-based speech recognition systems.

TABLE I
LARGEST, SMALLEST, MEAN AND STANDARD DEVIATIONS OF THE DISTORTIONS (DISTANCES) OF VQ, BQ, AND DTW ALIGNMENT IN INVESTIGATING THE PERFORMANCE OF DIFFERENT QUANTIZERS WITH DIFFERENT FEATURE VECTORS

Parameter	Method	Largest	Smallest	Mean	Deviation
LPCC	VQ	6.570039	1.000000	3.201547	0.016534
	BQ	3.590703	0.965371	2.119808	0.004738
	DTW	6.459890	0.112628	2.993226	0.009955
MFCC	VQ	3.848174	0.380152	1.617967	0.012703
	BQ	2.282323	0.535725	1.130031	0.003154
	DTW	4.468733	0.043639	1.289826	0.005594
LSP	VQ	0.187533	0.033657	0.085217	0.000448
	BQ	0.117624	0.028729	0.065580	0.000167
	DTW	0.234642	0.001939	0.078030	0.000264

The intraspeaker average distance between two repetitions of a word voice is defined as

$$||D_{\mathrm{DTW}}|| = \frac{1}{I}\min_{w(i)}\sum_{i=1}^{I} d[\boldsymbol{x}_i, \boldsymbol{y}_{w(i)}] \tag{9}$$

where $\boldsymbol{x}_i, i = 1, 2, \cdots, I$, are the feature vectors which belong to a word voice. $\boldsymbol{y}_m$, $m = w(i)$ and $m = 1, 2, \cdots, M$, are the feature vectors which belong to a repetition utterance of the same word. Both the utterances are uttered by the same speaker. The alignment path is constrained by the conditions in (10) and (11).

$$w(1) = 1 \qquad w(I) = M \tag{10}$$

$$w(i+1) - w(i) = \begin{cases} 0, 1, 2 & w(i) \neq w(i-1) \\ 1, 2 & w(i) = w(i-1). \end{cases} \tag{11}$$

For each utterance in test set, a $||D_{\mathrm{BQ}}||$, a $||D_{\mathrm{VQ}}||$, and a $||D_{\mathrm{DTW}}||$ are calculated for each kind of parameter, respectively. For each kind of parameter, the numbers of $||D_{\mathrm{BQ}}||$, $||D_{\mathrm{VQ}}||$, and $||D_{\mathrm{DTW}}||$ are 3360, respectively. The histograms of $||D_{\mathrm{BQ}}||$, $||D_{\mathrm{VQ}}||$, and $||D_{\mathrm{DTW}}||$ are drawn in Fig. 1, respectively. Fig. 1(a) is for LPCC, Fig. 1(b) for MFCC, and Fig. 1(c) for LSP. The largest, the smallest, the mean, and the standard deviation of each kind of average distortions are listed in Table I. From Fig. 1 and Table I, we can see that the average distortion of BQ is lower than that of VQ. The average distortion of BQ even falls below the intraspeaker average distance between two repetitions of the same word with DTW alignment. This means the binary quantization precision is high enough and BQ can be put into practical use because the large variability exists in speech signals, even in the repetitions of the same word uttered by the same speaker. From the view of statistical pattern recognition, the average distortion smaller than the intraspeaker average repetition distance of the same word is not very important and meaningful.

IV. SPEAKER IDENTIFICATION SCHEME

A block diagram of the BQHN-based speaker recognition system is given in Fig. 2. A speaker's template is described with the hierarchical model shown in Fig. 3. This model is

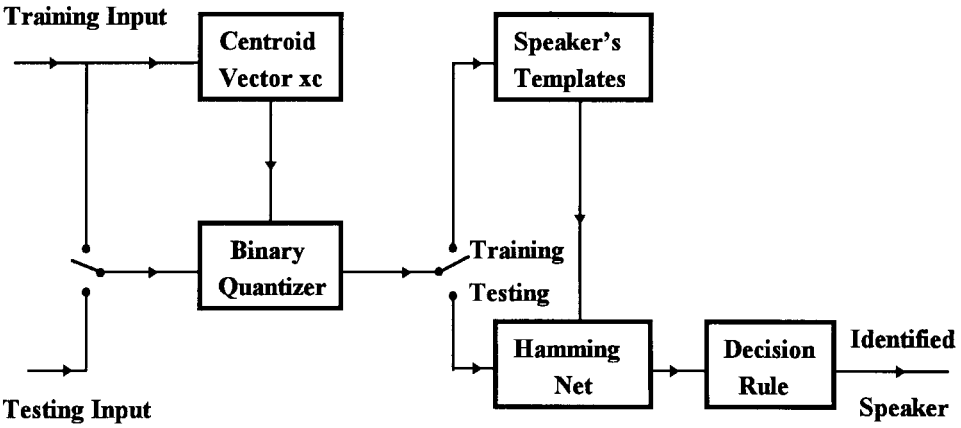


Fig. 2. Block diagram of BQHN-based speaker identifier.

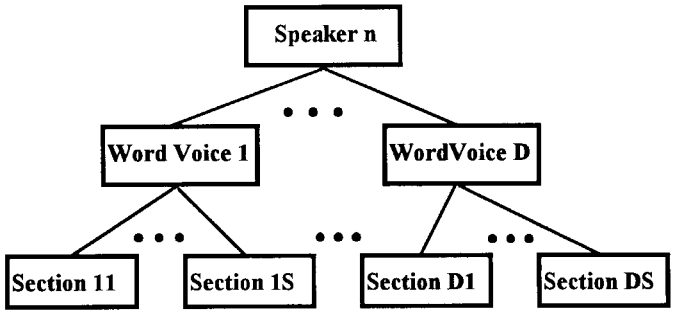


Fig. 3. Hierarchical model for describing the individual features of speaker n.

based on the fact that the speaker individual features reside in every word voice in the vocabulary. So, the recognizer operates as a text-independent (in digital vocabulary) speaker identifier. The identifier runs first in a training procedure to provide the centroid vector, $\boldsymbol{x}_c$, for binary quantizer and binary vectors to consist of each speaker's template. Five repetitions of each of ten digital voices of each of 42 speakers are used as training set for this purpose. The rest utterances are used as test data. Following are the steps for forming the binary reference patterns of each word for a speaker's template.

1) Each repetition of a word voice uttered by the speaker in the training set is divided into $\boldsymbol{S}$ sections in time sequence. If a training utterance has L frames, $(F_0, F_1, \cdots, F_L - 1)$, the number of frames in each section is

$$Ls = \text{INT}(L/S + 1). \tag{12}$$

The first frame of each section Fs

$$Fs = \begin{cases} \text{INT}((L^*s)/S) & s = 0, 1, \cdots, S-2 \\ F_{L-1} - Ls & s = S-1. \end{cases} \tag{13}$$

In adjacent sections, one or two frames overlapping may exist.

2) Merge the feature vectors from the corresponding sections of training utterances of the same word into a group, then, to average the vectors in this group, the centroid vector of the section, $x_s, l \le s \le S$, is obtained.

3) Each $\boldsymbol{x}_s$ is compared with $\boldsymbol{x}_c$ to get the binary representation, $\boldsymbol{b}_s$, $1 \le s \le S$, of the section s in terms of (4).

4) To arrange the S binary vector $\boldsymbol{b}_s$ in a binary sequence which is used as the reference $\boldsymbol{r}_d$ of the word d. Each component of $\boldsymbol{r}_d$ is $\boldsymbol{r}_{dk}$, $1 \le k \le K = S \times J$, $1 \le d \le D$, where S is the number of sections in a word voice, J is the number of components in a feature vector and D is the number of words in the vocabulary.

These steps are iterated for every word in the vocabulary for each speaker's template until the templates of 42 speakers (20 male and 22 female) are established.

In the identification phase, the feature vectors of the unknown speaker are first divided into S sections. After the vectors in each section are averaged, the centroid vector of each section is obtained. The centroid vectors are sent through the binary quantizer to get the binary representation of the input feature vectors i_k, $1 \le k \le K = S \times J$, as did in training phase. Then the distance between each word reference, $\boldsymbol{r}_{dk}$, from a speaker's templates and the input binary sequence, i_k, is measured to give a distance for each word reference by using the modified Hamming distance in the forward Hamming net

$$\|D_{\text{Hamming}}\|_d = \sum_{k=1}^{K} r_{dk} \oplus i_k \tag{14}$$

where $\oplus$ is the exclusive OR operator. From (14), we can see that the $\|D_{\text{Hamming}}\|_d$ equals the number of the different components between the word reference $\boldsymbol{r}_{dk}$ and the input binary pattern i_k. The smallest distance is chosen as the output of the speaker's template. The process and the forward Hamming net are illustrated in Fig. 4.

The decision rule chooses the speaker whose template gives the smallest distance.

$$\text{speaker}^* = \arg \min_{1 \le i \le N} \left\{ \min_{1 \le d \le D} \|D_{\text{Hamming}}\|_d^i \right\} \tag{15}$$

where D is the number of words in the vocabulary and N is the number of speakers registered in the system.

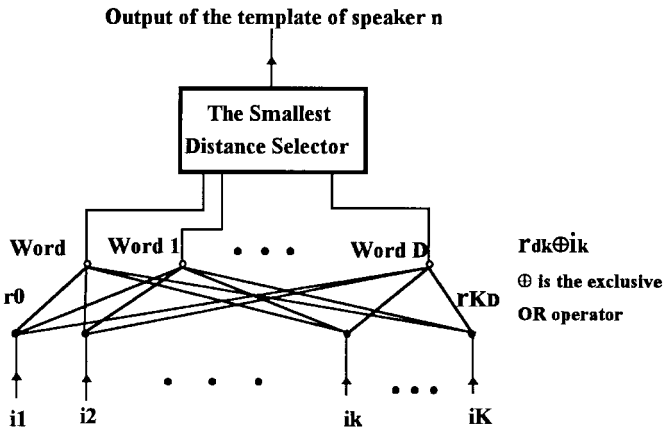


Fig. 4. Illustration of the forward Hamming net and distance measurement between word reference $\boldsymbol{r}_{dk}$ and input binary pattern i_k.

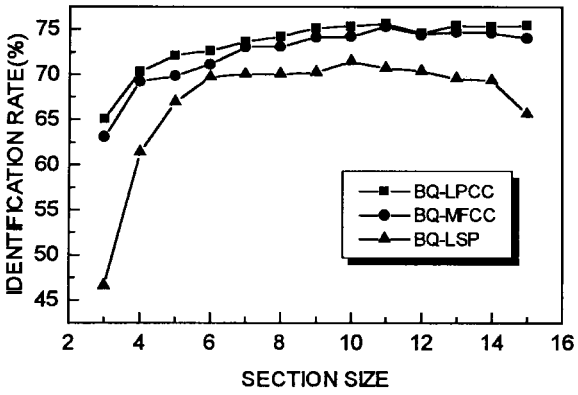
V. Experiments and Results

Since the speech data base was made by the authors (see Section III), a set of standard VQ-based speaker recognition experiments should be performed on this data base as "benchmarks." The VQ-based speaker recognition method chosen in this paper was proposed by Buck *et al.* [7]. They used multisection vector quantization (MSVQ) codebooks for text-dependent speaker recognition. We modify it in terms of the hierarchical model (see Fig. 3), so that the recognizer can operate as a text-independent speaker identifier. The section dividing method is the same one as discussed in Section IV. Multisection codebooks for each speaker are designed from five repetitions of each word and was represented by a rate-3 vector quantization codebook (quantization level $\boldsymbol{M}_{\mathrm{MSVQ}} = 2^3$). The codebook size is as large as that in paper [7]. The following experiments were performed to evaluate the effects of different system parameters on the recognition performance:

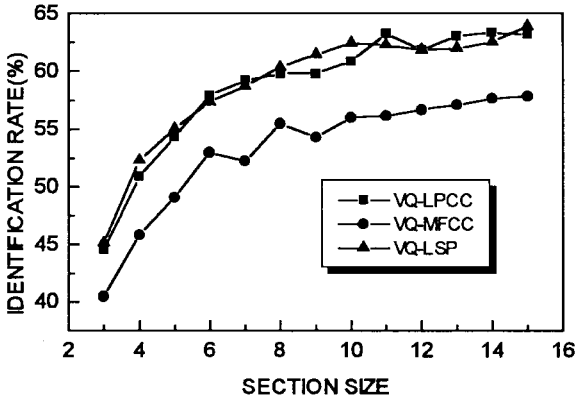
1) number of sections divided in a word voice (section size);
2) number of digits in test utterance;
3) Add Gaussian white noise to test utterance;
4) effect of different section size on robustness of the speaker identifiers.

A. Effects of Section Size

In training and testing phases, a word voice is divided into S sections in time sequence. The performance of BQHN-based identifier and MSVQ-based identifier is affected by the section size. This set of experiments seeks the effect of using different numbers of sections on the identification rates of the two identifiers. The experimental results are plotted in Fig. 5, where test utterance is a single digital voice, and section size varies from 3 to 15. Each point plotted in Fig. 5 represents 2100 speaker identifications—50 identifications per speaker for 42 speakers. Fig. 5(a) shows the performance of BQHN-based speaker identifier and Fig. 5(b) shows the performance of MSVQ-based identifier. From Fig. 5, the identification rates of the two identifiers increase rapidly when section size is less than ten. When section size is larger than ten, the identification rates of the two identifiers either increase very slow or decrease



(a)



(b)

Fig. 5. Effects of section size in a word voice on the identification performance. (a) Plots of the performance of BQHN-based speaker identifier versus section size. (b) Plots of the performance of MSVQ-based speaker identifier versus section size.

in some cases. Larger section size is of benefit to the MSBQ-based speaker identifier. Considering the computation and storage, all word voices used in both BQHN- and MSVQ-based speaker identifiers are divided into ten sections in Sections V-B and V-C.

B. Effects of Number of Digits in Test Utterance

The results in the first set of experiments were based on a single digital voice (for an average duration of 0.4 s) in the test utterance. By concatenating two or more, but different, digital voices randomly, different lengths of test utterances are obtained. The identification rate can be tested as a function of the duration of the test utterance. The results of this set of experiments performed on the BQHN-based identifier and MSVQ-based identifier are plotted in Fig. 6. For each point plotted in Fig. 6, 2100 identifications are performed to get the identification rate. Apparently, the identification rate increases as the duration of the test utterance increases.

The performance of BQHN-based identifier with LPCC as the front-end analysis is the best. For a test utterance of five digits, the identification rate is 99.24%. The performance of BQHN-based identifiers are better than that of MSVQ-based identifiers. The identification rates of BQHN-based identifiers are greater than 99% for test utterance concatenated five digits,

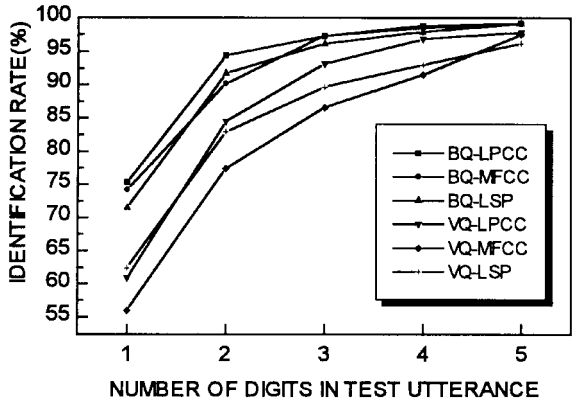


Fig. 6. Plots of speaker identification rate as a function of the number of digits in the test utterance.

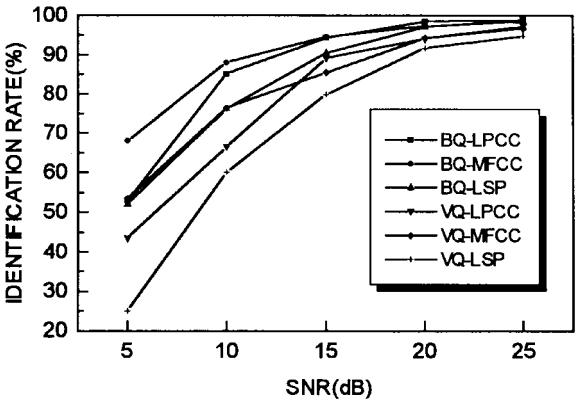


Fig. 7. Plots of speaker identification rate as a function of SNR.

while that of MSVQ-based identifiers are less than 98%, but greater than 97%, for the same duration of the test utterance.

C. Effects of Noise

The third set of experiments was concerned with the robust performance of the BQHN-based and MSVQ-based identifiers in noisy conditions. The identifiers were trained with clean speech and tested with additive Gaussian white noise-corrupted speech. The test speech and the noise are additive in the time domain. The test results are plotted in Fig. 7. The number of digits in the test utterance is five, and each point plotted in this figure represents 2100 identifications. Apparently, from Fig. 7, the performance of BQHN-based identifiers is better than that of MSVQ-based identifiers in a wide range of SNR. MFCC is more robust at low SNR.

To further illustrate the robustness of BQHN-based identifier in additive Gaussian noise, Fig. 8 shows the histograms of 2100 intraspeaker and 2100 interspeaker distortions of the 42 speaker identifications at three different SNR. The distortions given by BQHN-based identifier is uniformed by the distortions given by MSVQ-based identifier for comparison. In Fig. 8(a), the distortions were tested on clean speech. In Fig. 8(b), the distortions were tested at 15 dB SNR. In Fig. 8(c), the distortions were tested at 5 dB SNR. The feature parameters used for Fig. 8 are MFCC. The histograms plotted in the left column of Fig. 8 are from BQHN-based identifier and those in the right column are from MSVQ-based identifier.

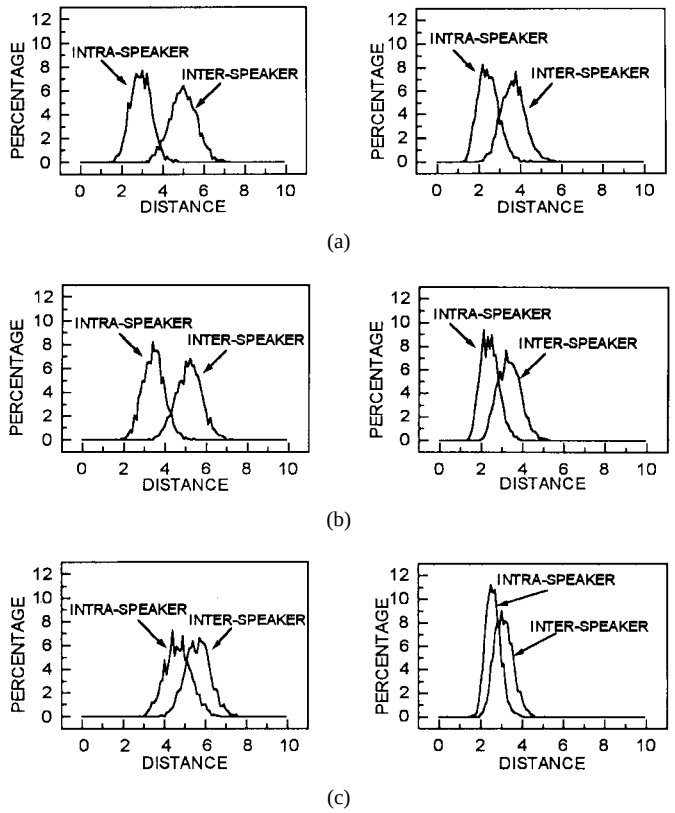


Fig. 8. Effects of noisy speech on the performances of BQHN-based speaker identifier (left column) and MSVQ-based speaker identifier (right column). (a) Clean speech. (b) SNR = 15 dB. (c) SNR = 5 dB.

The overlapping region between the histogram of intraspeaker and that of interspeaker increases as the SNR decreases. As the result of this increase of overlapping area, the recognition rate is decreased. The overlapping area derived from BQHN-based identifier is much lower than that derived from MSVQ-based identifier at corresponding SNR. These experiment results indicate why the performance of BQHN-based identifier is better than that of MSVQ-based identifier in noisy speech.

In addition to the above experiments in which the speaker identifiers were trained with clean speech, the identification rates were also tested with the same level of additive Gaussian white noise added to testing set as that added to training set, namely $\mathrm{SNR}_{test} = \mathrm{SNR}_{train}$. The percent identification accuracy of BQHN-based speaker identifier are listed in Table II. The results have indicated that the BQHN-based speaker identifier can achieve good identification performance if the conditions during training phase are similar to those during testing phase.

D. The Effects of Section Size on Robustness of the Speaker Identifiers

Results in Section V-C indicate that BQHN-based speaker identifier is more robust than MSVQ-based speaker identifier under additive Gaussian white noise. This experiment is designed to evaluate the effects of section size on robustness of the two speaker identifiers. The section size varies from 6 to 14, and the feature parameters used in this experiment are MFCC. 2100 speaker identifications are performed on a single digital voice in test utterances for each plotting point.

TABLE II
PERFORMANCE OF BQHN-BASED SPEAKER IDENTIFIER WHEN IT WAS TRAINED AND TESTED AT THE SAME LEVELS OF SNR

Number of Digits in Test Utterance	Parameter	Identification Rate(%) in different noise levels				
		5dB	10dB	15dB	20dB	25dB
1	LPCC	63.00	68.57	71.86	72.81	74.81
	MFCC	63.35	68.67	71.47	72.81	74.24
	LSP	60.80	65.33	66.38	69.09	70.85
2	LPCC	87.52	92.00	93.24	93.1	94.28
	MFCC	88.95	91.10	89.95	89.86	90.28
	LSP	84.47	87.23	89.04	90.19	91.00
3	LPCC	93.81	94.90	96.62	97.24	97.24
	MFCC	95.71	96.38	96.05	96.38	96.71
	LSP	92.76	94.09	96.09	96.04	96.52
4	LPCC	97.14	98.24	98.85	98.95	99.10
	MFCC	97.31	97.95	98.47	98.43	98.43
	LSP	96.04	97.61	98.19	98.30	98.42
5	LPCC	98.43	99.33	99.48	99.52	99.62
	MFCC	98.52	99.52	99.33	99.14	99.10
	LSP	97.95	99.00	99.09	99.33	99.19

The results are plotted in Fig. 9. Fig. 9(a) is for BQHN-based speaker identifier and Fig. 9(b) is for MSVQ-based speaker identifier. We can see clearly that the curves at different SNR plotted in Fig. 9 are almost parallel, which means the section size has no effect on the robustness of the two speaker identifiers. Comparing Fig. 9(a) with Fig. 9(b), it is demonstrated again that BQHN-based speaker identifier is more robust than MSVQ-based speaker identifier when test speech is corrupted by additive Gaussian white noise.

VI. STORAGE AND COMPUTATION ESTIMATION

Besides the robust performance of the BQHN-based identifier in additive Gaussian white noise, it is worthwhile roughly estimating the storage and the computation needed in designing the binary quantizer, in quantization phase, and in identification phase. In designing the binary quantizer, what needs to be done is to evaluate the centroid vector, $\boldsymbol{x}_c$, in (3), $J \cdot (I-1)$ additions and J divisions are required. Only $4 \cdot J = 4 \cdot 12 = 48$, $J = 12$ and 4 bytes are needed for storing a floating point datum, bytes storage are required for saving J floating point components of the vector, where J is the number of components in a vector, and I is the number of vectors in training set.

To get the binary representation of a feature vector, only J comparisons between the input vector and the centroid vector, $\boldsymbol{x}_c$, are required [see (4)]. If S is the number of sections divided in a word voice, D is the number of words in the vocabulary and N is the number of speakers registered in the identifier, $S \cdot D \cdot J \cdot N$ bits are required for saving the templates. For example, $S = 10$, $D = 10$, $J = 12$, $N = 42$,

$$S \cdot D \cdot J \cdot N \cdot = 10 \cdot 10 \cdot 12 \cdot 42 = 50\,400 \text{ bits} = 6300 \text{ bytes}.$$

For evaluating of a speaker's distance in forward Hamming net in identification phase,

$$S \cdot J \cdot D = 1200 \text{ exclusive OR operations}$$

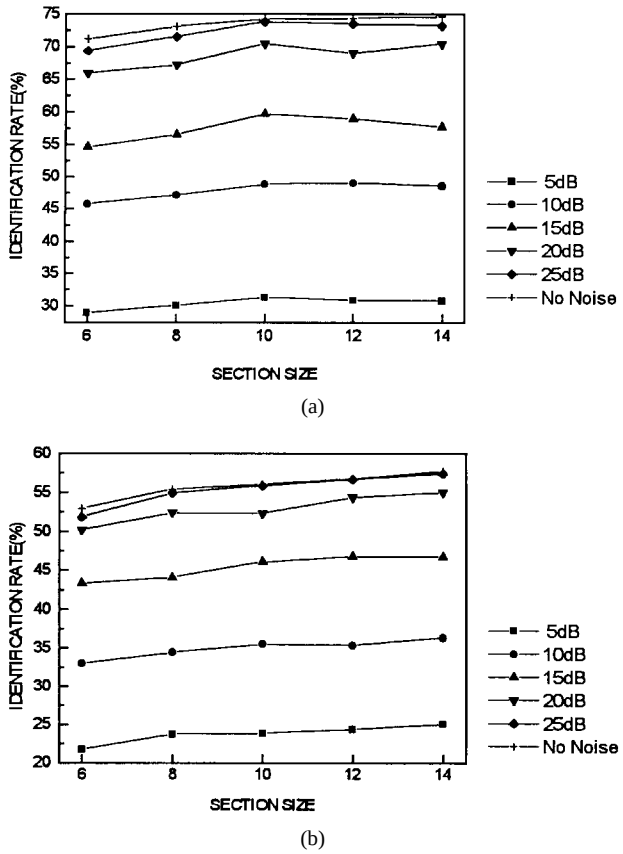


Fig. 9. Effects of section size in a word voice on the identification rate at different SNR. (a) Plots of robust performance of BQHN-based speaker identifier versus section size. (b) Plots of robust performance of MSVQ-based speaker identifier versus section size.

and $D = 10$ comparisons are needed. The decision rule needs $N = 42$ comparisons.

It should be noted that in our rough evaluation of computation, we have not included the computation of front-end analysis. This computation is necessary in any other identifiers.

If we compare the computation and the storage required in BQHN-based identifier with that required in any other VQ-based identifiers, we can see the computation and storage needed in BQHN-based identifier are much smaller.

VII. Summary

We have proposed an approach to quantize feature vectors: BQ. In order to demonstrate the effectiveness of BQ, the performance criterion of vector quantization, distortion (distance) measure, was employed. Results indicate that the average distortion caused by BQ is lower than that caused by VQ when $M_{\mathrm{VQ}} = 128$ and even lower than the intraspeaker average distance between two repetitions of the same word.

Since the output of BQ is a binary sequence, it is possible for us to combine it with forward Hamming net classifier. In terms of the idea of hierarchical model for describing individual features of a speaker, we have designed a new speaker identification system that operates as a text-independent speaker identifier in digital vocabulary. Experimental results show that the performance of the identifier is very good. In addition to the robust performance of the BQHN-based identifier, the little computation and the small storage required in the identifier are the other two advantages.

Acknowledgment

The authors would like to thank Prof. D. Gonghuan for his help and advice throughout this work, and the reviewers, whose excellent comments and suggestions helped improve the quality of the paper.

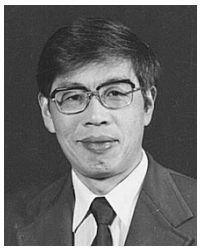
References

[1] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. 28, pp. 84–95, Jan. 1980.
[2] D. Y. Wong, B. H. Juang, and A. H. Gray, Jr., "An 800 bps LPC vector quantization vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 770–780, 1982.
[3] A. V. McCree and T. P. Barnwell, III, "Implementation and evaluation of a 2400 bps mixed excitation LPC vocoder," in *Proc. IEEE ICASSP'93*, pp. II159–II162.
[4] D. K. Burton and J. E. Shore, "Speaker-dependent isolated word recognition using speaker-independent vector quantization codebooks augmented with speaker specific data," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 440–443, 1985.
[5] D. K. Burton, J. E. Shore, and J. T. Buck, "Isolated-word speech recognition using multisection vector quantization codebooks," *IEEE Trans. Acoust., Speech, Signal Processing* vol. ASSP-33, pp. 837–849, 1985.
[6] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B. H. Juang, "A vector quantization approach to speaker recognition," in *Proc. IEEE ICASSP'85*, pp. 387–390.
[7] J. T. Buck, D. K. Burton, and J. E. Shore, "Text-dependent speaker recognition using vector quantization," in *Proc. IEEE ICASSP-85*, pp. 391–294.
[8] J. E. Shore and D. K. Burton, "Discrete utterance speech recognition without time alignment," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 473–491, 1983.
[9] L. R. Rabiner, S. E. 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, pp. 1075–1105, 1983.
[10] B. S. Atal, "Effectiveness of linear prediction characteristics of speech wave for automatic speaker identification and verification," *J.A.S.A.*, vol. 55, pp. 1304–1312, 1974.
[11] F. K. Soong and B.-H. Juang, "Optimal quantization of LSP parameters," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 15–24, 1993.
[12] Y. Bistritz and S. Peller, "Immittance spectral pairs (ISP) for speech encoding," in *Proc. IEEE ICASSP'93*, pp. II9-II12.
[13] R. J. Mannone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition: A feature-based approach," *IEEE Signal Processing Mag.*, vol. 13, pp. 58–71, 1996.
[14] F. Soong and B.-H. Juang, "Line spectrum pair and speech data compression," in *Proc. IEEE ICASSP'84*, pp. 1.10.1–1.10.4.
[15] Y. Zhong-Xuan, Y. Chong-Zhi, and F. Yuan, "Text-independent speaker identification using fuzzy mathematical algorithm," in *Proc. IEEE ICASSP'93*, pp. II403–II406.
[16] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals.* Englewood Cliffs, NJ: Prentice-Hall, 1978.
[17] Y. Zhong-Xuan, X. Bo-Ling, and Y. Chong-Zhi, "A kind of fuzzy-neural networks for text-independent speaker identification," in *Proc. IEEE ICASSP'96*, pp. 657–660.
[18] B.-H. Juang, D. Y. Wong, and A. H. Gray, Jr., "Distortion performance of vector quantization for LPC voice coding," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 294–304, 1982.
[19] T. Matsui and S. Furui, "A text-independent speaker recognition method robust against utterance variations," in *Proc. IEEE ICASSP'91*, pp. 377–380.
[20] L. Xu, J. Oglesby, and J. S. Mason, "The optimization of perceptually-based features for speaker identification," in *Proc. IEEE ICASSP'89*, pp. 520–523.
[21] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," *J.A.S.A.*, pt. 2, vol. 51, pp. 2044–2056, 1972.

Zhong-Xuan Yuan (S'95–A'98) received the M.S. and Ph.D. degrees in acoustics from Nanjing University, Nanjing, China, in 1993 and 1998, respectively.

Currently, he is an Assistant Professor with the Department of Electronic Science and Engineering, Nanjing University. His research interests include speech signal processing, pattern recognition, fuzzy logic, and the use of artificial neural networks for acoustical signal processing.

Dr. Yuan is a Committee Member of the Speech, Auditory, and Music Branch of the Acoustical Society of China.



Bo-Ling Xu received the M.Sc. degree from Nanjing University, Nanjing, China, in 1981.

Since then, he has been with the Institute of Acoustics, Nanjing University, where he is currently a Professor. From June 1985 to December 1986, he conducted research on digital signal processing at the Department of Electrical Engineering, University of South Florida, Tampa, as a Visiting Scholar. His research interests include signal processing, speech signal processing, and electroacoustics.

Mr. Xu is the Deputy Chairman of the Audio Engineering Society of China. He is also a member of the council of the Acoustical Society of China.

Chong-Zhi Yu graduated in physics from Amoy University, Amoy, China, in 1953. From 1955 to 1956, he took graduate training in acoustics at the Physics Department, Nanjing University, Nanjing, China.

From 1983 to 1984, he was a Visiting Scholar at the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA. Currently, he is a Professor at the Institute of Acoustics, Nanjing University. His research interests include speech processing, auditory modeling, and neural networks.