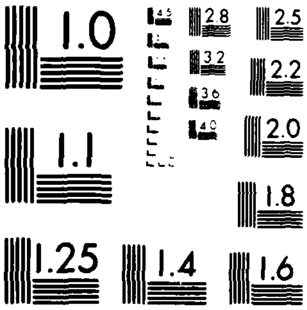

MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS-1963-A

2

NRL Memorandum Report 5367

Isolated-Word Speech Recognition Using Multi-Section Vector Quantization Code Books

D. K. BURTON, J. E. SHORE, AND J. T. BUCK

Computer Science and Systems Branch
Information Technology Division

July 13, 1984

NAVAL RESEARCH LABORATORY
Washington, D.C.

Approved for public release; distribution unlimited

REPORT DOCUMENTATION PAGE

1a REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

1b RESTRICTIVE MARKINGS

2a SECURITY CLASSIFICATION AUTHORITY

2b DECLASSIFICATION/DOWNGRADING SCHEDULE

3 DISTRIBUTION/AVAILABILITY OF REPORT
Approved for public release; distribution unlimited.

4 PERFORMING ORGANIZATION REPORT NUMBER(S)
NRL Memorandum Report 5367

5 MONITORING ORGANIZATION REPORT NUMBER(S)

6a NAME OF PERFORMING ORGANIZATION
Naval Research Laboratory

6b OFFICE SYMBOL (If applicable)
Code 7590

7a NAME OF MONITORING ORGANIZATION

6c ADDRESS (City, State and ZIP Code)
Washington, DC 20375

7b ADDRESS (City, State and ZIP Code)

8a NAME OF FUNDING/SPONSORING ORGANIZATION
Office of Naval Research

8b OFFICE SYMBOL (If applicable)

9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER

8c ADDRESS (City, State and ZIP Code)
Arlington, VA 22217

10 SOURCE OF FUNDING NOS

PROGRAM ELEMENT NO	PROJECT NO	TASK NO	WORK UNIT NO
61153N		RR014-09-41	75-0107-04

11 TITLE (Include Security Classification)
(See page ii)

12 PERSONAL AUTHOR(S)
Burton, D.K., Shore, J.E., and Buck, J.T.

13a TYPE OF REPORT
Final

13b TIME COVERED
FROM TO

14 DATE OF REPORT (Yr., Mo., Day)
July 13, 1984

15 PAGE COUNT
49

16 SUPPLEMENTARY NOTATION

17 COSATI CODES

FIELD	GROUP	SUB GR

18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Speech recognition
Isolated word recognition
Vector quantization
Multi-section code books
(Continues)

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

A new approach to isolated-word speech recognition using vector quantization (VQ) is examined. In this approach, words are recognized by means of sequences of VQ code books called multi-section code books. A separate multi-section code book is designed for each word in the recognition vocabulary by dividing the word into equal-length sections and designing a standard VQ code book for each section. Unknown words are classified by dividing them into corresponding sections, encoding them with the multi-section code books, and finding the multi-section code book that yields the smallest average distortion. For speaker-independent recognition of a 20-word vocabulary containing the digits, this approach achieves 95% recognition accuracy for the full vocabulary and 99% for the digits, in both causes with approximately 90% fewer distortion computations than typical dynamic-time-warping approaches. In addition, the approach achieves greater than 99% accuracy for speaker-dependent recognition of the digits with only 1 distortion computation per input frame per vocabulary word. The approach is described, detailed experimental results are presented and discussed, and computational requirements are analyzed.

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT
UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT □ DTIC USERS □

21 ABSTRACT SECURITY CLASSIFICATION
UNCLASSIFIED

22a NAME OF RESPONSIBLE INDIVIDUAL
D.K. Burton

22b TELEPHONE NUMBER (Include Area Code)
(202) 767-3490

22c OFFICE SYMBOL
Code 7590

11. TITLE (Include Security Classification)

Isolated-Word Speech Recognition Using Multi-Section Vector Quantization Code Books

18. SUBJECT TERMS (Continued)

Information theory
Computationally efficient
Nonlinear time alignment

CONTENTS

ISOLATED-WORD SPEECH RECOGNITION USING MULTI-SECTION VECTOR QUANTIZATION CODE BOOKS

I. INTRODUCTION

Vector Quantization (VQ) is a data compression principle [1] with several successful applications, including speech coding, [2, 3, 4] image coding [5, 6], and speech recognition [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17]. In previous work on speech recognition [8, 9, 16], we developed a method in which isolated words are classified by means of the average distortion that results from encoding them with VQ code books. In this paper, we present a generalization of that method. The generalization, which improves recognition performance and reduces computational requirements, was motivated by work of Martinez, Riviera, and Buzo [10].

In our previous approach [16], a VQ code book is generated for each word in the recognition vocabulary by applying an information-theoretic, iterative clustering technique [18] to a training sequence containing several repetitions of the vocabulary word. This clustering process removes all time-sequence information from the training sequence and represents each vocabulary word as a set of independent spectra. An input utterance is classified by encoding it with every code book and finding the code book that yields the smallest average distortion. Because the average distortion does not depend on the sequence of input speech frames, this approach performs isolated-word recognition entirely without time-alignment.

With just four spectra in each code book, our previous approach achieved 97.7% accuracy for speaker-dependent recognition of a twenty-word vocabulary [16]. With eight spectra in each code book, the accuracy increased to 98.8% [16]. These results showed that much more can be done without time-sequence information than is commonly assumed. For suitably chosen vocabularies,

Manuscript approved April 20, 1984.

characteristic spectra contain enough information for recognition, and information-theoretic clustering does a good job of extracting that information from training data.

To improve recognition performance and to decrease computational complexity, we have been investigating ways of incorporating time-sequence information into the recognition procedure. Here, we present results for a new method that incorporates time-sequence information by means of sequences of VQ code books that we refer to collectively as *multi-section* code books. A separate multi-section code book is designed for each word in the recognition vocabulary by dividing the words in the code book's training sequence into equal-length sections and designing a standard VQ code book for each section. Unknown words are classified by dividing them into appropriate sections, performing VQ on a section by section basis, and finding the multi-section code book that yields the smallest average distortion. The new approach reduces to our previous approach when the number of sections is reduced to one. Henceforth, we refer to our previous approach as the *single-section* case. Preliminary results for the multi-section approach were reported in [12, 17].

VQ has also been used by others to reduce the computational and memory requirements of existing isolated-word recognition approaches [7, 11, 13, 14, 15]. In these approaches, spectra from a single, large VQ code book are used to replace the spectra of both input speech frames and stored reference data. Our approach is quite different, both because we design separate code books for each word in the recognition vocabulary, and because we avoid standard methods of time alignment.

After explaining our speech recognition approach in Section II, we describe the data base and experiments in Section III. Section IV contains the results for speaker-independent recognition, and Section V contains results for speaker-dependent recognition. We discuss computational considerations in Section VI, and we present some general conclusions in Section VII.

II. APPROACH

In this section, we give background information and describe the multisection approach. We begin by describing VQ and explaining its role in our isolated-word recognition approach. We then discuss distortion measures, linear prediction parameters, and figures of merit.

A. Vector Quantization

VQ is an information-theoretic data compression principle introduced by Shannon in the late 1950's [19]. For a specified transmission rate, VQ's objective is to find the set of reproduction vectors, or code book, that represents an information source with minimum expected "distortion". The data compression is achieved by transmitting a reproduction vector index rather than the original source vector. In general, the selection of a perceptually meaningful distortion measure and the construction of an optimal code book are difficult problems. For speech, however, good choices exist [2, 3].

Speech coding by VQ is a narrow-bandwidth speech coding technique based on linear pı edictive coding (LPC) [2, 3]. Using estimates of the sample autocorrelation function that are measured in each frame, the shape of the speech spectrum in each frame is encoded as the index of a prestored set of LPC parameters that define an autoregressive model and is called a *codeword*. The LPC parameters used are the inverse filter gain squared σ^2 and the linear predictive coefficients a_i, $i=1, \cdots, M$, with $a_0=1$. The collection of possible codewords is called a *code book*. Let $\mathbf{C}=\{C_1, C_2, \cdots, C_N\}$ be a code book of N codewords C_i, each defining an autoregressive model and comprising a set of LPC parameters. Let S_j be the autocorrelation estimates from the jth frame of the speech to be coded. Then the spectrum shape of the jth frame is coded by identifying the codeword C_b that "best represents" S_j according to the "nearest-neighbor rule"

$$d(S_j, C_b) = \min_i d(S_j, C_i), \tag{1}$$

for some distortion measure d.

Vector quantization code books are designed to minimize the average distortion that results from encoding a long training sequence of speech frames. In particular, if T_j, $j=1, \cdots, L$ is such a training sequence, the code book **C** is designed so that

$$\frac{1}{L}\sum_{j=1}^{L} \min_i d(T_j, C_i) \tag{2}$$

achieves at least a local minimum. If the training sequence consists of typical speech and it is represented with a small average distortion by the code book, then **C** should encode new speech with a similarly small distortion. In practice, code books are designed by an iterative, clustering technique. The algorithm used here is based on the work in [18, 2]. Put simply, the L frames of the training sequence are divided into N clusters such that all frames in the same cluster have similar spectrum shapes. The N codewords are the centroids of these clusters.

B. VQ Word Recognition

In speech coding by VQ, a single code book is designed from a long training sequence that is representative of all speech to be encoded by the system. In the single-section approach to isolated word recognition [8, 9, 16], we used a separate code book for each word in the recognition vocabulary. We designed each code book from a training sequence containing repetitions of one vocabulary word. For example, a code book for the word "seven" would be designed by running the vector quantizer design algorithm on a training sequence of several repetitions of the word "seven". To classify an unknown word, it is first encoded using each of the code books and the average distortion for each code book is recorded. The unknown word is then classified according to the code book yielding the lowest average distortion.

Our new method, based on [10], represents each vocabulary word as a time-dependent sequence of section code books, which we call a multi-section code book. New words are classified by performing VQ and finding the multi-section code book that achieves the smallest average distortion.

To be more precise, let V be the number of words in the recognition vocabulary, and let T_k be the number of utterances in the training sequence used to design code book $\mathbf{C}_k$ for the k^{th} vocabulary word, where $k=1, \cdots, V$. Also, let F_{qk} be the number of frames in the q^{th} utterance in the training sequence for $\mathbf{C}_k$ where $q=1, \cdots, T_k$, and finally, let U_{mqk} be the m^{th} frame in the q^{th} training utterance for $\mathbf{C}_k$ where $m=1, \cdots F_{qk}$. Then there are V multi-section code books $\mathbf{C}_k$, each comprising a sequence of VQ *section code books* $\mathbf{C}_{kj}$. The section code book $\mathbf{C}_{kj}$ is designed using n frames from each training utterance for the k^{th} vocabulary word. That is, $\mathbf{C}_{kj}$ is designed from the frames U_{mqk}, where $m=(j-1)n+1, \cdots, jn$, and $q=1, \cdots T_k$. In particular, $\mathbf{C}_{k1}$ is designed from the first n frames of each training utterance for the k^{th} word in the recognition vocabulary, $\mathbf{C}_{k2}$ from the second n frames, etc. We call n the *compression factor* – it is the number of frames that are spanned per section. If, for a particular training utterance q, m is greater than F_{qk}, the corresponding frames U_{mqk} lie beyond the end of the word and are not included in the training sequence for $\mathbf{C}_{kj}$. Finally, let C_{kji}, $i=1,...,N_{kj}$ be codewords in section code book $\mathbf{C}_{kj}$. We call the V multi-section code books $\{\mathbf{C}_k\ ;\ k=1,\cdots,V\}$ a *code book set*.

Suppose a new utterance to be classified contains L frames, and P_l is the set of autocorrelation estimates from the lth frame $(l=1,\ldots,L)$. Now let D_k be the ***average distortion*** resulting from coding the unknown utterance with the code book $\mathbf{C}_k$,

$$D_k = \frac{1}{L}\sum_{j=1}^{S_k} d_{kj}, \tag{3}$$

where S_k is the number of section code books in $\mathbf{C}_k$, and

$$d_{kj} = \sum_{l=(j-1)n+1}^{\min[jn,\,L]} \min_i d(P_l, C_{kji}), \tag{4}$$

is the total distortion from coding the j^{th} section of the input with the j^{th} section code book $\mathbf{C}_{kj}$ of $\mathbf{C}_k$, and where n is the compression factor. Then the utterance is classified as the r^{th} word in the recognition vocabulary, where

$$D_r = \min_k D_k. \tag{5}$$

If desired, one can select a set of threshold values $D_{\min}$ and require $D_r < D_{\min}$ in (3) for a valid classification. This can improve classification reliability.

If, in the above description, all words are aligned at their beginnings, we call the approach ***left-aligned***. In the left-aligned case, variations in speaking rates often result in several sounds being included in the training sequences for individual section code books. To reduce this effect, we also tried linearly normalizing all training sequence and classification utterances to the same length. We call this approach ***length-normalized***.

In the length-normalized approach, the number of sections in the input word is always equal to the number of section code books. In the left-aligned approach, however, the input word can have more or less sections than the code books; we stop encoding a word in a code book when we run out of either input word frames or code book sections.

In the foregoing terms, the approach in [10] corresponds to left-alignment with $n=1$. For left-alignment with n greater than or equal to the maximum number of frames in all the training utterances, the multi-section approach reduces to our previous single-section approach. [8, 9, 16]

C. Multi-Section Code Books

Each classification code book $\mathbf{C}_k$ is designed from a separate training sequence containing repetitions of the kth word in the recognition vocabulary. A speaker-dependent code book is made from a training sequence spoken by one person. The resulting code books are then used to classify additional utterances from that speaker. For speaker-independent code books, the training sequence for each code book is spoken by several people and the code books are used to classify additional utterances from different people.

We used three types of multi-section code books:

(a) fixed-size code books;

(b) fixed-distortion code books;

(c) unclustered code books.

The three code book types are further discussed below.

As the name implies, in a ***fixed-size*** code book the section code book size N_{kj} is specified ahead of time and the design algorithm chooses N_{kj} codewords that minimize the average distortion resulting from encoding the training sequence for a particular section code book. Section code book sizes are limited for convenience to powers of 2, i.e., $N_{kj} = 2^{r_k}$, where r_k is called the *rate* of $\mathbf{C}_{kj}$. All section code books (and thus multi-section code books) in a fixed-size code book set have the number of code words.

For a *fixed-distortion* code book, the design algorithm increases the section code book size until it can design a section code book that encodes the training sequence with an average distortion that is less than or equal to a pre-specified value T. All section code books in a fixed-distortion code book set are generated with the same average distortion threshold and can therefore have different sizes. Like fixed-size section code books, the size of fixed-distortion section code book are limited to powers of 2.

The third type of code book is the *unclustered code book*. These are generated without the clustering algorithm, simply by making a codeword out of each frame in the training sequence. Our motivation for considering unclustered code books was twofold. The first was computational efficiency and convenience – generating them is much easier than generating clustered code books. The second was as a measure of performance. Since the clustering procedure attempts to find spectrum shapes that are representative of the training sequence, the effectiveness of clustering can be evaluated by comparing the performance of clustered and unclustered code books designed from the same training sequence.

D. Distortion Measures

In generating code books for voice coding, two distortion measures are effective [2, 20]. They are the *Itakura-Saito* (d_{IS}) and *gain normalized Itakura-Saito* (d_{GN}) distortion measures. For two power spectra $f(\vartheta)$ and $\hat{f}(\vartheta)$, the d_{IS} distortion between them is

$$d_{IS}(f,\hat{f}) = \int_{-\pi}^{\pi} \frac{d\vartheta}{2\pi}\left[\frac{f}{\hat{f}} - \ln\frac{f}{\hat{f}} - 1\right]. \qquad (6)$$

For power spectrum estimates f and $\hat{f}$ that have the autoregressive (LPC) form

$$f(\vartheta) = \frac{\sigma^2}{|A(z)|^2}, \qquad (7)$$

where

$$A(z) = \sum_{k=0}^{M} a_k z^{-k}$$

and $z = \exp(i\vartheta)$, the d_{GN} distortion is given by

$$d_{GN}(f,\hat{f}) \equiv d_{IS}\left(\frac{f}{\sigma^2},\frac{\hat{f}}{\hat{\sigma}^2}\right) = \frac{\alpha}{\sigma^2} - 1, \qquad (8)$$

where

$$\alpha = r(0)\hat{r}_a(0) + 2\sum_{n=1}^{M} r(n)\hat{r}_a(n),$$

$$\hat{r}_a(n) = \sum_{i=0}^{M-n} \hat{a}_i \hat{a}_{i+n},$$

and where $r(n)$ are the time-domain autocorrelations of $f(\vartheta)$.

Equations (6) and (7) show that d_{IS} depends on both the spectrum shape and the gain (σ^2). Thus, using it in (2) to design code books results in clusters that are sensitive both to spectrum shape and gain. Using d_{GN}, however, leads to clusters that depend only on spectrum shape. After extensive speech recognition experiments comparing the performance of these two distortion measures using single-section code books [16], we concluded that d_{GN} code books are better for speech recognition than d_{IS} code books, particularly when using small code books built from short training sequences. Thus, we used d_{GN} code books in the work reported herein.

For the classification distortion measure in (4), we considered three choices: d_{IS}, d_{GV}, and the *gain optimized Itakura-Saito* (d_{GO}) distortion measure.

$$d_{GO}(f,\hat{f}) \equiv \min_{\lambda>0} d_{IS}(f,\lambda\hat{f})$$

$$= \ln \int_{-\pi}^{\pi} \frac{d\vartheta}{2\pi}\left[\frac{f}{\hat{f}}\right] - \int_{-\pi}^{\pi} \frac{d\vartheta}{2\pi} \ln\left[\frac{f}{\hat{f}}\right] \qquad (9)$$

Like d_{GV}, d_{GO} is sensitive to spectral shape only. Properties of all three distortion measures are discussed in [21]. In our work with single-section code books [16], we found d_{GO} to be the best choice, and we used that same choice in the work reported herein. For LPC spectra of the form (7), d_{GO} can be expressed as

$$d_{GO}(f,\hat{f}) = \ln(\alpha) - \ln(\sigma^2). \qquad (10)$$

E. LPC Parameters

LPC parameters for both code book generation and utterance classification were generated using the autocorrelation method with Hamming windowing. Except for N, the number of points to shift between successive speech frames, we chose analysis conditions for compatibility with the Navy's 2.4-kbs LPC-10 system[22]: analysis window width = 180 points, filter order = 10, and pre-emphasis=94%. When using the length-normalized approach, N was adjusted to satisfy the normalization length requirement; however, when using the left-aligned approach, N=180 was used as is done for the Navy's LPC-10 system. The LPC analysis parameters used in classifications were always chosen to match those used in generating the code books.

F. Figures of Merit

The error rates reported in this paper are substitution error rates. We forced a choice for each utterance presented to the recognizer algorithm, and we presented only utterances that contained legitimate vocabulary words.

We used two figures of merit in evaluating the experiments. The first is simply the recognition accuracy. The second attempts to quantify the extent to which the classifications are correct or incorrect. In particular, suppose that the input utterance is the mth word in the recognition vocabulary. For correct classification, D_m should be the smallest of the average distortions (3) – i.e., $D_r = D_m$ (see (5)). Define

$$D^* = \min_{k \neq m} D_k \tag{11}$$

as the smallest average distortion of all code books except the co-rect one, and define

$$R = \frac{D^* - D_m}{D_m}. \tag{12}$$

If the classification is correct, $R>0$; if the classification is incorrect, $R<0$. For correct classifications, R is the fractional difference between the distortion of the correct code book, and the distortion of the next best choice – a large value of R means that the correct code book stands out clearly from the other choices. For each experiment, we computed the number of errors, the average value of R (R_{av}), and the standard deviation of R (R_σ).

III. EXPERIMENTAL BACKGROUND

Our experiments were conducted using a data base that was prepared by Texas Instruments, Inc. (TI) during a systematic test of discrete-utterance recognition devices [23]. A data base should be used solely for either tuning or testing a recognition algorithm. To balance the conflict between tuning and unbiased testing, we chose the following procedure. We first tuned the algorithm based on prior experience and on a speaker-independent, male-only parameter study. We then tested the tuned algorithm on the female speakers in the data base. In addition, we tested the tuned algorithm in a speaker-dependent mode on the entire TI data base.

Automatic endpoint detection for both training-sequence and classification utterances was used in our experiments. Our endpoint-detection algorithm is based on ideas presented in [24, 25], and is described in [16]. Briefly, the algorithm first analyzes the background noise to determine its average magnitude and then uses the results to set various thresholds that are used to find significant "energy clumps" in the data.

In the rest of this section we describe the data base, the experimental parameters, and the experiments.

A. TI Data Base

The TI data base [23] consists of twenty words: the digits *zero* through *nine* and the ten control words *yes*, *no*, *erase*, *rubout*, *repeat*, *go*, *enter*, *help*, *stop*, and *start*. Eight male and eight female speakers each recorded twenty-six repetitions of each word in the vocabulary, for a total of 8320 utterances. The data was recorded on analog tape under tightly controlled conditions: the noise level was low, the speech level was restricted to a ±3 dB range, the acoustic environment was unvarying, and all errors in the input words were eliminated. After collection, the data was low pass filtered and sampled at 12,500 samples per second. We received the data in digital form on magnetic tape. Each utterance, preceded and followed by short segments of ambient noise, was contained in a separate file. In a previous study using single-section code books [16], we used the data primarily at the 12,500 sampling rate. For the work reported here, the data was down sampled to 8000 samples per second. The down sampling procedure is described in [16].

B. Experimental Parameters

In this subsection, we describe the experimental parameters associated with code book generation and utterance classification. The code book generation parameters are as follows:

(a) number of utterances in the training sequence;

(b) energy threshold $E_{\min}$, where E is computed by

$$E = \sum_{i=1}^{W} x_i^2;$$

Here, W is the analysis window width, and x_i are the time-domain samples from a 12 bit A/D converter after pre-emphasis and Hamming windowing;

(c) left-alignment or length-normalized alignment;

(d) compression factor;

(e) code book type and size.

The energy threshold is used to ignore nearly-silent frames; frames with energy below this threshold are not used in designing code books or performing a classification. For all the work reported here, we used $E_{\min}$=250.

The parameters associated with utterance classification are as follows:

(a) compression factor;

(b) utterance alignment;

(c) energy threshold.

For consistency these values were chosen to match those used in the code book generation.

C. List of Experiments

In this subsection, we list the experiments reported in the remainder of the paper. The following speaker-independent experiments are listed according to the corresponding subsection of Section IV:

A. Complete male-data-base study of recognition accuracy as a function of compression factor and section code book rate;

   Comparison of recognition performance using unclustered and clustered code books when using the "best" compression factor;

   Study of recognition accuracy as a function of the normalization length;

   Recognition accuracy comparison using fixed-size and fixed-distortion code book sets;

   Recognition accuracy comparison of left-aligned and length-normalized approaches;

B. A female-only experiment using parameters that did best during the male parameter study;

C. Classification of 4 speakers using code books designed from both male and female speakers.

Section V. contains the results of speaker-dependent experiments. The experiments are listed according to the corresponding subsection of Section V:

A. Comparison of multi-section and single-section recognition performance on the sixteen speaker data base;

B. A rate-0 multi-section study;

C Recognition results for fixed-size code books with short training sequences;

Recognition results for unclustered code books with short training sequences.

IV. SPEAKER-INDEPENDENT EXPERIMENTS

In this section, we describe three sets of experiments. The first set were parameter studies done on just the male speakers – we varied the compression factor, section code book rate, utterance alignment, and code book design method. Based on the results, we give guidelines for parameter selection. In the second set of experiments, the parameters were fixed based on the results of the first set, and speaker-independent classification experiments were done for the female speakers. In the last set, a combined male and female recognition experiment was done.

A. Male Parameter Study

For all parameter studies, the LPC parameters are those specified in section II.E. We considered each of the 8 male speakers in turn. For each male speaker, we classified 520 utterances using code books designed from the first 9 utterances from each of the other 7 males. We used multiple repetitions by speakers in the training sets because of the small number of speakers we had available, not because we believe it to be an efficient way to train a recognizer.

In the first parameter study we examined the relationships among compression factor, section code book rate, and recognition accuracy. We used a 24-frame, length-normalized approach — 24 frames was approximately the average length of the words in the recognition vocabulary. We used fixed-size, section code books with rates 2,3, and 4 together with compression factors 1,2,3,4,6,8, and 12. The results are plotted in Figure 1. Note that each point on the plot represents 4160 speaker-independent classifications - 520 classifications per speaker for 8 speakers.

Based on Figure 1, we make the following observations:

(a) at each compression factor, the error spread is less than 2% for all section code book rates;

(b) the difference in error rates between section code book rates 2 and 3 is generally small, but it is consistent and significant;

(c) there is no significant difference in error rates for section code book rates 3 and 4;

(d) a *compression factor* between 3 and 6 appears best.

To gain insight into any relationship among word complexity (such as the number of syllables or phonemes), compression factor, and error rate, we examined the number of errors as a function of compression factor for the nondigit words. We had conjectured that simplier words like *no*, *go*, and *yes* would be easier to recognize using larger compression factors, and that more complex words like *repeat*, *rubout*, and *start* would require smaller compression factors. The data, however, showed no obvious correlation between word complexity, error rate, and compression factor.

Previously [16], we performed a similar speaker-independent classification experiment on these same 8 male speakers. There we used the single-section approach and the original 12500 samples per second data. The training method

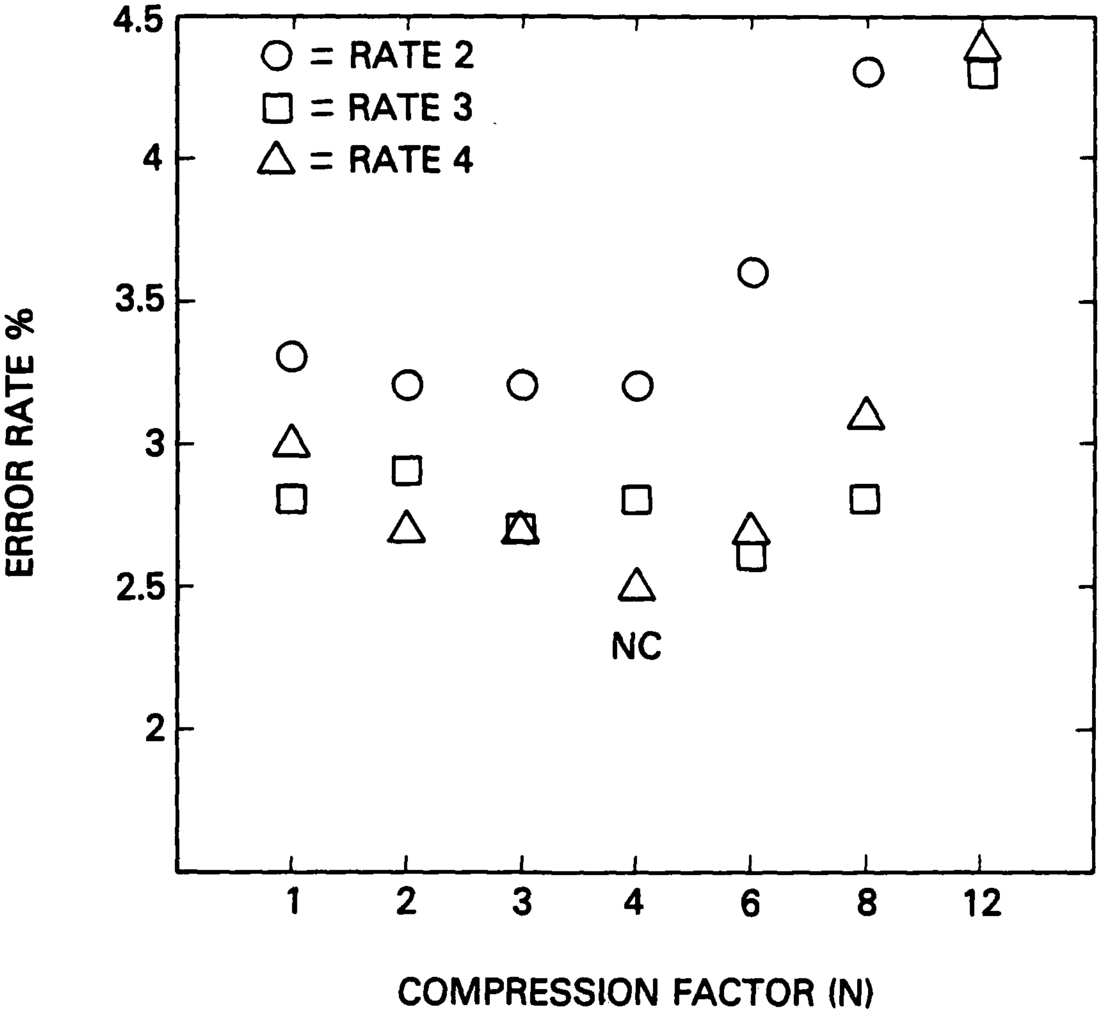

Figure 1. Relationship among compression factor, error rate, and section code book rate for speaker independent recognition.

was the same as used here: nine utterances from each of the seven speakers not being classified were used to build code books. The analysis conditions consisted of the following: N = 250 (20 milliseconds), analysis window = 250 points, filter order = 16, pre-emphasis = 90%, and Hamming windowing. As in this study, the autocorrelation method of LPC was used. Using rate-5, single-section code books, an average recognition accuracy of 88% was achieved, as opposed to the 97.5% achieved by the current approach. Thus by using the multi-section approach, the number of distortion computations per classification has been reduced (by a factor of 4 for rate-3 section code books), and the number of errors has been reduced by about a factor of 4.

As stated earlier, unclustered code books are generated by making a codeword out of each frame in the training sequence, and the effectiveness of clustering can be evaluated by comparing the performance of unclustered and clustered code books designed from the same training sequence. We built unclustered code books using a compression factor of 4 and the same LPC analysis parameters as specified for the clustered code books. The result is marked by *NC* in Figure 1. The degradation in recognition performance using rate-3 clustered code books instead of unclustered code books is small – about .5%. Since the rate-3, multi-section code books are only about 1/30 the size of the unclustered code books and the error rates for the two are close, it is apparent that the clustering procedure performs an effective data compression function.

Next we studied the effect of normalization length on recognition accuracy. We felt that, in general, longer normalization lengths would result in higher recognition accuracies. Doubling the normalization length, however, also doubles the number of distortion computations needed to compare an input utterance with a code book. We were searching for the shortest normalization length that did not significantly degrade the recognition accuracy. To study this, we chose normalization lengths of 12, 24, and 36. We used rate-3 section code books, and the compression factor was adjusted in each case so that there were 6 section code books per word. Note that for a fixed analysis window width, increasing normalization length increases the overlap between adjacent analysis frames.

The results, listed in Table I, show that the average recognition accuracy increases gradually with increases in the normalization length. The question remains, however, whether the increase is significant.

To test for statistical significance, we used the two-sample Wilcoxon rank sum test [26]. For this test, let $F(x)$ be the probability distribution function describing the recognition accuracy x of a multi-section approach with a specific set of multi-section parameters (compression factor, section code book rate, normalization length, etc.). In the normalization length study described above, let $F_s(x)$ be the probability distribution function describing the recognition performance of one of the shorter length-normalized approachs, and let $F_l(x)$ be the probability distribution function for an approach with a longer normalization length. Also, let μ_s be the mean recognition accuracy corresponding to $F_s(x)$, and let $F_l(x)$ have a mean μ_l. The null hypothesis for our test is $F_s(x)=F_l(x)$ for all x: thus, $\mu_s = \mu_l$. The alternative hypothesis is $F_s(x)=F_l(x+\Delta)$ for some positive Δ, or $F_s(x)$ is shifted to the left of $F_l(x)$. This implies $\mu_s < \mu_l$.

We performed the Wilcoxon test for all three length combinations: 12 vs. 24, 12 vs. 36, and 24 vs. 36. The significance levels for rejection of the null hypothesis of equal mean recognition accuracies were .186, .104, and .397

respectively. Based on the Wilcoxon test results and the average recognition accuracies, we believe the increase in computations in going from 12 frames to 24 frames is justified, but the increase in going to 36 frames is not justified.

Previously [16], we compared the performance of fixed-distortion and fixed-size code books using the single-section approach. Although in that study the fixed-size code books performed better than the fixed-distortion code books, we felt this might not hold true when using multi-section code books. One reason is that each section code book represents only a small portion of a word instead of the whole word as in the single section approach. This restriction might reduce the types of confusions that earlier caused fixed-distortion code books to perform worse than fixed-size code books. The possible advantages of fixed-distortion code books are that each fixed-distortion code book is only as large as necessary to satisfy the distortion criterion. Thus it follows that fixed-distortion code books might lead to the same classification performance as fixed-size code books but with fewer total codewords. This could lead to smaller memory requirements and faster classification performance.

We chose T = .45 and T = .30 as distortion thresholds, and we designed fixed-distortion code books sets using the same conditions as used in the previous fixed-size code book studies. For the T =.45 threshold, the average section code book size was 7.35 codewords; for the T =.30 threshold, it was 15.99 codewords.

The average recognition accuracy using the fixed-distortion code books with T = .45 was 96.5%. With T = .30, the recognition accuracy was 96.8%. The fixed-size, rate-3 and -4 code book sets had recognition accuracies of 97.2% and 97.5% respectively. So, as before [16], the fixed-size code books discriminate better in word recognition than do fixed-distortion code books.

So far, the experiments used length-normalized code books. We tested the left-aligned approach using a compression factor of 4, a section code book rate of 3, and, except for N (the number of points to shift between successive speech frames), the same analysis conditions as before. In the left-aligned experiment, N was fixed at 180. Left alignment was used both to design code books and to classify input utterances.

The left-aligned results together with the rate-3, compression factor 4, length-normalized results are shown in Table II. The length-normalized approach is clearly superior. This conclusion is also supported by the Wilcoxon test: the significance level is .012 for rejecting the null hypothesis of equal mean recognition accuracies.

The foregoing results suggest the following guidelines:

(a) length normalization should be used with analysis conditions that provide frame overlap;

(b) the compression factor should correspond to roughly 20% of the normalized length;

(c) fixed-size section code books of at least rate-3 should be used.

Although the speakers in these studies possessed several of the major dialects, the speaker sample was small and homogeneous - 8 male speakers living in Texas. Thus, the rate-3 section code books might be too small. In the next two sections we further evaluate this issue by studying a female speaker sample and a combined male and female speaker sample.

B. Female Results

Using a compression factor of 4 and 24-frame length normalization, we studied speaker-independent recognition using the 8 female speakers. As in the male study, we classified 520 utterances for each speaker using code books designed from the first 9 utterances of each speaker not being classified. The rate-3 and -4 results are listed in Table III. The rate-4 code books performed better that the rate-3 code books, but the difference does not appear to be significant – the Wilcoxon test yields a large significance level of .318 for rejecting the null hypothesis of equal average recognition accuracies.

The average recognition accuracy of 93.8% for females is significantly less than the 97.2% found for males. About half of the female errors, however, were for two speakers: SAS and DFG. On examining the data we found that most of the errors for DFG occurred for words on which the endpoint detector had grossly misidentified the endpoints: her voice had a breathy, nasal quality that was unlike the other speakers. This was not the case for SAS, however. There seemed to be nothing obviously unusual about her speech, yet it was difficult to recognize.

To see if the addition of new speakers to the training sequence would improve the recognition performance, we recorded data from 10 additional female speakers. The speakers were chosen arbitrarily. Each new speaker provided 1 utterance of each vocabulary word. The new data was down sampled to 8000 samples per second using the same procedure as used on the TI data, and it was added to the previous training data. No analysis or experimental conditions were changed. The results using the expanded training sequences are shown in Table IV.

The average recognition accuracy increased to 95.1% (98.5% for just the digits), but more interesting, the improvement was restricted to the two hardest speakers: SAS and DFG. Thus, adding more training data improved the recognition performance for the speakers that were poorly represented by the original training sequence and neither degraded nor improved the results for the rest of the speakers. Table V contains the confusion matrix for the female experiments using the expanded 17-speaker training sequence. Each row contains the results for classifying all utterances of one word in the recognition vocabulary: the columns correspond to the different classification decisions. The most frequent errors were *no*←→*go* and *stop*→*five*. The *no* and *go* errors were generally caused by their spectral and temporal similarities. The rest of the errors are not so easily categorized, but they usually could be attributed to inadequacies in the training data or to inaccurate endpoint detection.

C. Combined Male and Female Results

The separate results for males and females suggest that a rate-3, multisection code book is adequate for recognition purposes. This may not be the case for mixed populations, however. Because general differences in male and female vocal tracts sizes lead to characteristic formant shifts for the same speech sounds, larger code book sizes might be required to maintain performance for mixed populations. We examined this issue by performing a recognition experiment on a 4 speaker subset of the TI data base (2 males: RLD and GRD, and 2 females: SAS and ALK). We used code books designed from the remaining 12 speakers – each speaker provided 9 utterances of each word as training data.

The results for section code book rates 1 through 5 are shown in Table VI. For this small speaker sample, no significant improvement resulted from a section code book rate greater than 3. Table VII contains the individual rate-3 results for the combined male-female training data experiment and the earlier rate-3 single-sex experiments. A large increase in recognition accuracy for SAS offset small decreases in recognition accuracy for the rest of the speakers, and the average recognition accuracy using the combined-sex training sequences was about the same as that using the single-sex training sequences. The spread in recognition accuracies, however, using the combined-sex training sequences has been dramaticly reduced. The reduced spread in recognition accuracies suggests the 12-speaker training sequences characterize the general population better than the 7-speaker training sequences used earlier, and it gives evidence that increased stability of performance would result from using richer training sequences.

V. SPEAKER-DEPENDENT EXPERIMENTS

In this section, we describe the results of speaker-dependent experiments. In the first experiment, the multi-section approach was tested on the full TI data base. In the second, two multi-section rate-0 approaches were compared, and in the final experiment, the effect of short training sequences was examined. All the experiments described in this section used the 24-frame, length-normalized approach.

A. Multi-Section Results

In the speaker-independent study described in the last section, good recognition performance required a section code book rate of at least 3. It seems reasonable, however, that a smaller section code book rate might suffice for speaker-dependent recognition. To evaluate this possibility, we performed speaker-dependent recognition experiments using the 16 speakers in the TI data base For each speaker, the first 10 utterances of each word were used as a training sequence. We used a compression factor of 4 and section code book rates 0, 1, and 2.

Table VIII contains the results for all 16 speakers. The first 8 are male and the last 8 are female, and the male results are slightly better than the female results. As one would expect, the average recognition accuracy improves with increases in section code book rate. Using the two-sample Wilcoxon test to compare the rate-0 vs. rate-1, rate-1 vs. rate-2, and rate-0 vs. rate-2 results, the significance levels for rejection of the null hypotheses of equal average recognition accuracies were .138, .133, and .031 respectively. Based on the Wilcoxon test results and the average recognition accuracies, we believe the use of rate-2 section code books significantly increases the recognition accuracy compared to rates 0 and 1.

The average recognition accuracy obtained with the rate-2 section code books was 98.7%. A confusion matrix for these results is shown in Table IX. The most frequent errors were *go*←→*no*, *stop*→*five*, and *start*→*five*. Most of the *go* and *no* classification errors were due to their spectral and temporal similarities. Many of the other classification errors can be attributed to time alignment problems caused by inadequacies of the endpoint detector.

To be more specific, we examined the errors made using the rate-2 section code books: there were 66 words incorrectly classified. The endpoints had been misidentified on 42 of those 66 words. We hand labeled the endpoints on those 42 words and reclassified them in the original code books. Thirty-eight of the 42 words were now correctly identified, and the average recognition accuracy increased to 99.5%. This improvement points out the importance of accurate endpoint detection.

In our previous single-section work [16], we performed a similar speaker-dependent classification experiment on the TI data base. In that work, the 12500 samples per second data was used together with the following analysis conditions: N = 250 points, analysis window = 250 points, analysis filter order = 16, pre-emphasis = 90%, and Hamming windowing. As in this study, we used the autocorrelation method of LPC and the first 10 utterances of each word for each speaker as training data. The recognition accuracy using single-section, rate-3 code books on the full bandwidth data was about the same as using multi-section, rate-2 code books on the narrow bandwidth data: 98.8% and 98.7% respectively. Based on reductions in both the analysis filter order and the section code book rate, incorporating time-sequence information reduced the computational requirements by slightly more than a factor of 3, at the expense of doubling the memory required.

B. Rate-0 Multi-Section Study

The most remarkable aspect of the above speaker-dependent results is the high recognition accuracy of the rate-0 code books. The multi-section code book for each word consists of only 6 codewords – one codeword per section – and the classification of an input utterance requires only one distortion computation per input frame for each vocabulary word. Moreover, the code book generation consists simply of computing autocorrelations and averaging them, which is also easy to do quickly. Yet, despite these major simplifications, a recognition accuracy of 97.8% was achieved. Considering only the digits, the recognition accuracy was 99.5%.

Building references by linearly normalizing the training utterances to the same length, and then computing the average of a set of parameters for each frame in the normalized word, is an approach that many researchers evaluated before the introduction of dynamic programming and whole-utterance clustering techniques. Our rate-0, compression factor 4 (ROC4) approach is a modification of that normalize-the-utterance and average-each-frame (NUAF) approach using autocorrelations as the parameters. Because of the similarity between the two approaches, it is reasonable to ask if our ROC4 approach is any better than the old NUAF approach.

In the terminology of this paper, the NUAF approach corresponds to using rate-0, compression factor 1 (ROC1) code books. So, we designed ROC1 code books and evaluated them. Based on the speaker-independent parameter study results, we expected the larger compression factor code books (ROC4) to perform better than the smaller compression factor code books (ROC1).

Table X contains the ROC1 results along with the previous ROC4 results from Table VIII. Each compression factor 4 result is better than or equal to the compression factor 1 result except for speaker WMF, and using the Wilcoxon test on the two samples, the significance level for rejection of the null hypothesis of equal average recognition accuracies is .159. We believe the improved performance using a compression factor of 4 is because of two things: the slowly varying nature of speech spectra and the freedom from strict time alignment that a compression factor of 4 allows. Apparently, averaging the spectra in the training sequence over small sections of a word produce reference spectra that characterize a speaker's variation in pronunciation better then averaging over a single frame. Although the significance level for rejection of the null hypothesis is somewhat large, the amount of storage for each code book is reduced and the recognition accuracies are better using a compression factor of 4.

C. Short Training Sequences

Many speaker-dependent isolated word recognition devices on the market today use from 1 to 3 training utterances to train the system [27]. Although our previous results [16] suggested the inadequacy of short training sequences, we confirmed this expectation. Using a compression factor of 4 and the first 2 utterances of each word as the training sequence, we classified the same 320 utterances as above for each of the 16 speakers. The average recognition accuracies were 94.6%, 95.6%, and 95.7% for rate-0, rate-1, and rate-2 multi-section code books respectively. This is a decrease of about 3% at each rate relative to the results using 10-utterance training sequences (see Table VIII).

Finally, we performed a recognition experiment on 4 speakers using 1-utterance training sequences. We used unclustered code books to retain all the information in the training data, and we used a compression factor of 4. These results along with the 2- and 10-utterance training sequence, rate-2 results are shown in Table XI. The effect of using only one training utterance is dramatic. The average recognition accuracy for this 4 speaker subset has fallen to 90.9%.

These re-ults using short training sequences simply emphasize what is commonly known: there is much variability in a speaker's pronunciation of a particular word.

VI. COMPUTATIONAL AND MEMORY CONSIDERATIONS

It is interesting to compare the computational and memory requirements of the multi-section VQ approach to those of DTW for the classification of an unknown input utterance. As we pointed out earlier, the requirements for the DTW approach can be substantially reduced by incorporating VQ into the DTW procedure, but we do not consider that case here. Our intention is to compare the computational and memory requirements of the multi-section VQ with that of "classical" DTW [28]. Savings obtained by tracking the average distortion during classification to reject several of the hypotheses or using table-storage and look-up are also not considered.

In this analysis, we consider only the length-normalized approach. Let M be the LPC analysis filter order, N_{SC} be the number of codewords per section code book, n be the compression factor, and L_N be the normalization length. Then the memory required for a multi-section code book is

$$N_{SC}\ \mathrm{ceil}\left[\frac{L_N}{n}\right](M+1)$$

real numbers, where ceil $[X]$ is the smallest integer greater than or equal to X. Since the input word is normalized to L_N frames, classification requires $N_{SC}L_N$ distortion computations per multi-section code book.

In DTW approaches, the reference template and the input utterance are often linearly normalized to the same length L before doing DTW [28]. High recognition accuracies can then be achieved with αL^2 distortion computations per reference template, where α is in the range .20 to .35 [28]. Each reference template requires L storage locations, and to achieve high recognition accuracies, several reference templates per vocabulary word are normally stored. For speaker-dependent recognition, the number of reference templates Q, is usually one or two; for speaker-independent recognition, Q is normally about ten [29].

It follows that the ratio D of the number of distortion calculations required by the VQ approach to the number required by the DTW approach is about $D \approx N_{SC}L_N / \alpha L^2 Q$. For fixed-size code books with $N_{SC}=2^{R_{SC}}$, where R_{SC} is the section code book rate, and for a nominal value of $\alpha \approx .25$, the ratio becomes

$$D \approx 2^{R_{SC}+2} L_N / L^2 Q.$$

We shall assume that both normalization lengths are $L = L_N = 32$ frames (640 milliseconds at 20 milliseconds per frame) – this is perhaps too large, but it is conveniently a power of two. It follows that the ratio of distortion calculations becomes

$$D \approx \frac{2^{R_{SC}-3}}{Q}. \tag{13}$$

For our best speaker-dependent results – 98.7% correct using a section code book rate $R_{SC}=2$ – (13) shows the ratio of distortion computations to be $1/2Q$. Since Q is usually 1 or 2 for the speaker-dependent case, this shows that the multi-section VQ approach requires fewer distortion computations than DTW. The 98.7% speaker-dependent recognition accuracy of the multi-section approach is comparable with that achieved by other approaches on this data base [23]. For speaker-independent recognition, the multi-section approach required the rate $R_{SC}=3$. For this case, (13) shows the ratio of distortion computations to be $1/Q$. Since Q is approximately 10 for the speaker-independent case, this shows that the multi-section approach requires an order of magnitude fewer distortion computations than DTW.

The ratio W of memory locations required by the multi-section approach to the number required by the DTW approach is

$$W \approx \frac{N_{SC}\ \mathrm{ceil}\left[\frac{L_N}{n}\right]}{L_N Q},$$

where the length of a DTW reference L has been assumed equal to the normalization length L_N. Using a $L_N=32$, a $n=.2L_N$ and substituting $2^{R_{SC}}$ for N_{SC},

$$W \approx \frac{2^{R_{SC}-2.7}}{Q}. \qquad (14)$$

Equation (14) shows that for reasonable values of Q and R_{SC}, speaker-dependent recognition using the multi-section approach requires about one-half the memory that DTW requires, and for speaker-independent recognition, the multi-section approach requires only 1/8 the memory that DTW requires.

During classification, the input speech frames provide the argument f in (9). It follows that both the time-domain autocorrelations $r(n)$ and the LPC gain squared σ^2 must be known for each input frame, which in turn means that an LPC analysis must be done. For the d_{CO} distortion measure, however, the gain enters as a constant term ($\ln(\sigma^2)$) that contributes a constant term in the computation of the average code book distortions (3). The classification can therefore be done without this term, so no LPC analysis of the input utterance is required – only autocorrelations need be computed.

The software for these experiments was written in FORTRAN-77 and run on a DEC VAX11/750 with a floating point accelerator. Starting with the autocorrelations from a 63-utterance training sequence, generating the fixed-size, rate-3, multi-section code books required about 2 minutes of execution time each. Classification of a single utterance with these code books took about 0.1 second per code book – about ten times faster than our previous approach to speaker independent recognition [16]. The speedup is the result of a combination of factors: the section code books are smaller than the previous single-section code books (8 code words instead of 32 code words), the narrower bandwidth data (4000 Hz. vs 6250 Hz.) allowed a reduction in the LPC filter order from 16th to 10th, and autocorrelations were computed over a 16 millisecond window instead of a 20 millisecond window. Since all the software was designed for research purposes, specially designed programs should run considerably faster.

VII. SUMMARY AND DISCUSSION

In comparison to our previous single-section results [18], the incorporation of time-sequence information into the VQ recognition procedure has improved average recognition performance. For male speaker-independent recognition, the average recognition accuracy for the 20-word vocabulary increased from 88% to 97% with a factor of 4 reduction in computational complexity For female speakers, the average speaker-independent recognition accuracy was 95% on the 20-word vocabulary, and it was 98.5% on just the digits. For speaker-dependent recognition, the multi- and single-section approach performed approximately the same, but the multi-section approach required only half the number of distortion computations. The costs for the computational and accuracy improvements of the multi-section approach are a slightly more complicated control structure and an increase in memory for code book storage.

Perhaps the most remarkable multi-section VQ result was the 97.8% (99.5% for digits) speaker-dependent recognition accuracy for the rate-0 section code books. Only six spectra are used to characterize each vocabulary word, classification requires only one distortion computation per input speech frame per vocabulary word, and the code book design requires no clustering.

The memory requirements and computational complexity of the speaker-dependent, multi-section approach are about 1/2 to 1/4 those of the DTW approach. For speaker-independent recognition, the multi-section approach requires only about 1/8 the memory and 1/10 the distortion computations of DTW. It follows that the multi-section approach will be particularly useful when the computational and memory burden of multiple templates cannot be afforded.

As general conclusions about the multi-section VQ approach, we offer the following:

(a) all utterances should be length normalized before processing;

(b) the normalization length should be as long as computational constraints permit (up to the maximum word length expected);

(c) the analysis conditions should provide frame overlap;

(d) for speaker-independent recognition, a section code book rate of at least 3 is required;

(e) for speaker-dependent recognition, a section code book rate of at least 2 is required;

(f) short training sequences cannot be used;

(g) accurate endpoint detection is important.

The success of the multi-section approach is due primarily to two things. First, VQ code books are an efficient representation of the training data. Second, multi-section code books allow flexibility in the time alignment of an input utterance with a code book, but they enforce sectional time alignment. In fact, there is an analogy in the time alignment procedures of DTW and multi-section VQ. Neither enforces a strict sequential frame by frame comparison of the input and references, and both find locally a best path through the reference. The analogy quickly breaks down, but it is clear that the nonlinear time alignment allowed by both approaches contributes to their success.

Our results are encouraging, but they were for a small, homogeneous set of speakers. How multi-section VQ will perform on a larger, more diverse population is an open question, which we intend to investigate.

Our original single-section VQ approach tried to model each vocabulary word as a discrete memoryless source. Although the results were good, this model is, of course, naive. A better source model for an isolated word is a Markov model, and many researchers have used this idea [30, 31, 15]. Multi-section VQ is an ad hoc way of incorporating memory. It can be viewed as a one-step Markov model with transition probabilities that are either zero or one for moving to the next state or section. It would be more satisfying, and we suspect more accurate, if the states and the state representations for a word were determined by the same criterion as that used in designing a memoryless VQ code book – minimizing the distortion between the training data and the representation. Some steps in this direction have been made.

Ostendorf and Gray have developed an algorithm for designing both a separate zero memory quantizer for each of a finite number of states and a set of next-state functions depending only on the current state and codeword to update the state [32]. Using this algorithm, a separate finite-state vector quantizer could be designed for each vocabulary word, and an unknown input utterance could be classified by encoding it in each of the finite-state vector quantization code books, just as is now done with the multi-section code books. Since time-sequence information is implicit in the next-state function, and since a state code book is likely to be smaller than a section code book, the recognition accuracy should improve and the computational complexity should decrease.

Acknowledgments: We thank A. Buzo and R. Johnson for helpful discussions; we thank R. M. Gray for providing some of the vector quantization software; and we thank T. Schalk for his help in obtaining the data base.

Table I. Male Speaker-Independent Recognition: Length-Normalization Study

		Length = 12 Frames			Length = 24 Frames			Length = 36 Frames		
Speaker	No. Class.	% Correct	R_{av}	R_σ	% Correct	R_{av}	R_σ	% Correct	R_{av}	R_σ
TBS	520	97.2	.589	.359	97.7	.581	.301	97.3	.600	.310
WMF	520	94.0	.645	.464	96.7	.639	.396	97.9	.629	.405
RLD	520	95.6	.685	.516	95.4	.653	.472	96.4	.672	.482
GRD	520	93.1	.469	.342	95.8	.459	.336	95.8	.475	.338
KAB	520	95.8	.667	.480	96.0	.652	.415	95.2	.662	.455
MSW	520	98.5	.869	.562	98.1	.835	.478	98.7	.822	.452
REH	520	98.5	1.107	.575	98.9	1.030	.523	99.0	1.089	.517
RGL	520	97.9	.927	.594	99.4	.874	.487	99.8	.850	.459
all	4160	96.3	.745	.531	97.2	.715	.465	97.5	.725	.467

Table II. Male Speaker-Independent Recognition: Left-Aligned vs. Length-Normalized Code Books

Speaker	No. Class.	Left Aligned				Length Normalized			
		Errors	% Right	R_{av}	R_σ	Errors	% Right	R_{av}	R_σ
WMF	520	34	93.5	.591	.407	17	96.7	.639	.396
RLD	520	21	96.0	.629	.398	24	95.4	.653	.472
RGL	520	14	97.3	.819	.496	3	99.4	.874	.487
MSW	520	22	95.8	.517	.407	10	98.1	.835	.478
GRD	520	25	95.2	.422	.314	22	95.8	.459	.336
TBS	520	20	96.2	.584	.349	12	97.7	.581	.301
KAB	520	34	93.5	.608	.429	21	96.0	.652	.415
REH	520	25	95.2	.957	.512	6	98.9	1.030	.523
all	4160	195	95.3	.641	.448	115	97.2	.715	.465

Table III. Female Speaker-Independent Recognition: Section Rates 3 and 4

Speaker	No. Class.	Section Rate 3				Section Rate 4			
		Errors	% Right	R_{av}	R_σ	Errors	% Right	R_{av}	R_σ
ALK	520	29	94.4	.556	.389	23	95.6	.587	.407
CJP	520	15	97.1	.558	.366	13	97.5	.567	.389
DFG	520	56	89.2	.363	.320	51	90.2	.358	.305
GNL	520	33	93.7	.705	.487	21	96.0	.702	.514
HNJ	520	30	94.2	.535	.419	27	94.8	.561	.420
JWS	520	13	97.5	.813	.592	15	97.1	.795	.566
SAS	520	73	86.0	.545	.528	61	88.3	.561	.537
SJN	520	9	98.3	.644	.421	14	97.3	.683	.435
all	4160	258	93.8	.590	.465	225	94.6	.602	.470

Table IV. Female Speaker-Independent Recognition: 17-Speaker Training Data, Section Rates 3 and 4

Speaker	No. Class.	Section Rate 3				Section Rate 4			
		Errors	% Right	R_{av}	R_{σ}	Errors	% Right	R_{av}	R_{σ}
ALK	520	26	95.0	.553	.376	21	96.0	.565	.384
CJP	520	13	97.5	.542	.339	15	97.1	550	365
DFG	520	40	92.3	.352	.300	42	91.9	.348	308
GNL	520	23	95.6	.745	.515	15	97.1	.742	.489
HNJ	520	32	93.9	.546	.407	23	95.6	.581	.413
JWS	520	23	95.6	.803	.560	19	96.4	.800	.594
SAS	520	36	93.1	.569	.485	47	91.0	.557	.515
SJN	520	12	97.7	.654	.398	15	97.1	.664	.416
all	4160	205	95.1	.595	.450	197	95.3	.601	.462

Table V. Confusion Matrix for Female Speaker-Independent Recognition: Compression Factor = 4, Section Rate = 3, 17-Speaker Training Sequence

	0	1	2	3	4	5	6	7	8	9	EN-TER	ER-ASE	GO	HELP	NO	RUB-OUT	RE-PEAT	STOP	START	YES
0	204		1	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	1	
1	.	200		.	.	1	.	.		6		.	.	1		.		.		
2	2	.	200	.		.	.	.	.	.	.	.	.		8	.	.	.		
3	.	.	.	203	.	.		.	.		.	3	.	.	.	.	2	.		
4	.	.	.	.	208	.	.	.		.	.	.	.	.	.	.	.	.	.	
5	.	.	.	.	.	203	.			1	.	.				.	.	4		.
6	.	.	.	.	.	.	207		.	.	.	.	.	.	.	.	.		1	
7	.	.	1	.	.	1	1	205	.	.	.								.	
8	.	.	.	1	.	.	4	.	192	.	4	5	.	.	.		2		.	.
9	.	3	.	.	.	1	.	.	.	203	.	.	.		1	.			.	.
ENTER	1	.	.	.	.	.		.	1	.	206	.	.		.	.		.		
ERASE	.		.	.	.	.	.	.	.	.	.	208	.	.			.	.		
GO	4	1	.	.	4	.				.	.		158	4	37					
HELP	.		.	.	1	.		.			.			207		.				
NO	2		.	.	.	.	.	.	.	2	.	.	31	2	171					
RUBOUT	1				.	.	.	.	.		.	1	.	.		208				
REPEAT	.	.	.	3	.		.		.			5					200			
STOP	.	.	.	.	.	18	.	3			.	.		2	.		.	175	9	
START	1	.	.	.	1	1	1	.	.	.	.	1		.	.	.		6	197	
YES	.	.	.	.	.	.	.	.	1	.	1	3	.	.			1		.	202

Table VI. Results Using Combined Male and Female Training Data: Compression Factor = 4

Section Rate	No. Class.	Errors	% Right	R_{av}	R_{σ}
1	2080	192	90.8	.355	.305
2	2080	155	92.6	.407	.334
3	2080	138	93.4	.429	.334
4	2080	131	93.7	.457	.361
5	2080	131	93.7	.484	.388

Table VII. Comparison of Results for Single-Sex and Combined Training Data: Compression Factor = 4, Section Rate = 3

Speaker	No.	Combined-Sex				Single-Sex			
	Class.	Errors	% Right	R_{av}	R_σ	Errors	% Right	R_{av}	R_σ
RLD	520	29	94.4	.469	.344	24	95.4	.653	.472
GRD	520	38	92.7	.333	.255	22	95.8	.459	.336
SAS	520	34	93.5	.473	.372	73	86.0	.545	528
ALK	520	37	92.9	.438	.336	29	94.4	.556	.389
all	2080	138	93.4	.429	334	148	92.9	.553	443

Table VIII. Section Rate Study For Speaker-Dependent Recognition

		Comp. Fact. = 4 Section Rate = 0			Comp. Fact. = 4 Section Rate = 1			Comp. Fact. = 4 Section Rate = 2		
Speaker	No. Class.	% Correct	R_{av}	R_σ	% Correct	R_{av}	R_σ	% Correct	R_{av}	R_σ
TBS	320	98.8	1.00	.56	100.0	1.45	.77	100.0	1.69	.95
WMF	320	98.8	.95	.46	98.8	1.26	.69	99.1	1.41	.76
RLD	320	97.5	.79	.52	98.1	1.15	.78	99.4	1.35	.90
GRD	320	95.6	.73	.47	95.9	1.06	.69	96.3	1.23	.80
KAB	320	99.7	.78	.42	99.4	1.06	.59	99.7	1.22	.66
MSW	320	98.4	1.02	.51	98.8	1.51	.73	99.1	1.71	.81
REH	320	97.8	1.17	.61	98.8	1.76	89	99.1	2.08	1.03
RGL	320	100.0	1.13	.52	100.0	1.63	.79	100.0	1.89	.98
CJP	320	95.9	.94	.53	97.8	1.36	.74	97.8	1.60	.86
DFG	320	95.3	.52	.30	97.5	.84	.47	99.1	1.03	57
ALK	320	99.4	.95	.55	99.4	1.44	.84	99.7	1.79	1.01
HNJ	320	95.3	.78	.48	95.6	1.20	.74	96.3	1.45	.82
GNL	320	97.8	1.24	.96	98.8	1.79	1.29	98.8	2.19	1.49
JWS	320	98.1	.98	.57	98.8	1.63	.99	99.4	1.90	1.13
SJN	320	99.7	1.06	.61	99.7	1.60	.86	99.7	2.03	1.09
SAS	320	96.3	.83	.54	96.9	1.31	.87	96.3	1.58	.98
all	5120	97.8	.93	58	98.4	1.38	.86	98.7	1.64	1.01

Table IX. Full Data Base Speaker-Dependent Confusion Matrix: Compression Factor = 4, Section Rate = 2

	0	1	2	3	4	5	6	7	8	9	EN-TER	ER-ASE	GO	HELP	NO	RUB-OUT	RE-PEAT	STOP	START	YES
0	255												1							
1		251								3				2						
2			256																	
3				254													2			
4			1		255															
5						254												2		
6							256													
7							2	251		1				1					1	
8				1					252		1	1					1			
9			1							255										
ENTER											255								1	
ERASE										1		254					1			
GO					1								246	1	8					
HELP									1					253				2		
NO.		2											5		248	1				
RUBOUT																256				
REPEAT																	256			
STOP						8		1						2				245		
START				2		5	2					1							246	
YES																				256

Table X. Compression Factor Study For Speaker-Dependent Recognition: Section Rate = 0

		Comp. Fact. = 1 Section Rate = 0			Comp. Fact. = 4 Section Rate = 0		
Speaker	No. Class.	% Correct	R_{av}	R_σ	% Correct	R_{av}	R_σ
TBS	320	97.8	1.37	.94	98.8	1.00	.56
WMF	320	99.1	1.17	.71	98.8	.95	.46
RLD	320	96.9	1.04	.79	97.5	.79	.52
GRD	320	94.7	.97	.69	95.3	.73	.47
KAB	320	99.1	1.01	.69	99.7	.78	.42
MSW	320	98.4	1.46	.87	98.4	1.02	.51
REH	320	97.8	1.69	1.05	97.8	1.17	.61
RGL	320	99.4	1.58	.87	100.0	1.13	.52
CJP	320	95.6	1.35	.85	95.9	.94	.53
DFG	320	95.3	.73	.44	95.3	.52	.30
ALK	320	98.1	1.47	1.01	99.4	.95	.55
HNJ	320	94.1	1.13	.85	95.3	.78	.48
GNL	320	96.9	1.91	1.67	97.8	1.24	.96
JWS	320	97.2	1.51	1.07	98.1	.98	.57
SJN	320	98.8	1.67	1.11	99.7	1.06	.61
SAS	320	95.9	1.35	1.06	96.3	.83	.54
all	5120	97.2	1.34	1.00	97.8	.93	.58

Table XI. Speaker-Dependent Training Sequence Study: Compression Factor = 4

		1 Utterance Training Seq. (Unclustered)			2 Utterance Training Seq. (Clustered)			10 Utterance Training Seq. (Clustered)		
Speaker	No. Class.	% Correct	R_{av}	R_{σ}	% Correct	R_{av}	R_{σ}	% Correct	R_{av}	R_{σ}
TBS	320	95.0	1.088	.753	95.9	1.342	.900	100.0	1.694	.952
WMF	320	89.7	.894	.807	97.8	1.160	.684	99.1	1.412	.756
RLD	320	88.8	.790	.695	92.8	.945	.719	99.4	1.353	.900
CJP	320	90.3	.893	.769	93.4	1.194	.845	97.8	1.602	.861
all	1080	90.9	.916	.765	95.0	1.161	.802	99.1	1.515	.881

References

1. T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Prentice Hall, Englewood Cliffs, NJ (1971).

2. A. Buzo, A. H. Gray, Jr., R. M. Gray, and J. D. Markel, "Speech coding based upon vector quantization," *IEEE Trans. Acoust., Speech, Signal Processing* **ASSP-28**, pp. 562-574 (Oct. 1980).

3. R. M. Gray, A. H. Gray, Jr., G. Rebolledo, and J. E. Shore, "Rate-distortion speech coding with a minimum discrimination information distortion measure," *IEEE Trans. Inform. Theory* **IT-27**, pp. 708-721 (Nov. 1981).

4. G. Rebolledo, R. M. Gray, and J. P. Burg, "A multirate voice digitizer based upon vector quantization," *IEEE Trans. Commun.* **COM-30**, pp. 721-727 (April, 1982).

5. A. Gersho and B. Ramamurthi, "Image coding using vector quantization," pp. 428-431 in *Proceedings of ICASSP 1982, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Paris, France (May, 1982). IEEE 82CH1746-7.

6. R. L. Baker and R. M. Gray, "Image compression using non-adaptive spatial vector quantization," *Conf. Record of the Sixteenth Asilomar Conference on Circuits, Systems, and Computers*, pp. 55-61 (October, 1982).

7. R. Hamabe, Y. Yamada, M. Murata, and T. Namekawa, "A speech recognition system using inverse filter matching technique," *Proc. Annual Conf. Inst. of Television Engineers*, Kyushu University, (in Japanese) (June 1981).

8. J. E. Shore and D Burton, "Discrete utterance speech recognition without time normalization," pp. 907-910 in *Proceedings of ICASSP 1982, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Paris, France (May, 1982). IEEE 82CH1746-7.

9. J. E. Shore and D. Burton, "Discrete utterance speech recognition without time normalization – recent results," *Proceedings 1982 6th Int. Conf. Pattern Recognition*, pp. 582-584, IEEE 82CH1801-0 (Oct. 1982).

10. A. Buzo, C. Riviera, and H. Martinez, "Discrete utterance recognition based upon source coding techniques," pp. 539-542 in *Proceedings of ICASSP 1982, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Paris, France (May, 1982). IEEE 82CH1746-7.

11. R. Billi, "Vector quantization and Markov source models applied to speech recognition," pp. 574-577 in *Proceedings of ICASSP 1982, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Paris, France (May, 1982).

12. D. K. Burton, J. E. Shore, and J. T. Buck, "A generalization of isolated word recognition using vector quantization," pp. 1021-1024 in *Proceedings of ICASSP 1983, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Boston, MA (April, 1983). IEEE 83CH1841-6.

13. N. Sugamura, K. Shikano, and S. Furiu, "Isolated Word Recognition Using Phoneme-Like Templates," pp. 723-726 in *Proceedings of ICASSP 1983, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Boston, Mass. (April, 1983).

14. R. Pieraccini and R. Billi, "Experimental Comparison Among Data Compression Techniques In Isolated Word Recognition," pp. 1025-1028 in *Proceedings of ICASSP 1983, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Boston, Mass. (April, 1983).

15. 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," *The Bell Systems Technical Journal* **Vol. 62, No. 4**, pp. 1075-1105 (April, 1983).

16. J. E. Shore and D. K. Burton, "Discrete utterance speech recognition without time alignment," *IEEE Trans. Inform. Theory* **IT-29**, pp. 473-491 (July, 1983).

17. D. K. Burton, J. T. Buck, and J. E. Shore, "Parameter selection for isolated word recognition using vector quantization," in *Proceedings of ICASSP 1984, IEEE International Conference on Acoustics, Speech, and Signal Processing*, San Diego, Ca. (March 1984). IEEE 84CH1945-5.

18. Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.* **COM-28**, pp. 84-95 (Jan. 1980).

19 C. E. Shannon, "Coding Theorems for a Discrete Source with a Fidelity Criterion," pp. 93-126 in *Information and Decision Processes*, ed. R. E. Machol, McGraw-Hill, New York (1960).

20. 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* **ASSP-30**, pp. 294-303 (April, 1982).

21. R. M. Gray, A. Buzo, A. H. Gray, Jr., and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing* **ASSP-28**, pp. 367-376 (August 1980).

22. T. E. Tremain, "The government standard linear predictive coding algorithm: LPC-10," *Speech Technology* **1**, pp. 40-49 (April 1982).

23. G. R. Doddington and T. B. Schalk, "Speech recognition: turning theory to practice," *IEEE Spectrum* **Vol 18, No. 9**, pp. 26-32 (Sept. 1981).

24. L. R. Rabiner and M. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Syst. Tech. J.* **54**, pp. 297-315 (Feb., 1975).

25. L. Lamel *et al.*, "An improved endpoint detector for isolated word recognition," *IEEE Trans. Acoustics, Speech, & Signal Processing* **ASSP-29**, pp. 777-785 (Aug., 1981).

26. Robert V. Hogg and Elliot A. Tanis, *Probability and Statistical Inference*, Macmillan Publishing, New York (1977).

27. W. A. Lea, "Selecting the best speech recognizer for the job," *Speech Technology* **1**, pp. 10-22, 27-29 (January/February 1983).

28. C. S. Myers, L. R. Rabiner, and A. E. Rosenberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," *IEEE Trans. Acoustics, Speech, Signal Proc.* **ASSP-28**, pp. 623-635 (Dec. 1980).

29. L. R. Rabiner and J. G. Wilpon, "Speaker-Independent Isolated Word Recognition for a Moderate Size (54 Word) Vocabulary," *IEEE Trans. Acous., Speech, and Signal Processing* **ASSP-27**, pp. 583-587 (Dec., 1979).

30. J. K. Baker, "The DRAGON System - An Overview," *IEEE Trans. Acoust., Speech, Signal Processing* **ASSP-23**, pp. 24-29 (Feburary 1975).

31. F. Jelinek, "Continuous Speech Recognition by Statistical Methods," *Proceedings of the IEEE* **Vol. 64**, pp. 532-556 (April, 1976).

32. M. Ostendorf and R. M. Gray, "An algorithm for the design of labeled-transition finite-state vector quantizers," *IEEE Transactions on Communications* (1984). to appear.

9