THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 5, May–June 1983
Printed in U.S.A.

The Effects of Selected Signal Processing Techniques on the Performance of a Filter-Bank-Based Isolated Word Recognizer

By B. A. DAUTRICH, L. R. RABINER, and T. B. MARTIN

(Manuscript received December 3, 1982)

To implement an isolated word recognizer based on filter bank techniques, decisions must be made as to how to condition the speech signal prior to the filter bank analysis (preprocessing), how to condition the feature vector at the output of the filter bank analysis (postprocessing), and how to perform the time alignment in the pattern comparison between an unknown test pattern and previously stored reference patterns (registration and distance computation). In the past most designers of such word recognition systems made arbitrary choices about how the various signal processing operations were to be carried out. This paper presents results of a systematic study of the effects of selected signal processing techniques on the performance of a filter bank isolated word recognizer using telephone-quality speech. In particular, the filter bank analyzer was a 13-channel, critical-band-spaced filter bank with excellent time resolution (impulse response durations of from 3 to 30 ms) and poor frequency selectivity (highly overlapping filters with ratios of center frequency to 3-dB bandwidth of about 8 for each band). Among the signal processing techniques studied were: preemphasis of the speech signal; time and frequency smoothing of the filter bank outputs; thresholding, quantization, and normalization of the feature vector; principal components analysis of the feature vector; local and global distance computations for use in the time alignment procedure; and noise analysis in both training and testing. Each of the signal processing techniques was studied individually; hence no tests were run in which several of the techniques were used together. Results showed that some fairly simple signal processing operations provided the best overall performance in the noise-free case; in noisy conditions performance degraded significantly for signal-to-noise ratios less than about 24 dB.

I. INTRODUCTION

To implement an isolated word recognizer based on a filter analysis, decisions must be made as to how to preprocess the speech signal prior to the filter bank analysis, how to postprocess the feature vectors obtained at the output of the filter bank analysis, and how to perform the time alignment and distance computation in the pattern comparison between an unknown test pattern and previously stored reference patterns. Often such decisions are made arbitrarily based on experience, heuristic procedures, or sometimes a few brief tests with the system. To our knowledge no one has attempted to systematically examine the effects of various signal processing techniques on the performance (as measured in word error rate) of a filter-bank-isolated word recognizer. This paper provides such a comparison by examining several of the most popular signal processing techniques and showing how they affect the performance of a particular filter bank word recognizer using telephone-quality speech.[1]

There are two inherent problems with any study that attempts to find the best signal processing techniques for a system via experimental means. The first is that the results presented are highly dependent on the signal processing techniques that were studied. Hence, the "optimal" way of processing the signal may not even have been investigated (due to lack of knowledge, etc). With our limited knowledge we know of no way to avoid this difficulty. The second problem is that, of necessity, each of the various signal processing techniques is studied independently of any other (thereby tacitly assuming independence of the various methods). Hence, any interactions between the techniques studied will go unnoticed. Again, we know of no practical way of studying the interactions between signal processing operations; the processing, assuming independence of operations, took about four full months on a modern minicomputer system!

The results to be presented in this paper are an extension of a previous study[1] that examined different filter bank structures and compared their performance to that of a conventional linear predictive coefficient (LPC) word recognizer.[2,3] The key results of this earlier work were:

(*i*) The best performance in word recognition tests was achieved by both a 13-channel, critical-band-spaced filter bank, and a 15-channel, uniformly spaced filter bank. Both filter banks had composite frequency responses without gaps at the band edges. The 13-channel filter bank had highly overlapping filters; the 15-channel filter bank had filters with almost no overlap.

(*ii*) There were significant performance differences between talkers (especially female as opposed to male talkers).

(*iii*) Performance of the LPC and the best filter bank recognizers were comparable for a simple vocabulary of the 10 digits using telephone-quality speech (with no extra noise degradation) over a dialed-up telephone line using a local private branch exchange (PBX).

(*iv*) Performance of the LPC recognizer was superior to that of the best filter-bank recognizers for a complex vocabulary of the alphabet, digits, and three command words, again using telephone-quality speech.

A key question arising from these results was whether any of the proposed signal-processing techniques for the filter bank system could bring up the performance to that of the LPC system for the complex alpha-digits vocabulary. Unfortunately, we will see that none of the proposed methods was able to significantly improve filter-bank performance. (However, some were able to keep performance the same while reducing required storage.)

Two other implementational aspects of word recognizers were studied. The first involves the use of the normalize-and-warp procedure proposed by Myers et al.[4] In this procedure a fixed-length pattern is created for both test and reference patterns prior to time alignment. In this manner the largest warping area is obtained, and the computational aspects of implementing the time-warping procedure are greatly simplified. Instead of considering just the word average length for warping, we studied the effects (for both the filter bank and LPC systems) of warping to prespecified lengths of various amounts. It was found that a large amount of compression could be made before system performance degraded by a significant amount.

The second implementational aspect studied was the effects of additive noise on the performance of both the LPC and filter bank recognizers. We considered cases in which both the training and testing occurred in the noisy background, and when only the testing occurred in the noisy background. It was found that far superior performance was obtained, at all signal-to-noise ratios, when both training and testing occurred in the noisy background. Furthermore, performance of both types of recognizers degraded for signal-to-noise ratios less than or equal to 24 dB. Also, for signal-to-noise ratios greater than 6 dB, the LPC recognizer outperformed the filter bank recognizer.

An overview of the work presented in this paper is as follows. In Section II we review the general implementation of the filter bank isolated word recognizer. In Section III we discuss the signal processing methods that were studied in conjunction with the filter bank. In Section IV we discuss the noise studies. In Section V we describe the experiments performed and give word error rates for the various tests. Finally, a discussion of the results is given in Section VI.

II. THE FILTER-BANK-ISOLATED WORD RECOGNIZER

Figure 1 shows a block diagram of the overall filter bank word recognizer. The input speech signal is recorded off a dialed-up telephone line, band-limited to 3200 Hz, and digitized at a 6.67-kHz rate. The digitized speech signal, $s(n)$, is first sent to a preprocessor to condition the signal for the filter bank analyzer. Preprocessing is basically a spectra-shaping operation (e.g., linear filtering) for increased immunity to finite word-length processing in the remainder of the system.[5] The preprocessed signal, $\hat{s}(n)$, is then sent to a filter bank analyzer whose structure is shown in Fig. 2. The filter bank contains a set of Q parallel bandpass filters that cover the speech band of

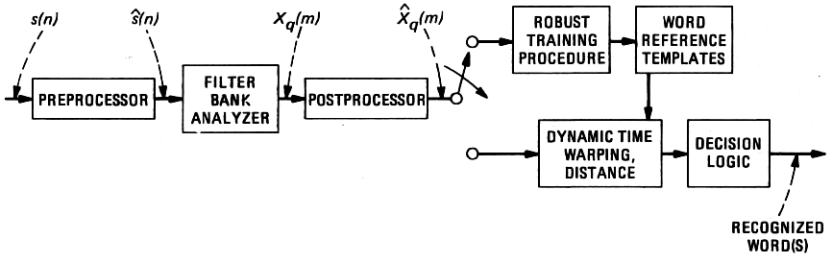

Fig. 1—Filter bank word recognizer with both preprocessing and postprocessing operations.

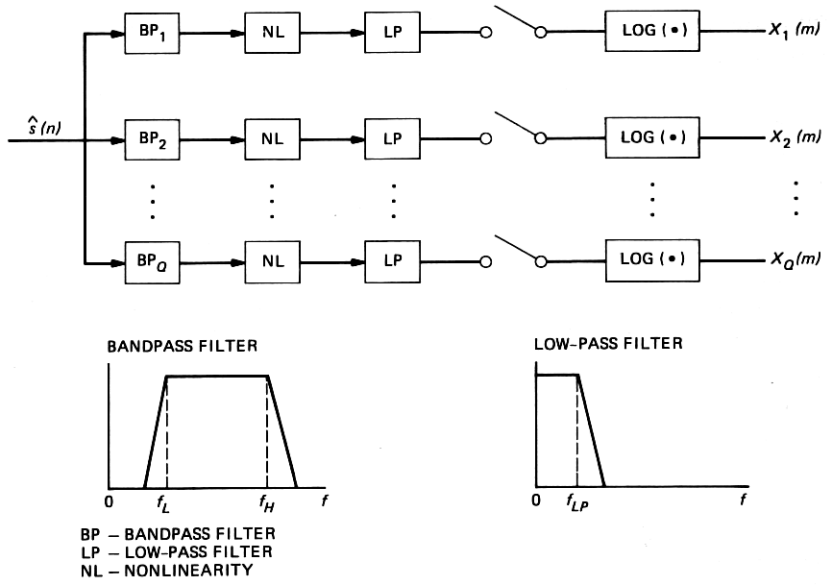

Fig. 2—Structure of filter bank analyzer.

interest (100 to 3200 Hz for telephone speech). Each bandpass filter is followed by a nonlinearity (NL), a low-pass (LP) filter, a sampler, and a logarithmic compressor. The output of the filter bank at time m is a vector

$$X(m) = [X_1(m), X_2(m), \cdots, X_Q(m)], \tag{1}$$

whose components $X_i(m)$ represent the energy in the speech signal in channel i at time m.

In our previous work[1] we studied the effects of different types of filter banks on recognizer performance and found that the highest accuracy was obtained with two types of filter banks, namely:

(*i*) A 13-channel, critical-band-spaced filter bank with higher overlapping channels. This filter bank had excellent time resolution (on the order of 10 ms) but poor frequency resolution.

(*ii*) A 15-channel, uniformly spaced filter bank with essentially no overlap between channels. This filter had poor time resolution but excellent frequency resolution. The composite spectrum of this filter bank was flat to within fractions of a dB.

Both of these filter banks used a magnitude nonlinearity and a 3-pole, Bessel, low-pass filter with a 30-Hz cutoff frequency. The sampling rate of the output feature vector was 67 Hz—i.e., adjacent feature estimates were spaced 15 ms in time.

The output of the filter bank was sent to a postprocessor, which performed one or more of the following operations:

(*i*) Time smoothing of feature vectors
(*ii*) Frequency smoothing of channels within a feature vector
(*iii*) Normalization of the feature vector
(*iv*) Thresholding and/or quantization of the feature vector
(*v*) Principal components analysis of the feature vector.

The output of the postprocessor was the input pattern to either the training mode (a robust training procedure[6]), or to the testing mode. In the training mode a set of word reference templates were created based on consistent matches of tokens of a word to previously analyzed tokens of the same word. In the testing mode the test pattern, T, consisting of the sequence of feature vectors

$$\mathbf{T} = \{\hat{X}(1), \hat{X}(2), \cdots, \hat{X}(M)\} \tag{2}$$

was compared to the reference pattern, R^i, for the ith vocabulary word using a dynamic time-warping alignment procedure.[2] For the ith reference pattern, a total average distance, D_i, between it and the test pattern was computed, and simple decision logic was used to make the word choice for recognition.

To implement the word recognizer of Fig. 1, one has to choose the types of processing to go into the preprocessor, the filter bank analyzer,

the postprocessor, and the dynamic time-warping algorithm. Based on the earlier study[1] we limited the filter bank to the 13-channel critical-band filter bank. However, for each of the remaining signal-processing blocks we tried to choose one or more possibilities and then did experiments to evaluate its usefulness to the overall word recognizer. We discuss our choices in detail in Section III. In addition we chose to study the effects of both noise addition and length quantization of both reference and test patterns on overall performance. These experiments are described in Section IV.

III. SIGNAL PROCESSING CHOICES IN THE RECOGNIZER

3.1 Preprocessor

The function of the preprocessor is to spectrally shape the speech signal to achieve some desired gross spectral shape. The most common form of preprocessing is simple preemphasis, which is used to compensate the inherent 6-dB per octave falloff in the speech spectrum. In such a case a simple first-order network of the form

$$H(z) = 1 - \alpha z^{-1} \tag{3}$$

has been found adequate[3] for recognition purposes. Thus, the difference equation relating $\hat{s}(n)$ to $s(n)$ is of the form

$$\hat{s}(n) = s(n) - \alpha s(n-1), \tag{4}$$

where a value of $\alpha = 0.95$ has been used previously.

3.2 Postprocessing

We denote the output of the qth channel of the filter bank at frame m as $X_q(m)$, $m = 1, 2, \cdots, M$, $q = 1, 2, \cdots, Q$. All of the postprocessing operations can be expressed in terms of signal processing on $X_q(m)$ to give the signal $\hat{X}_q(m)$. We have considered the operations in Sections 3.2.1 through 3.2.5.

3.2.1 Thresholding and energy normalization

The purpose of channel thresholding is to clamp low-level noise signals from channels at times when essentially no speech signal is present. This is done by applying a threshold so that channel signals below threshold are clamped at the threshold value. In this way much less sensitivity to background noise is achieved. In particular, this is achieved by determining X_q^{MAX}, the peak signal level for the qth channel, for each word as:

$$X_q^{\text{MAX}} = \max_{1 \le m \le M} [X_q(m)]. \tag{5}$$

Then the threshold for the qth channel is set at

$$T_q^* = X_q^{\text{MAX}} - T^*, \tag{6}$$

where T^* is a parameter of the recognition system. A typical range of T^* is from 30 to 50 (dB).[1] The thresholded channel signal is then given as

$$\hat{X}_q(m) = \max[X_q(m), T_q^*] \tag{7}$$

for all q and m.

The purpose of frame energy normalization is to compensate for variations in speech level from utterance to utterance. We have considered two distinct normalization methods, which we call average and peak normalization. For average normalization we calculate the frame average, $\bar{X}(m)$, as

$$\bar{X}(m) = \frac{1}{Q} \sum_{q=1}^{Q} X_q(m), \tag{8}$$

and for peak normalization we calculate the peak as

$$\bar{X}(m) = \max_{1 \le q \le Q} [X_q(m)]. \tag{9}$$

The energy-normalized feature vector is then given as†

$$\hat{X}_q(m) = X_q(m) - \hat{X}(m). \tag{10}$$

It can readily be shown that both peak and average normalization have the property that if a feature set T is derived from the speech signal $s(n)$, then the feature set T' derived from

$$s'(n) = \gamma s(n) \tag{11}$$

will be identical to T after the normalization of eq. (10) is carried out. Hence, gain variations are normalized out of the processing as desired.

3.2.2 *Time smoothing of feature vectors*

The purpose of time smoothing of feature vectors is to reduce the variability in channel outputs by averaging adjacent time frames. The cost of such smoothing is a decrease in time resolution achieved by the recognizer. If we assume that M^* adjacent frames are to be overlapped and smoothed, then time smoothing can be expressed as

$$\hat{X}_q(m) = \frac{1}{M^*} \sum_{\hat{m}=m-M^*+1}^{m} X_q(\hat{m}), \qquad m = M^*, \cdots, M. \tag{12}$$

† The reader is reminded that the channel signals are logarithmically encoded. Hence, normalization takes the form of subtraction.

It should be noted that the first $(M^* - 1)$ frames are eliminated and are used as initial conditions for the smoothing.

3.2.3 Frequency smoothing of channel outputs

As in time smoothing, the purpose of frequency smoothing is to reduce the variability in channel outputs by averaging adjacent channels for a given time frame. Again, the cost of this smoothing is a loss in frequency resolution. If we assume that Q^* adjacent channels are to be overlapped and smoothed, then frequency smoothing can be expressed as

$$\hat{X}_q(m) = \frac{1}{Q^*} \sum_{\hat{q}=q-Q^*+1}^{q} X_{\hat{q}}(m), \qquad q = Q^*, \cdots, Q. \tag{13}$$

It should again be noted that the first $(Q^* - 1)$ channels are eliminated and are used as initial conditions for the smoothing.

3.2.4 Quantization of channel outputs

The purpose of quantizing the channel outputs is to reduce the storage requirements of the recognizer both for reference patterns and for the test pattern. If we use a B-bit quantizer, and we assume the channel signals are in the range $[0, -T^*]$ because of the thresholding and energy normalization operations, then with a uniform quantizer we have a quantization width of

$$\Delta E = \frac{T^*}{2^B} \tag{14}$$

and we can express the quantized output signal as

$$\hat{X}_q(m) = \lfloor X_q(m)/\Delta E \rfloor \cdot \Delta E, \tag{15}$$

where $\lfloor x \rfloor$ is the greatest integer less than x, and it is assumed that $\hat{X}_q(m)$ is already thresholded and energy normalized.

3.2.5 Principal components analysis

The last form of postprocessing that we considered was a principal components analysis[7] in which the Q-dimensional filter bank feature vector is transformed into a new P-dimensional feature vector, where $P < Q$, such that all the important information in the original vector is retained. The purpose of reducing the feature dimensionality is to reduce the required storage for reference and test patterns. This method has also been used by Pols[7] to recognize vowels with good success.

The way in which the principal components analysis was performed was as follows. The first step is to collect a large number of filter bank

feature vectors and to compute the covariance matrix, Λ, between dimensions as

$$\Lambda_{ij} = \frac{\sum_{\cdot} [X_i(\cdot) - \bar{X}_i(\cdot)][X_j(\cdot) - \bar{X}_j(\cdot)]}{\left\{\sum_{\cdot} [X_i(\cdot) - \bar{X}_i(\cdot)]^2 \sum_{\cdot} [X_j(\cdot) - \bar{X}_j(\cdot)]^2\right\}^{1/2}}, \tag{16}$$

where the summation is over the training set of feature vectors. The principal components analysis then determines a new dimension, which is a linear combination of the original Q dimensions, that contains as much of the total variance as possible. Then a second new dimension is determined such that it is orthogonal to the first new dimension and contains as much of the remaining variance as possible. This new dimension is again a linear combination of the Q old dimensions. This process is continued until we have P new orthogonal dimensions, all of which are linear combinations of the original Q dimensions. Hence, if we denote the transformed set as $\hat{X}_{\hat{q}}(m)$, the transformation to the new dimensions is of the form

$$\hat{X}_{\hat{q}}(m) = \sum_{q=1}^{Q} \beta_{\hat{q}}(q) X_q(m), \qquad \hat{q} = 1, 2, \cdots, P, \tag{17}$$

where $\beta_{\hat{q}}(q)$ is the coefficient vector for dimension $\hat{q}$.

The resulting P-dimensional space of the principal components analysis contains as much of the total variance of the original space as is possible in P dimensions. The new space is obtained formally by doing an eigenvector analysis of the original covariance matrix, Λ. The resulting eigenvectors are the coefficient vectors for the transformation of eq. (17).

3.3 Dynamic time-warping considerations

Once the feature vectors have been obtained, the recognizer must compare the unknown test pattern, T, to each word reference pattern, R_i, $i = 1, 2, \cdots, V$, for a V-word vocabulary. For this comparison the technique of dynamic time warping (DTW) is used.[2,8,9] If we denote the test pattern, T, as

$$T = \{T(1), T(2), \cdots, T(M)\} \tag{18}$$

and the ith reference, R_i, as

$$R^i = \{R^i(1), R^i(2), \cdots, R^i(N_i)\}, \tag{19}$$

then the DTW algorithm determines a warping path

$$n = w(m) \tag{20}$$

such that the total distance, $D(T, R^i)$, defined as

$$D(T, R^i) = \frac{1}{M} \sum_{m=1}^{M} d\{T(m), R^i[w(m)]\} \tag{21}$$

is minimized, where $d(T, R^i)$ is the local distance between test and reference frames.

We have considered several variations on the conventional DTW algorithm. First we have modified the global distance of eq. (21) to include a time weighting of the form

$$\hat{D}(T, R^i) = \frac{\sum_{m=1}^{M} W^T(m) d\{T(m), R^i[w(m)]\}}{\sum_{m=1}^{M} W^T(m)}, \tag{22}$$

where $W^T(m)$ is the weight applied to the local distance at frame m. It should be noted that in eq. (22) the weight is a function of only the test pattern, T.

We have also considered a variety of types of local distance calculations of the form

$$d(T, R) = \frac{\left\{ \sum_{q=1}^{Q} (W_q^T)^p [|T(q) - R(q)|^p] \right\}^{\frac{1}{p}}}{\left[\sum_{q=1}^{Q} (W_q^T)^p \right]^{\frac{1}{p}}}, \tag{23}$$

where W_q^T is a frequency weighting curve dependent only on the test pattern, T, and p is the distance power for emphasizing the frequency variations. Typical values for p are 1 (magnitude distance), 2 (squared distance), and 1/2 (square root distance). Again, it should be noted that the frequency weight of eq. (22) is only a function of the test frame.

An alternative form of distance weighting was suggested by Silverman and Dixon[10] and is of the form

$$d(T, R) = \frac{1}{Q} \sum_{q=1}^{Q} |T(q) - R(q) - f(|\bar{T} - \bar{R}|)|, \tag{24}$$

where

$$f(y) = 1 - \frac{y}{y^{\text{MAX}}} \tag{25}$$

and y^{MAX} is the largest value that y can attain. The form of eq. (24) is similar to that of the average normalization [eq. (8)] discussed earlier in that the means of T and R (over channels) are essentially subtracted

from each $T(q)$ and $R(q)$ component. However, this is only the case when $|\bar{T} - \bar{R}| \approx 0$, in which case $f(|\bar{T} - \bar{R}|) \approx 1$. For cases in which $|\bar{T} - \bar{R}|$ is large (i.e., close to the maximum difference of T^*), then $f(|\bar{T} - \bar{R}|) \approx 0$ and no mean correction is used. Thus, the weighting of eq. (24) places extra emphasis on regions of high average energy difference, and less emphasis on regions of low average energy difference. We denote the distance measure of eq. (24) as the Silverman-Dixon (SD) distance measure.

The last variation on the conventional DTW algorithm that we have investigated is the relaxation of the endpoint constraints on the warping path. Normally we use the simple constraints that

$$w(1) = 1 \qquad \text{Initial Point} \tag{26a}$$

$$w(M) = N_i \qquad \text{Final Point,} \tag{26b}$$

i.e., the first test frame is mapped to the first reference frame and the last test frame is mapped to the last reference frame. We have considered relaxation of both endpoint constraints of eq. (26) to the form

$$1 \le w(M_B) \le \delta_{\text{BEG}}, \qquad 1 \le M_B \le \delta_{\text{BEG}} \tag{27a}$$

$$N_i - \delta_{\text{END}} \le w(M_E) \le N_i, \qquad M - \delta_{\text{END}} \le M_E \le M. \tag{27b}$$

The new endpoint constraints say that the warping path can begin anywhere within a square of size $\delta_{\text{BEG}} \times \delta_{\text{BEG}}$ at the origin of the test-reference plane, and end anywhere with a square of size $\delta_{\text{END}} \times \delta_{\text{END}}$ at the upper right-hand corner of the test-reference plane. This situation is depicted in Fig. 3. By using local path constraints, which keep the slope of the warping path greater than 1/2 and less than 2, the warping path becomes constrained to lie within the shaded area of the test-reference plane.

3.4 The normalize-and-warp procedure

The conventional DTW algorithm works quite well for most cases of interest. However, in cases when the length of the test pattern, M, is significantly different from the length of a reference pattern, N, then the region in the test-reference plane in which the warping path can lie often becomes very small. To handle such cases the normalize-and-warp procedure was devised,[4] and it basically consists of linearly prenormalizing both the test and reference patterns to a fixed length, $\hat{N}$, and then performing the DTW on these equal length patterns. In this manner the area in the test-reference plane in which the warping path can lie is maximized; hence we have the best chance of finding a good time-alignment path.

The normalize-and-warp procedure has been successfully used in a number of tests with an LPC recognizer[4,11-13] with very good results.

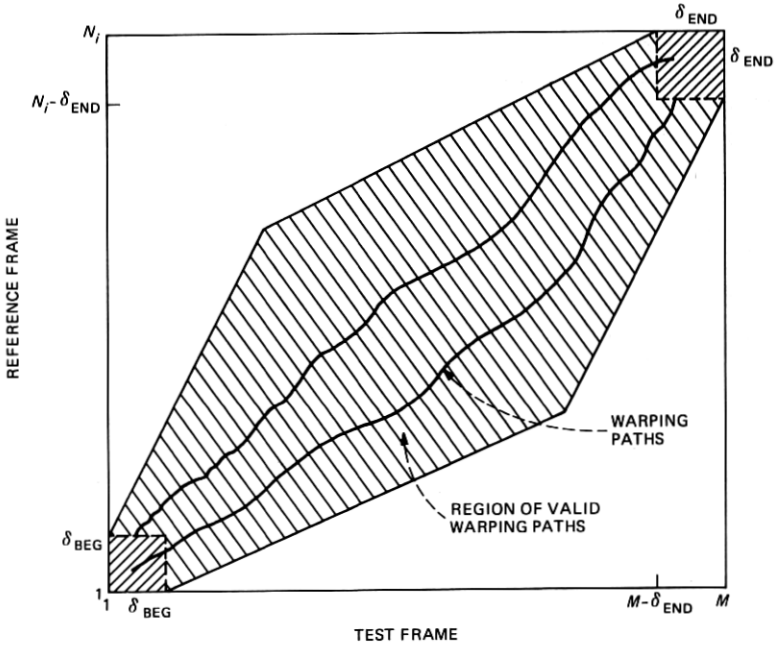

Fig. 3—The allowable warping region in the "test-reference" plane when the warping path has a beginning region of size $\delta_{BEG} \times \delta_{BEG}$ and an ending region of size $\delta_{END} \times \delta_{END}$.

For all these systems the fixed length to which all patterns were warped was the average duration of all words in the vocabulary. In this study we consider use of the normalize-and-warp procedure with the fixed length parameter a free variable.

3.5 Summary of signal processing choices

In this section we have enumerated a number of ways of implementing the signal processing of a filter bank isolated word recognizer. These factors include

(*i*) Spectral preemphasis
(*ii*) Thresholding of channel signals
(*iii*) Energy normalization of channel signals
(*iv*) Time smoothing of feature vectors
(*v*) Frequency smoothing of channel outputs
(*vi*) Quantization of channel signals
(*vii*) Principal components analysis
(*viii*) Time weighting of local DTW distances
(*ix*) Frequency weighting of channel signals in DTW computation
(*x*) Local distance metric for DTW computation
(*xi*) Loosened endpoint constraints in DTW computation

(*xii*) Use of the normalize-and-warp procedure for DTW computation.

In Section V we give the actual choices that were studied for each of the above factors. First, however, we describe the tests of the noise immunity of both filter bank and LPC word recognizers.

IV. NOISE STUDIES WITH THE ISOLATED WORD RECOGNIZER

Almost all tests of isolated word recognizers are made in a laboratory environment with a high signal-to-noise ratio on the recordings (e.g., greater than 35 dB is typical). There are a wide variety of applications (namely those of the military) in which the word recognizer is required to operate in noisy environments [e.g., signal-to-noise ratios (s/n) around 0 to 20 dB]. Thus, an important consideration in the evaluation of an isolated word recognizer is how the performance degrades as the background goes from laboratory conditions to highly noisy conditions.

When a word recognizer must operate in high-background-noise environments, an important issue arises, namely, whether it is better to train the system in a noise-free environment (and test in the noisy background), or to train and test in the noisy background. We have attempted to study these questions by artificially adding uncorrelated, zero mean, white noise to the speech signals at a specified signal-to-noise ratio, and then performing the required word recognition tests on both the filter bank and LPC word recognizers. A discussion of the test conditions and the results is given in the next section.

When one is concerned with using a word recognizer in an environment with a poor signal-to-noise ratio, another important consideration is whether one would "cancel" any of the noise by using a noise spectral estimation technique and subtracting the noise spectrum out. These techniques have been investigated in the context of voice coding[14-16] and have achieved various degrees of success. For sufficiently stationary noise backgrounds it seems reasonable to expect that a high degree of noise cancellation could be obtained. For such cases it would be of interest to understand how such noise cancellation algorithms work in the context of word recognition.

V. EXPERIMENTAL RESULTS ON ISOLATED WORD RECOGNITION

To evaluate the effects on performance (word error rate) of each of the recognition system factors of Sections III and IV, a series of tests were run with the following specifications:

Vocabulary	—	39 word alpha-digits
Number of talkers	—	2 male, 2 female
Training	—	7 replications for each word
Testing	—	10 replications for each word.

Table I—Word error rates for filter bank and LPC word recognizers

Talker	Candidate Position 1	2	3	4	5
(a) Rates for baseline filter bank recognizer as a function of talker and candidate position					
1 (Male)	9.0	4.1	0.5	0.5	0.0
2 (Male)	5.4	2.3	1.0	0.5	0.3
3 (Female)	13.1	2.8	0.5	0.3	0.3
4 (Female)	18.7	8.5	4.4	2.1	1.3
Average	11.6	4.4	1.6	0.9	0.5
(b) Rates for LPC recognizer as a function of talker and candidate position					
1 (Male)	5.1	0.5	0.0	0.0	0.0
2 (Male)	4.1	2.3	0.8	0.3	0.3
3 (Female)	10.3	2.3	1.3	1.0	0.5
4 (Female)	11.8	6.7	3.3	1.3	0.8
Average	7.8	3.0	1.4	0.7	0.4

All recordings were made over dialed-up telephone lines, and the test and training replications were obtained in different recording sessions. The speaker-dependent training used the robust training method[6] to give a single reference pattern for each vocabulary word.

The filter bank used was the 13-channel, critical-band spacing system that gave essentially the best performance in earlier tests.[1] A "baseline" filter bank recognizer was defined that had the following signal processing options:

(*i*) No preemphasis—$\alpha = 0$.

(*ii*) Channel thresholding at $T^* = 50$ dB below the peak in each channel.

(*iii*) Average energy normalization.

(*iv*) No time smoothing—$M^* = 1$.

(*v*) No frequency smoothing—$Q^* = 1$.

(*vi*) No quantization of channel signals (i.e., floating point accuracy)—$B = \infty$.

(*vii*) No principal components analysis.

(*viii*) Uniform time weighting of local distances—$W^T(m) = 1$, all m.

(*ix*) Uniform frequency weighting of local distances—$W_q^T = 1$, all q.

(*x*) Magnitude local distance—$p = 1$.

(*xi*) No opening up of DTW endpoint regions—$\delta_{\text{BEG}} = \delta_{\text{END}} = 1$.

(*xii*) No length prenormalization prior to the DTW.

(*xiii*) No additive noise—s/n(*Test*) = s/n(*Train*) = ∞.

The performance results of this baseline system are given in Table Ia and are shown plotted in Fig. 4a. Both the table and the figure show the word error rate as a function of candidate position for all four

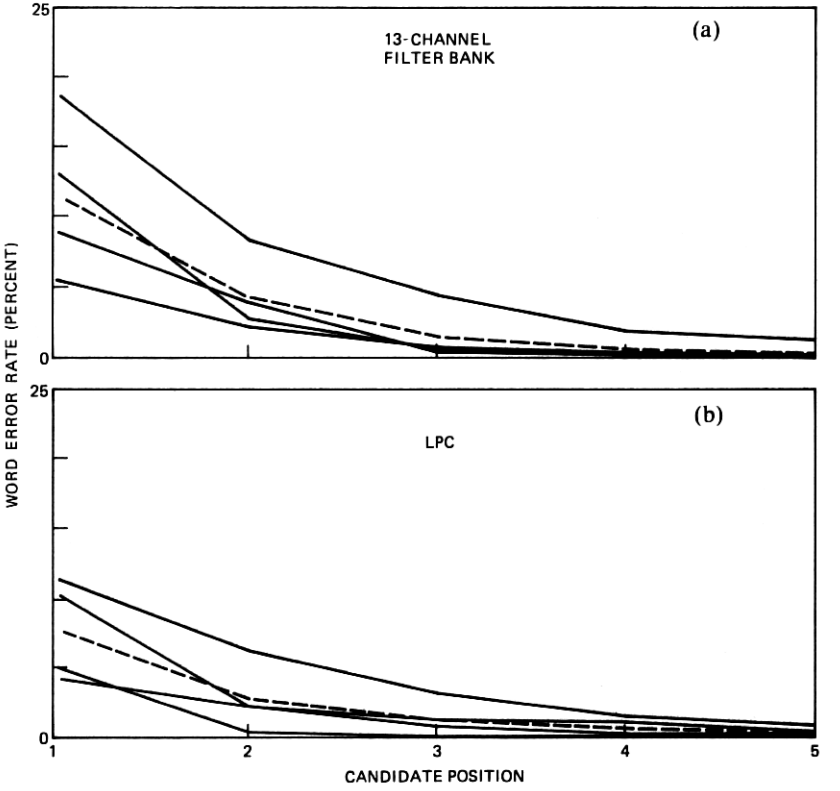

Fig. 4—Plots of word error rate versus candidate position for each of the four talkers and the average for (a) the baseline 13-channel filter bank system, and (b) the LPC word recognizer.

talkers (the solid curves in Fig. 4a) and the average (the dashed curve). These results show an average error rate of 11.6 percent in the top candidate position, with a high degree of variability in error rate across talkers. For comparison purposes, Table Ib and Fig. 4b show similar results on the LPC word recognizer. The average error rate for the top candidate position is about 4 percent lower for the LPC recognizer than for the 13-channel filter bank recognizer. Again we see a fair degree of variability in error rate scores across talkers for the LPC recognizer.

In the following sections we present results of tests designed to measure changes in performance of the filter bank recognizer as the factors noted above are varied. As discussed earlier we have been forced to use the expedient of only varying one parameter at a time; hence all information about interactions between two or more parameters is unavailable.

5.1 Effects of simple preemphasis

Only one value of the preemphasis constant, α, was studied, namely, $\alpha = 0.95$. This is the value used in previous work on the LPC recognizer.[3] The results of recognition with the use of the preemphasis network are given in Table IIa (only results for top candidate position are included). It can be seen by comparing these results to those of the baseline system that a small improvement in average accuracy was obtained. This improvement was not statistically significant at the 0.9 confidence level.

5.2 Effect of clipping threshold

The values studied for the clipping threshold were $T^* = 40$ and $T^* = 30$ (dB). The resulting recognition scores are given in Table IIb for the top candidate position. The results for $T^* = 40$ are comparable to those for $T^* = 50$ in the baseline system, whereas for $T^* = 30$ a significant loss (2.4 percent) in word accuracy is obtained. Hence, a 30-dB range is deleterious to the channel signals in that useful recognition information is lost by clamping the signals at too high a level.

5.3 Effects of peak energy normalization

The results of using peak (rather than average) energy normalization of the channel signals are given in Table IIc.[†] The results show a large increase in word error rate for all talkers, thereby indicating a lack of stability of the peak in each frame and therefore its inappropriateness to be used as an energy normalization aid.

5.4 Effects of time smoothing

The value used for the smoothing duration, M^*, were 2 and 3 (frames). The recognition results for this condition are given in Table IId for the top candidate position. For $M^* = 2$ an insignificant increase in average error rate occurred, while for $M^* = 3$ there was a very significant increase in error rate. The results show that time smoothing produced far worse scores for female talkers (3 and 4) than for male talkers (1 and 2). This was undoubtedly due to the high variability in channel signals for the females (due to the high pitch frequency), which often led to smearing a "good" frame with a "bad" adjacent frame. The results indicate that time smoothing should not be done.

5.5 Effects of frequency smoothing

The results of smoothing across $Q^* = 2$ adjacent frequency channels are given in Table IIe. It can be seen that a uniform increase of about

[†] Recall that average normalization is one of the standard options used in the recognizer.

Table II—Word error rates for several signal processing techniques in the filter bank recognizer

			Talker			
(a) Rate with preemphasis ($\alpha = 0.95$)						
α			1	2	3	4	Average
0.95			7.7	7.7	11.8	17.4	11.2
(b) Rates as a function of threshold parameter, T^* (dB)						
T^*			1	2	3	4	Average
40			8.7	6.9	13.3	17.4	11.6
30			9.5	8.5	17.2	21.0	14.0
(c) Rates for peak energy normalization						
Method			1	2	3	4	Average
Peak			23.1	19.5	14.4	40.5	24.4
(d) Rates as a function of number of frames over which smoothing occurred, M^*						
M^*			1	2	3	4	Average
2			7.9	5.6	13.1	21.3	12.0
3			7.7	5.6	26.9	29.7	17.5
(e) Rates as a function of number of frequency channels over which smoothing occurred, Q^*						
Q^*			1	2	3	4	Average
2			12.6	7.4	15.4	20.5	14.0
(f) Rates as a function of B, the number of bits used to quantize the channel signals						
B			1	2	3	4	Average
6			6.7	5.6	14.4	17.4	11.0
4			11.3	7.2	16.9	18.7	13.5
(g) Rates as a function of P, the dimensionality of the principal components analysis						
P			1	2	3	4	Average
12			11.3	9.9	25.1	29.5	18.7
6			11.3				
4			11.5				
2			18.7				
(h) Rates using a nonuniform time weighting in the DTW algorithm						
			1	2	3	4	Average
			9.2	6.5	14.5	19.5	12.4
(i) Rates using a nonuniform frequency weighting in the DTW algorithm						
Weight			1	2	3	4	Average
Fig. 5			8.7	8.7	16.2	21.3	13.7
SD			8.2	7.4	16.4	17.4	12.4
(j) Rates as a function of p, the power in the local distance computation						
p			1	2	3	4	Average
2			12.3	7.1	21.4	23.1	16.0
½			10.0	6.2	13.1	20.8	12.5
(k) Rates as a function of the opening region parameters δ_{BEG} and δ_{END} of the DTW algorithm							
δ_{BEG}	δ_{END}	Region	1	2	3	4	Average
2	0	Square	8.7	5.6	13.1	22.3	12.4
0	4	Square	7.9	5.4	13.8	21.5	12.2
2	4	Square	8.5	5.6	14.4	22.8	12.8
2	4	Line	8.9	6.2	24.6	22.1	15.5

2.5 percent in word error rate is obtained for all four talkers. Thus we conclude that smoothing across channels leads to a loss in information for recognition and therefore should not be used.

5.6 Quantization of channel signals

The results on channel signal quantization are presented in Table IIf. It can be seen that quantization of the channel signals to 6 bits actually decreases the average error rate by 0.6 percent; however, a further reduction to 4 bits leads to a 1.9-percent increase in error rate over the baseline system. Hence the results indicate that 6-bit quantization is adequate for the channel signals.

5.7 Results using principal components analysis

The results obtained using the principal components analysis for a single talker are presented in Table IIg. It can be seen that for $P = 12$ a 7.1-percent increase in average error rate is obtained; however, small increases in error rate were attained for reductions in P down to 4. In fact, for talker 1 the word error rate increased by 0.2 percent in going from 12 to 4 principal components.

An explanation of why the $P = 12$ principal components analysis led to such large increases in error rate is as follows. The transformation of the feature vector used in the principal components analysis has the property that it is invariant to a quadratic distance measure. The distance measure used in the baseline system was an absolute value distance; hence a significant decrease in accuracy resulted. We will show in Section 5.10 that using a quadratic distance measure gave much worse recognition accuracy than the absolute distance. Thus it would appear that the principal components analysis is not a useful tool, at least for this particular filter bank word recognizer.

5.8 Results using time weighting in the global distance for DTW

The results using nonuniform time weighting in the DTW global distance calculation are given in Table IIh. The actual weighting function, W^T, was a function only of the energy in the test pattern, of the type shown in Fig. 5. The test energy, E^T, was estimated as the sum of the individual channel energies. A high correlation ($\approx$0.94) was measured between this estimate of test energy, and the actual test energy (as computed from the raw speech samples). For frames in which the test energy was within 20 dB of the peak energy (suitably normalized to 0 dB), the frame weight was 1.0; for frames with E_T less than 40 dB below the peak, the weight was set to 0.01; a linear interpolation of the weight was used for frames with $-40 \leq E_T \leq -20$ dB.

The results in Table IIh show a small increase in word error rate for

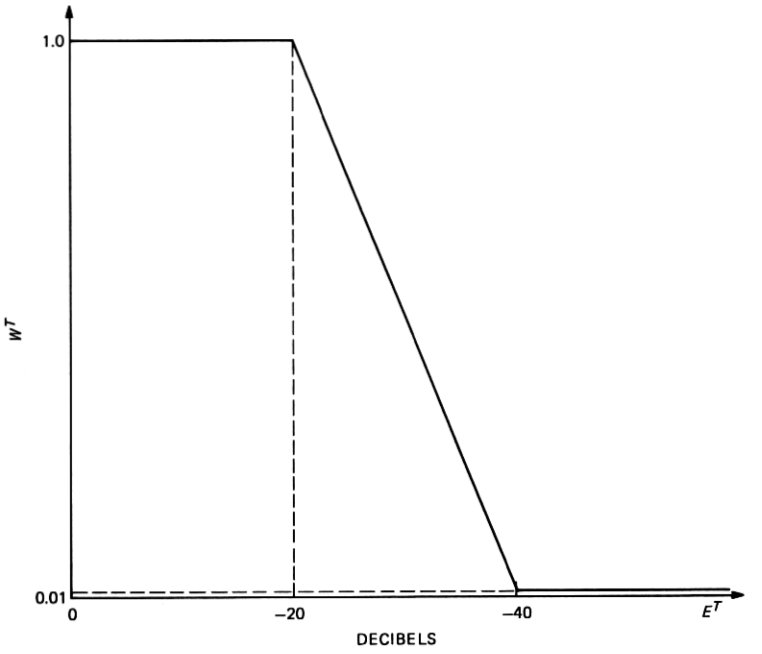

Fig. 5—The nonlinear weighting function, W^T, on the local distance as determined from the test energy, E^T, estimated from the sum of channel outputs.

each talker. Hence we conclude that the addition of time weighting (of the form of Fig. 5) in the DTW distance calculation is unnecessary.

5.9 Results using frequency weighting in the local distance for DTW

The results of using a nonuniform frequency weighting for local distances in the DTW algorithm are shown in Table IIi. The frequency-weighting characteristic was identical to the time-weighting characteristic of Fig. 5, except that the abscissa was the individual channel energy (relative to the peak channel energy for the word) and the ordinate was frequency weight W_q. It can be seen from Table IIi that a 2.1-percent increase in average word error rate is obtained using the nonuniform frequency weighting. Hence we again conclude that such weighting should not be used for this particular filter bank recognizer.

The results of using the SD weighting proposed by Silverman and Dixon (based on both reference and test frame energies) are also shown in Table IIi. Although the average word error rate is lower than for the nonuniform weighting of Fig. 5, it is still about 0.8 percent higher than obtained using simple uniform weights.

5.10 Effects of different local distance computations

The results of using different local distance computations in the DTW algorithm are given in Table IIj. Values for p of 2 (squared distance) and 1/2 (square root distance) were considered. It can be seen that for $p = 2$ a 4.4-percent increase in average word error rate is obtained; for $p = 1/2$ the increase in average word error rate is 0.9 percent. These results indicate that the magnitude distance ($p = 1$) is the best compromise between giving extra weight to very different channel energies ($p = 2$) and giving a small weight to very different channel energies ($p = 1/2$).

5.11 Results of opening up the DTW starting and ending regions

The results of opening up the beginning and/or ending region of the DTW speech regions are given in Table IIk. Results are given for an initial or final square search region, as well as for an initial or final line search region (i.e., the path had to begin or end at the first or last frame of *either* the test or reference; it could not begin or end at a noninitial or nonfinal frame of *both*). It can be seen that all the cases studied led to a small (for square regions) or a large (for the line region) increase in average word error rate. This result is anticipated from previous results, which have shown that opening up the DTW search region consistently aids false matches (reference and test different) as much or more than true matches (reference and test the same).[3,4,17]

5.12 Results using length normalization prior to DTW

The results of using fixed-length word normalization prior to the DTW (the normalize-and-warp procedure) are given in Table III and Fig. 6. Table IIIa and Fig. 6a show results for the 13-channel filter bank and Table IIIb and Fig. 6b show results for the LPC-based recognizer. The results show that for a broad range of warping lengths (from 20 to 45 frames) the average word accuracy does not change significantly. Significant degradation in performance is obtained only for the shortest warping lengths considered (i.e., 10 and 15 frames). Hence the results indicate that the normalize-and-warp procedure is suitable for a sizeable range of warping lengths so long as the length used does not become too small.

5.13 Results on noise studies

The results of the noise studies are given in Table IV and plotted in Fig. 7. Results are given for three cases:

(*i*) Signal-to-noise ratio (s/n)(Test) = s/n(Train), where s/n varied from ∞ down to 0 dB (Table IVa, Fig. 7a).

Table III—Word error rates for the normalize-and-warp procedure applied to both the filter bank and LPC word recognizers

		Length of Reference and Test					
(a) Rates as a function of warping length of reference and test prior to DTW for the 13-channel recognizer							
Talker	Variable	40	30	25	20	15	10
1	9.0	8.7	9.5	8.7	8.7	10.3	12.6
2	5.4	5.6	5.4	5.9	5.6	6.9	11.5
3	13.1	13.1	14.6	14.1	14.6	15.4	19.5
4	18.7	21.8	20.8	21.8	20.8	22.8	23.3
Average	11.6	13.1	12.6	12.6	12.7	13.9	16.7

(b) Rates as a function of warping length of reference and test prior to DTW for the LPC-based recognizer									
Talker	Variable	45	40	35	30	25	20	15	10
1	5.1	4.9	6.2	4.6	4.6	5.1	4.9	8.5	11.8
2	4.1	4.6	3.8	4.9	5.1	4.9	5.6	6.9	7.4
3	10.3	8.2	9.0	7.9	8.5	10.0	10.5	12.6	12.8
4	11.8	11.3	12.1	12.6	11.5	12.6	12.8	14.1	20.0
Average	7.8	7.3	7.8	7.5	7.4	8.2	8.5	10.5	13.0

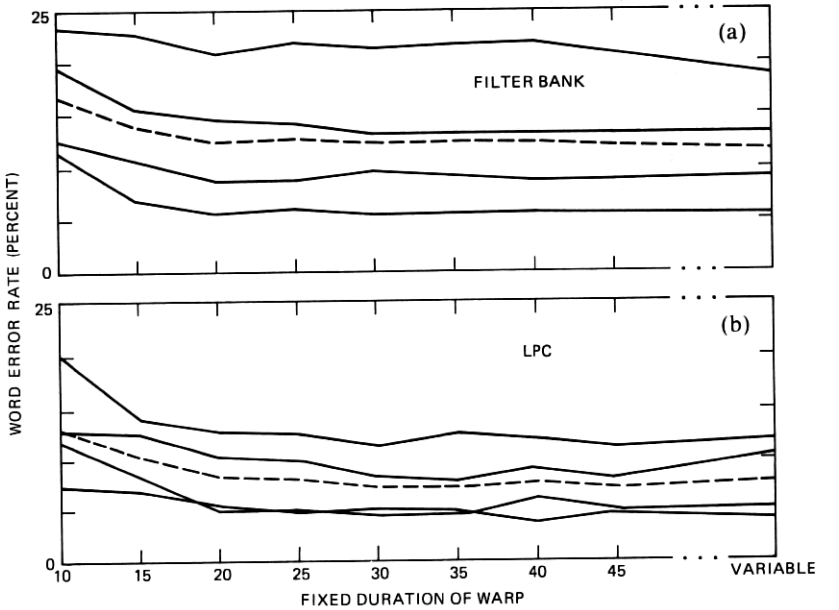

Fig. 6—Plots of word error rate versus fixed frame duration for linear prewarp prior to DTW alignment for (a) the filter bank system, and (b) the LPC system.

(*ii*) s/n(Train) = ∞, s/n(Test) variable from ∞ down to 0 dB (Table IVb, Fig. 7b).

(*iii*) s/n(Test) = 18 dB, s/n(Train) variable from ∞ down to 0 dB (Table IVc, Fig. 7c).

The first case represents the situation when both training and testing

Table IV—Word error rates for noise studies

	s/n (dB)						
(a) Rates for filter bank (FB) and LPC word recognizers as a function of s/n for case when noise added to both test and reference signals							
System	∞	30	24	18	12	6	0
FB	9.0	9.5	11.0	13.1	13.6	16.7	21.0
LPC	5.1	3.8	6.2	7.4	11.3	16.7	23.6
(b) Rates for filter bank (FB) and LPC word recognizers as a function of s/n for the case when noise added to test only—i.e., s/n of reference for training was ∞							
System	∞	30	24	18	12	6	0
FB	9.0	14.4	17.3	32.6	61.0	82.3	92.3
LPC	5.1	10.0	17.4	37.2	65.1	76.9	90.8
(c) Rates for filter bank (FB) and LPC word recognizers as a function of s/n for case when noise added to reference at variable s/n, and with test s/n set to 18 dB							
System	∞	30	24	18	12	6	0
FB	32.6	12.3	11.8	13.1	14.1	18.5	89.5
LPC	37.2	10.3	7.2	7.4	10.0	17.2	31.5
(d) Rates for filter bank (FB) and filter bank with noise removal (FB/NR) for case when noise added to reference at variable s/n, and with test s/n set to 18 dB							
System	∞	30	24	18	12	6	0
FB	32.6	12.3	11.8	13.1	14.1	18.5	89.5
FB/NR	26.2	12.6	11.8	13.1	14.4	18.7	28.2

of the word recognizer are done in the same noisy background; case (*ii*) represents the situation when there is "clean" training (no noise) but the test words are spoken in the noisy background; case (*iii*) represents the situation when there is noise in both training and testing; however there may be a mismatch in s/n.

The results in Table IV and Fig. 7 show that:

(*i*) For case (*i*), the LPC system performs as well as or better than the filter bank (FB) system for s/n $\geq$ 6 dB. The filter bank (FB) system outperformed the LPC system only at a 0 dB s/n.

(*ii*) For case (*i*), there was little degradation in performance down to s/n's of close to 24 dB for either the FB or LPC recognizer.

(*iii*) For case (*ii*) the performance of both the FB and LPC recognizers was significantly worse at all s/n's than for case (*i*). Hence we see that using clean training data with noisy test data leads to badly degraded system performance for s/n $\leq$ 30 dB.

(*iv*) For case (*iii*) the results indicate that when s/n(Test) and s/n(Train) differ by as little as 6 dB (or more) degraded performance results.

The results of Table IV and Fig. 7 indicate that it is mandatory that both the training (reference) and testing data be obtained in the same background noise conditions for best word recognition performance.

A test was also conducted on the filter bank recognizer to determine whether the effects of additive (background) noise could be lessened by subtracting out an (estimated) average noise spectrum prior to recognition. A one-second average was calculated for each channel

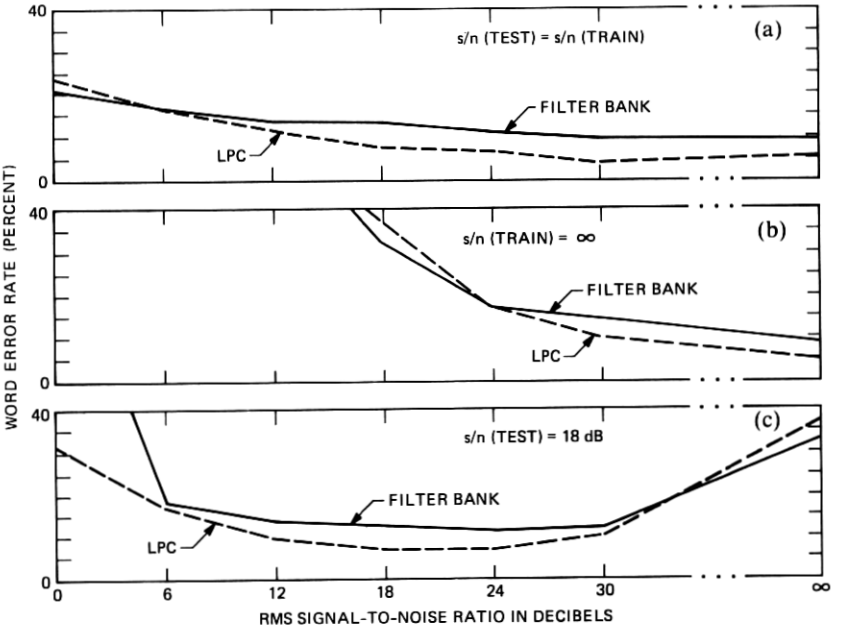

Fig. 7—Plots of word error rate (for a single talker) versus signal-to-noise ratio for both the LPC and filter bank recognizers for, (a) s/n(Test) = s/n(Train); (b) s/n(Train) = ∞, s/n(Test) variable; and (c) s/n(Test) = 18 dB, s/n(Train) variable.

signal of the filter bank when only the additive white noise was present at the input. Each of these 13-channel average noise values was then subtracted from the corresponding channel signal to form a new channel signal, which was used in the recognition processing. For these tests the signal-to-noise ratio for the test data was held constant at 18 dB, and the noise level in the reference data was varied to give signal-to-noise ratios between 0 dB and ∞.

The results of this experiment are given in Table IVd. These results show that the effects of this simple noise-cancelling arrangement are to broaden the range of signal-to-noise ratios over which the filter bank recognizer can operate. It can be seen that recognizer performance is not changed significantly between 30-dB and 6-dB signal-to-noise ratios from that obtained without the noise cancellation. However, for signal-to-noise ratios of ∞ and 0 dB, a considerable reduction of error rate is obtained with the noise cancellation method. The conclusion from this test is that noise cancelling is useful for reducing the effects of variations in noise level between reference and test data.

VI. DISCUSSION AND CONCLUSIONS

The results presented in Section V lead to the following conclusions:

(*i*) Essentially none of the proposed signal processing techniques

for use in the filter bank word recognizer led to an improvement in performance of the system (i.e., reduced word error rate). At best any single technique led to a small (insignificant) increase or decrease in word error rate; at worst it led to a significant increase in word error rate.

(*ii*) The filter bank coefficients (for telephone inputs) needed only about 6 uniform bits for a representation with no increase in word error rate. Hence, the storage requirements on the $Q = 13$ channel recognizer were about 78 bits per frame using this 6-bit coding scheme.

(*iii*) The normalize-and-warp procedure was an effective method for reducing storage and processing requirements in the DTW computation in that fixed duration linear prewarps of size as small as 20 frames per word did not increase word error rate significantly for either the LPC or FB recognizers.

(*iv*) The best strategy for using a word recognizer in a noisy background was to both train and test the recognizer in the same noise background.

(*v*) The LPC word recognizer gave error rates the same or lower than the FB word recognizer for s/n $\geq$ 6 dB.

Our initial goal was to find signal processing techniques to enhance the performance of the FB word recognizer so as to come closer to that of an existing LPC word recognizer. Our results indicate that we have not succeeded in attaining this goal. Hence our main question is whether we failed because we tried the wrong things, or because there is no way of doing consistently better with the FB limitations. There is no simple answer to this question. Perhaps our best response is that we tried a wide range of techniques that encompassed those methods previously proposed and studied in other FB recognizers. The lack of any significant improvement in performance for any of the proposed techniques indicates to us that perhaps the only way to improve accuracy is by some heuristic based on linguistic knowledge of the vocabulary words. We have meticulously avoided such techniques as they change the nature of the recognizer from a vocabulary-independent system to one that depends on the specific words to be recognized.

Another possible objection to the conclusions as drawn from the results given in Section V is that we studied each proposed signal processing technique independently. As such we avoided interactions between techniques that could have led to improved accuracy. Again we iterate our speculation that since no individual technique led to a real performance improvement, we are skeptical that combinations of techniques would lead to real improvements. Of course we have no concrete evidence that this is indeed the case.

Our noise analysis results dispel the common notion that LPC recognizers "fall apart" in noisy backgrounds while FB recognizers

degrade gracefully. Our results show that with proper training the LPC system outperforms the FB system at all reasonable signal-to-noise ratios.

Finally, the noise results show that training and testing should *always* be done in the same acoustic backgrounds. If there are gross differences in acoustic backgrounds, significant degradation in performance results.

VII. SUMMARY

We have presented results of a study to measure the effects of selected signal processing techniques on the performance of a filter bank word recognizer. We have shown that a fairly simple set of signal processing techniques led to the best overall performance of the word recognizer in the noise-free case. In noisy conditions the performance of the recognizer degraded significantly for signal-to-noise less than about 24 dB.

REFERENCES

1. B. A. Dautrich, L. R. Rabiner, and T. B. Martin, "On the Effects of Varying Filter Bank Parameters on Isolated Word Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-31*, No. 4 (August 1983).
2. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-23* (February 1976), pp. 67–72.
3. L. R. Rabiner and S. E. Levinson, "Isolated and Connected Word Recognition—Theory and Selected Applications," IEEE Trans. on Communication, *COM-29* (May 1981), pp. 621–59.
4. C. S. Myers, L. R. Rabiner, and A. E. Rosenberg, "Performance Trade-offs in Dynamic Time Warping Algorithms for Isolated Word Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-28* (December 1980), pp. 623–35.
5. J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*, New York, NY: Springer-Verlag, 1976, Chapter 9, pp. 212–26.
6. L. R. Rabiner and J. G. Wilpon, "A Simplified Robust Training Procedure for Speaker Trained, Isolated Word Recognition Systems," J. Acoust. Soc. Amer., *68* (November 1980), pp. 1271–6.
7. L. C. W. Pols, "Real-Time Recognition of Spoken Words," IEEE Trans. Comput., *C-20* (September 1971), pp. 972–8.
8. H. Sakoe and S. Chiba, "Dynamic Programming Optimization for Spoken Word Recognition," IEEE Trans. Acoust., Speech, and Signal Processing, *ASSP-26* (February 1978), pp. 43–9.
9. L. R. Rabiner, A. E. Rosenberg, and S. E. Levinson, "Considerations in Dynamic Time Warping Algorithms for Discrete Word Recognition," IEEE Trans. Acoust., Speech, and Signal Processing, *ASSP-26*, No. 6 (December 1978), pp. 575–82.
10. H. F. Silverman and N. R. Dixon, "A Comparison of Several Speech-Spectra Classification Methods," IEEE Trans. Acoust., Speech, and Signal Processing, *ASSP-24*, No. 4 (August 1976), pp. 289–95.
11. J. G. Wilpon, L. R. Rabiner, and A. Bergh, "Speaker Independent Isolated Word Recognition Using a 129 Word Airline Vocabulary," J. Acoust. Soc. Am., *72* (August 1982), pp. 390–6.
12. M. K. Brown and L. R. Rabiner, "An Adaptive Ordered Graph Search Technique for Dynamic Time Warping for Isolated Word Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-30*, No. 4 (August 1982), pp. 535–44.
13. M. K. Brown and L. R. Rabiner, "On the Use of Energy in LPC-Based Recognition of Isolated Words," B.S.T.J., *61*, No. 10, Part 1 (December 1982), pp. 2971–87.
14. S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction,"

IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-29,* No. 2 (April 1979), pp. 113–20.
15. J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," Proc. IEEE, *67,* No. 12 (December 1979), pp. 1586–604.
16. M. M. Sondhi, C. E. Schmidt, and L. R. Rabiner, "Improving the Quality of a Noisy Speech Signal," B.S.T.J., *60,* No. 8 (October 1981), pp. 1847–59.
17. L. R. Rabiner, "Note on Some Factors Affecting Performance of Dynamic Time Warping Algorithms for Isolated Word Recognition," B.S.T.J., *61,* No. 3 (March 1982), pp. 363–73.