On the Effects of Varying Filter Bank Parameters on Isolated Word Recognition

BRUCE A. DAUTRICH, LAWRENCE R. RABINER, FELLOW, IEEE, AND THOMAS B. MARTIN

Abstract—**The vast majority of commercially available isolated word recognizers use a filter bank analysis as the front end processing for recognition. It is not well understood how the parameters of different filter banks (e.g., number of filters, types of filters, filter spacing, etc.) affect recognizer performance. In this paper we present results of performance evaluation of several types of filter bank analyzers in a speaker trained isolated word recognition test using dialed-up telephone line recordings. We have studied both DFT (discrete Fourier transform) and direct form implementations of the filter banks. We have also considered uniform and nonuniform filter spacings. The results indicate that the best performance (highest word accuracy) is obtained by both a 15-channel uniform filter bank and a 13-channel nonuniform filter bank (with channels spacing along a critical band scale). The performance of a 7-channel critical band filter bank is almost as good as that of the two best filter banks. In comparison to a conventional linear predictive coding (LPC) word recognizer, the performance of the best filter bank recognizers was, on average, several percent worse than that of an eighth-order LPC-based recognizer. A dicussion as to why some filter banks performed better than others, and why the LPC-based system did the best, is given in this paper.**

I. INTRODUCTION

SINCE the early 1970's, researchers have been working on building machines that have the ability to communicate with man in his natural method of communication. One research area that has developed from this work is that of speech recognition. The general goal of speech recognition is to understand normal human speech and then to be able to perform some task based on this understanding. This is a very natural goal in that it requires machines to adapt to humans rather than vice versa. In this way speech recognition would provide a convenient method of communication with machines (e.g., computers) via terminals and ordinary telephone handsets.

Progress has been made toward the general goal of speech recognition by imposing some restrictions on the speech input. These restrictions are usually in the form of limits placed on the vocabulary, the set of allowable users, or the mode of the input. The purpose of this last limitation (probably the most severe one) is to restrict the form of input speech to a set of isolated word commands, instead of continuous speech, in order to achieve reliable recognition. With these restrictions speech recognition has made major strides forward in the past decade and several commerical systems have appeared [1]-[6]. These systems are predominantly isolated word speaker-trained systems. The availability of these systems has led to an increased interest in the possibility of producing terminal equipment that uses this new technology. In this type of environment, what is needed is a recognition system that is modular, low-cost, and highly accurate. At present none of the available systems fulfill all of these requirements. However, in the near future, with improvements in both the algorithms used to perform the recognition task and the integrated circuit technology used to construct these systems, practical systems will clearly become feasible.

The speech recognition system developed in the Acoustics Research Department at Bell Laboratories [7]-[10] meets two of the three essential requirements for the terminal equipment market. That is, the system is modular and highly accurate. However, the system is still too costly for practical and widespread use. In order for the Bell Laboratories system to be applicable in standard communication tasks, the cost of the system must be reduced substantially. There are two possible ways of accomplishing this cost reduction. The first of these is to change the recognition algorithm to reduce system cost while trying to maintain high accuracy. An alternative is to change the hardware used by the system to take advantage of current integrated circuit technology.

This paper describes work on an alternative low-cost feature analysis system for the Bell Laboratories word recognizer. The current recognizer uses linear prediction coefficients (LPC) to represent the input speech. At present, the calculation of these coefficients requires relatively expensive hardware (≈$100-1000). The proposed method of reducing the cost of the recognizer is to replace the LPC representation by a filter bank analysis to represent the input speech. Because it is possible to construct an inexpensive (≈$10) integrated circuit version of a filter bank, this alternative is very attractive [11].

Besides its low cost, there are two good reasons for examining filter bank recognizers. First, the ear is known to process speech using a structure similar to a filter bank (albeit with highly nonuniform filter spacing and with filter characteristics that would be difficult to match with conventional filters) [12]. Second, previous research has shown that an LPC-based recognition system and a filter bank recognition system [using wide-band speech (0-10 kHz)] could achieve essentially identical recognition accuracy on some standard word vocabularies [13]. This suggested that implementation of the less expensive filter bank system need not necessarily cause a degradation in recognition accuracy.

Although a number of different filter bank structures have been proposed for recognition, there is no simple guideline for choosing an optimal filter bank for a particular application. By this we mean that, to date, there have been only a small set of comparisons of the effects on performance (word error rate) of different filter bank structures in an automatic speech

Manuscript received August 9, 1982; revised November 19, 1982 and February 28, 1983.

The authors are with Bell Laboratories, Murray Hill, NJ, 07974.

0096-3518/83/0800-0793$01.00

recognizer [14], [15]. Even simple questions such as the type of filter bank (FIR or IIR filters), the filter spacing (uniform or nonuniform, nonoverlapping or overlapping), the number of filters, the filter types, etc., have not been systematically investigated for any common vocabulary or recognition system. Other important questions of interest are the ways in which filter bank feature sets are preprocessed and postprocessed for use in conventional dynamic time warping (DTW) structures. The purpose of this paper is to investigate several of the issues mentioned above in order to more fully understand the factors affecting performance of filter bank analyzers in word recognition systems.

An overview of the work presented in this paper is as follows. In Section II the general implementation of a filter bank feature measurement system is described. This section includes a brief description of the overall recognizer as well as the filter bank analyzers. The details of the implementation structures for the different filter banks are discussed in Section III. This is then followed by a description, in Section IV, of the particular filter banks that were studied. In Section V we describe the experiments performed to evaluate the word recognizer, and give word error rate scores for several filter bank analyzers. Finally, a discussion of the results is given in Section VI.

II. General Implementation of a Word Recognizer Based on Filter Bank Feature Sets

Fig. 1 shows a block diagram of a filter bank feature measurement system [16]. In this system the speech is first passed through a bank of Q bandpass filters. For the purposes of this study all filters were implemented digitally to faciliate making design changes. This bank of bandpass filters separates the frequency spectrum of interest (in our case 100-3200 Hz since we are concerned primarily with telephone-based systems[1]) into various frequency bands. In existing systems (i.e., commerical or experimental word recognizers) the number of filters (Q) has varied from 3 to 32. In part, this is because many of the filter bank analysis systems for speech recognition are based on designs used in vocoder applications. The spacing of these filters is usually such that they are continuous over the frequency spectrum and the composite spectrum of the overall filter bank is essentially flat (i.e, no sharp valleys between adjacent filters). This assures that equal weighting is given to all frequencies of the spectrum of interest. The frequency spacing of the filters in the filter bank can be determined in a number of ways. A fairly standard technique is to divide the frequency spectrum uniformly and to space the filters on a uniform frequency scale. Other possibilities are to space the filters equally on a logarithmic frequency scale (e.g., octave or $\frac{1}{3}$ octave spacing) or along a frequency scale related to a speech information measure such as the articulation index.

As shown in Fig. 1, the output of each bandpass filter is generally passed through a nonlinearity such as a square-law detector or a full-wave rectifier. This nonlinearity has the effect of nonuniformly distributing the original band-limited

[1]It should be noted that very few commerical word recognizers have been designed for use over the telephone system.

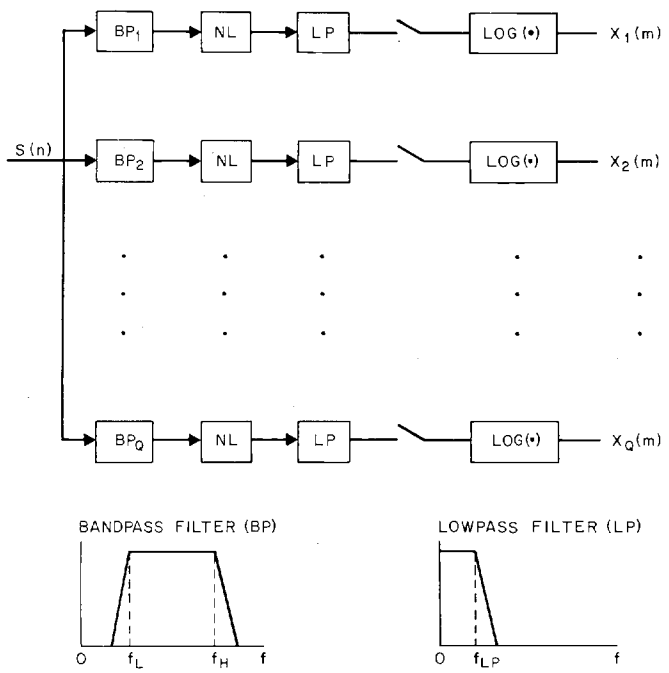

Fig. 1. Block diagram of filter bank feature measurement system.

signal energy over the entire frequency spectrum. However, the signal energy at low frequencies is generally proportional to the total band-limited signal energy. Thus, when this nonlinearity is followed by a low-pass filter, the output of the low-pass filter is a measure of the energy of the speech signal in the particular frequency band. This low-pass filter is then decimated at a rate twice that of the low-pass filter cutoff frequency, typically 40-60 Hz. For purposes of dynamic range compression, the energy is encoded by a logarithmic transformation. The set of energy values, at each instant of time, constitute a Q-dimensional feature vector. The time variation of these feature vectors defines a pattern for the speech. Thus, if we denote the signal, at time m, for filter channel i, as $X_i(m)$, then the feature vector at time m is

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

and the pattern T, defined for $m = 1, 2, \cdots, M$, is

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

Fig. 2 shows a block diagram of the overall word recognition system based on a filter bank analysis. Following filter bank analysis the pattern is subjected to a postprocessor which provides some time and/or frequency normalization to the filter bank output vectors. Although a variety of techniques could be used in the postprocessor, we have used only two of them, namely channel thresholding (to further reduce dynamic range of the channel energies) and energy normalization. We describe these procedures in Section II-A.

Following postprocessing, the modified pattern is either used in a training mode to obtain speaker-dependent, word reference templates, or in a testing mode to compare against stored reference patterns (using a DTW alignment and distance algorithm), and to give word distance scores which are then passed on to a decision box to choose the "recognized" word. The word endpoint detector, training procedure, DTW alignment

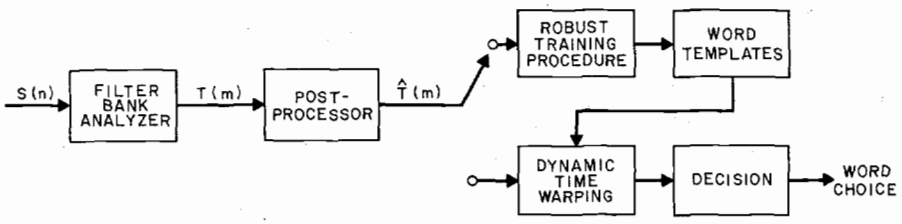

Fig. 2. Block diagram of word recognition system.

algorithm, and decision boxes are similar to those used previously in the LPC-based word recognizer [7]-[10].

A. Channel Thresholding and Energy Normalization

As mentioned above, the first step in the postprocessing of the channel signals is to apply a threshold to limit the dynamic range of the channel signals. The purpose is to prevent a channel signal from varying too much at times when essentially no speech signal is present in that band. At such times the channel output varies greatly depending on the background noise level. By applying a threshold so that signal levels below the threshold are clamped at the threshold value, much less sensitivity to background noise is achieved.

In a practical isolated word recognizer, channel thresholding is achieved automatically by the quantization inherent in each channel of the recognizer. We have used the channel clamping at a fixed threshold to model just the finite dynamic range of such a practical quantizer. We have found, in practice, that for clamping thresholds on the order of 50 dB below peak signal, as used here, the clamping threshold is applied only rarely, and has the effect of clamping large spectral differences for extremely low level channel signals.

In particular, for each channel and for each word, the peak signal level in each band, $X_i^{\max}$ is obtained as

$$X_i^{\max} = \max_{1 \le m \le M} [X_i(m)] \tag{3}$$

and the threshold for the ith channel is set at

$$T_i^* = X_i^{\max} - T^* \tag{4}$$

where T^* is a parameter of the recognition system. For telephone inputs, where the average signal-to-noise ratio is about 35 dB, and typical peak signal-to-noise ratios are 50 dB, a value of $T^* = 50$ (dB) is used. The specific choice of T^* is not terribly important as long as T^* is in the range of 50 (dB). The major effect of a finite value of T^* (rather than $T^* = \infty$) is to eliminate gross recognition errors due to widely varying signal energies in bands with no speech energy.

Fig. 3 illustrates the use of thresholding on a typical channel output. Fig. 3(a) shows the original channel energy (for a typical channel for the word /REPEAT/) and Fig. 3(b) shows the thresholded channel output (in (b) the peak energy is normalized to 0 dB). It can be seen that three regions of the pattern were clamped at the threshold, two within the word (from T_1 to T_2 and from T_3 to T_4) and one at the end of the word (from T_5 to T_6). It was found that without such thresholding, errors in matching in the regions of low energy (the clipping regions) led to gross recognition errors in some cases.

The second function of the postprocessor is a level normalization to compensate for variations in speech level from

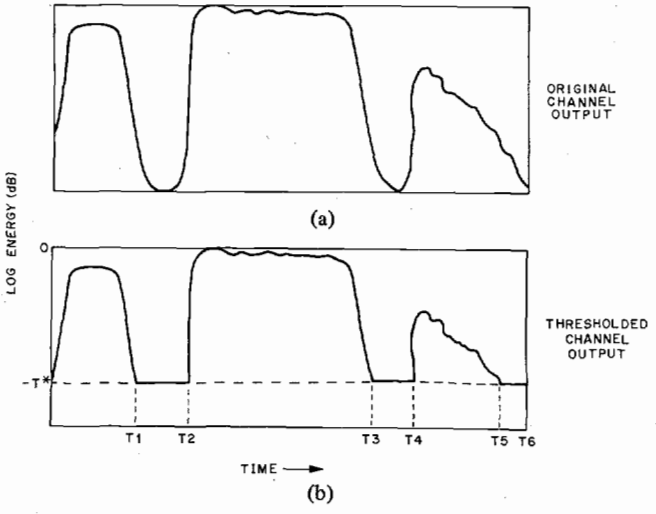

Fig. 3. Example of channel thresholding for the word /REPEAT/.

utterance to utterance. For each frame $X(m)$, the average value $\bar{X}(m)$ is calculated as

$$\bar{X}(m) = \frac{1}{Q} \sum_{i=1}^{Q} X_i(m) \tag{5}$$

and the "average normalized" feature vector $\hat{X}(m)$ is given as

$$\hat{X}(m) = X(m) - \bar{X}(m) \tag{6}$$

with components

$$\hat{X}_i(m) = X_i(m) - \bar{X}(m). \tag{7}$$

It should be clear that if a feature set T is derived from the speech signal $s(n)$, the feature set T' derived from

$$s'(n) = \alpha s(n) \tag{8}$$

is identical to T after the average normalization of (5)-(7) is carried out. Thus, this processing correctly handles simple gain variations. It should also be clear that the ordering of the thresholding and average normalization operations is important. The thresholding procedure should be done before the average normalization to more closely approach equal average levels for both reference and test frames.

The reader should be aware that the mean normalization technique described here has been found to perform well in practice with our recognition system. However, we make no claims as to its optimality; hence, alternative normalization schemes may work as well or better for this and other recognizers.

B. Modifications of the DTW Algorithm

As mentioned previously, the DTW algorithm designed for the LPC-based feature set was used in the filter bank recognizer. Only one simple modification was required in the local distance calculation to replace the log likelihood distance measure used for LPC coefficients. The local distance between the test T at frame m, and a reference R at frame n is the simple absolute value distance

$$d_{mn} = d(T(m), R(n)) = \sum_{i=1}^{Q} |T_m(i) - R_n(i)|. \tag{9}$$

This distance metric was chosen because of the simplicity of implementation (no multiplies, Q additions and subtractions) and because it has been shown to work well in practice [17], [18]. Alternative distance metrics for speech recognition have been proposed by Klatt [19], but have not been studied here.

III. General Design of the Analysis Filter Banks

A variety of considerations goes into the choice of filters for the filter bank of Fig. 1. The first issue that had to be resolved was the type of filter used for the bandpass filters in the structure. The possible choices include finite impulse response (FIR) and infinte impulse response (IIR) filters. Because of their linear phase properties and because simple implementations are possible, FIR filters were chosen for the bandpass filters [20].

Once we have decided on using FIR filters for the bandpass filters, the next question is the number of filters, Q, and the filter spacing. The choice of a value for Q depends upon the intended application of the spectrum; values of Q from 10 to 32 have typically been used in vocoder applications [21]. For (axis crossing) estimation of formant frequencies or gross measures of the spectrum, smaller values of Q (3-7) have been used. For recognition purposes it is not clear just how many filters are required.

The second issue in the design of the analysis filter bank is the filter spacing. One standard method is to design a uniform filter bank in which the center frequency f_i of the ith channel is

$$f_i = \frac{F_s}{N} \cdot i \qquad i = 1, 2, \cdots, Q \tag{10}$$

where F_s is the sampling rate of the input, N is the number of filters that span the baseband frequency of the signal, and Q satisfies the property

$$Q \leqslant N/2 \tag{11}$$

since channels for $i > N/2$ are mirror images of those for $i < N/2$.

An alternative filter bank design is to choose channel bandwidths equally spaced on a logarithmic frequency scale. If we define channel bandwidths ΔF_i as

$$\Delta F_i = \alpha \Delta F_{i-1} \qquad i = 2, 3, \cdots, Q \tag{12a}$$

$$\Delta F_1 = C, \tag{12b}$$

then channel center frequencies are given as

$$F_i = \sum_{j=1}^{i-1} \Delta F_j + F_0 + \frac{\Delta F_i}{2} \tag{13}$$

where F_0 is the lower frequency of the first band. Fig. 4(a) illustrates an octave band filter bank design ($\alpha = 2$) for $Q = 4$ with $F_s = 6400$ Hz. For this example $C = 200 = F_0$. Clearly, F_0 could be lower (i.e., the first channel could go from say 100 to 300 Hz) and all other properties of the filter bank would be preserved. In Fig. 4(b) a 12-channel $\frac{1}{3}$ octave ($\alpha = 4/3$) filter bank design is also given with $F_0 = 200$, and $C \cong 50$.

Alternative ways of choosing filter spacings are available including the so-called critical band [22] filter banks (with channels uniform until about 1000 Hz and then logarithmic above 1000 Hz), and arbitrarily spaced filter banks where other considerations are used in designing the individual filters. Fig. 5 illustrates a 7-channel critical band filter bank design. We will discuss these filter banks more in Section IV.

Once we have designed the necessary bandpass filters, the next step is to choose the nonlinearity and design the required low-pass filter. The nonlinearity chosen for this study was a full-wave rectifier. This is standard for most filter bank applications. For the low-pass filter, an infinite impulse response (IIR) filter was chosen because of the narrow bandwidth of the filter. An FIR filter would have required a prohibitively long impulse response. The cutoff frequency of the low-pass filter was chosen to be 30 Hz to allow for sampling the channel outputs at a rate of 67 Hz. The desired low-pass filter was realized using a third-order Bessel IIR filter. The impulse and frequency responses of this filter are shown in Fig. 6.

A. Implementations of the Bandpass Filters

Since each of the bandpass filters is an FIR design, the entire analysis filter bank can be implemented in a direct form structure in which the bandpass output signal is obtained as the convolution of the input signal with the filter impulse response. Thus, if we define the input to the ith bandpass filter as $s(n)$, the impulse response of the filter as $h_i(n)$, and the output $y_i(n)$, we get

$$y_i(n) = \sum_{m=0}^{L_i - 1} h_i(m)\, s(n - m) \tag{14}$$

where L_i is the duration of the impulse response $h_i(m)$. Technically $y_i(n)$ need only be computed at a rate twice the bandwidth of the bandpass filter. However, if we use highly nonuniform bandwidths for the individual filters, and if we desire to keep the sampling rates of each channel the same, then we can only achieve a 2-to-1 reduction in sampling rate for some of the filter banks (in particular the octave-spaced design). For such filter banks there is very little gain in lowering the sampling rate of the bandpass filter outputs. Since we follow the bandpass filtering by a nonlinearity, it was decided to keep the sampling rate at the original sampling rate to minimize the frequency distortions due to the nonlinearity.

For the uniform filter banks, however, an alternative, more efficient implementation is possible using discrete Fourier transform (DFT) structures. Each bandpass filter response

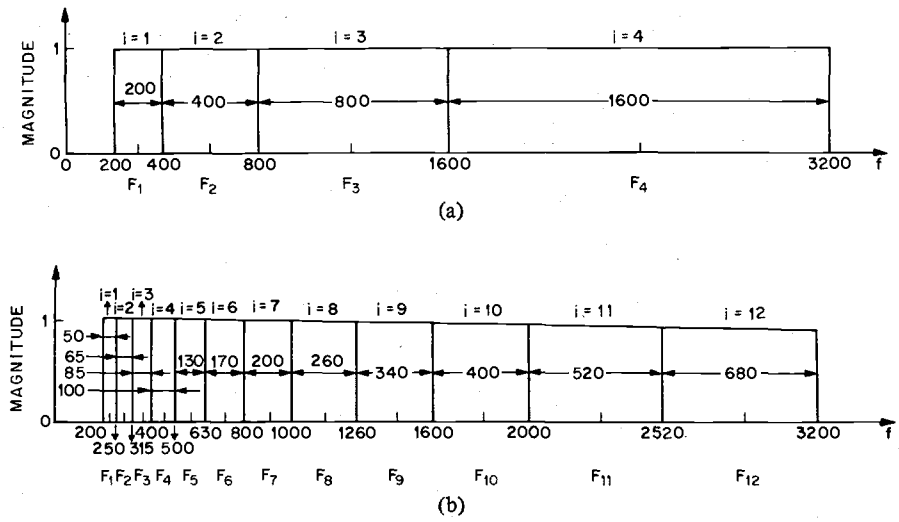

Fig. 4. Ideal octiave and $\frac{1}{3}$ octave filter banks over baseband of interest (200~3200 Hz).

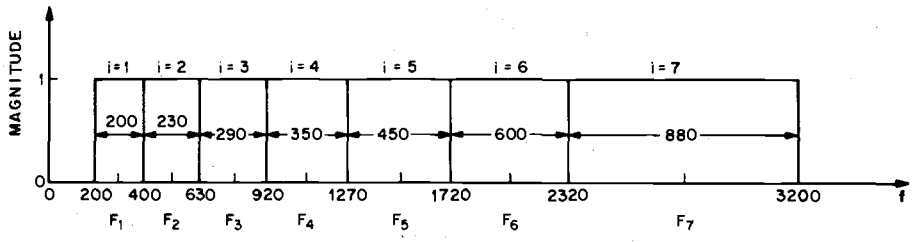

Fig. 5. Ideal critical band spaced filter bank over baseband of interest (200–3200 Hz).

$h_i(n)$ can be represented as a frequency modulated low-pass filter, i.e.,

$$h_i(n) = w(n)\, e^{j2\pi f_i n} \tag{15}$$

where $w(n)$ is the FIR prototype low-pass filter, and the factor $e^{j2\pi f_i n}$ modulates the center frequency of the filter from 0 frequency to frequency

$$f_i = \left(\frac{F_s}{N}\right) \cdot i. \tag{16}$$

Then we can write the response of the ith bandpass filter as

$$y_i(n) = \sum_{m=-\infty}^{\infty} x(m)\, w(n-m)\, e^{j(2\pi/N) F_s i(n-m)} \tag{17a}$$

$$= e^{j(2\pi/N) F_s i n} \sum_{m=-\infty}^{\infty} [x(m)\, w(n-m)]\, e^{-j(2\pi/N) F_s i m}. \tag{17b}$$

Equations (17a) and (17b) show that $y_i(n)$ could be computed by multiplying the input signal $x(n)$ by the time reversed and shifted window $w(n-m)$, taking an N-point DFT of the product, and then modulating the resulting signal by the factor $e^{j(2\pi/N) F_s i n}$ (to give a bandpass signal). If we consider all N DFT outputs of (17) we see that the terms for $i = 1, 2, \cdots, Q$ define the Q desired filter outputs (in complex form) and the terms for $i = 0$ and for $i > Q$ are for frequency bands that are of no interest. Hence, we have an efficient implementation of a Q channel uniform filter bank if the center frequencies obey (16), if the filters obey (15), and if the required value of N is a power of 2.

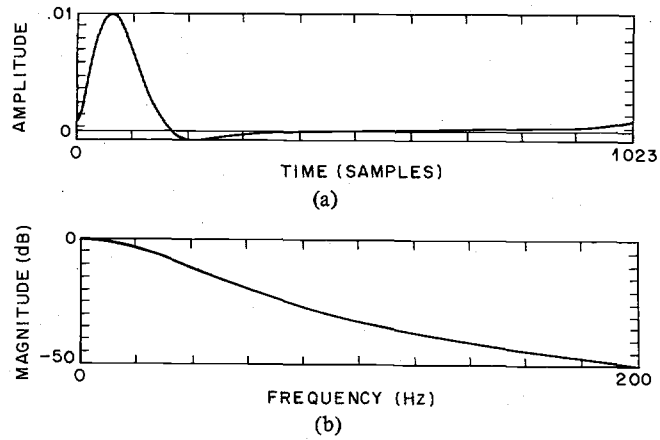

Fig. 6. Time and frequency responses of low-pass filter.

Fig. 7 shows a summary of the required processing for implementing the uniform filter bank based on the DFT structure. The input signal is multiplied by the time reversed and shifted low-pass window w, an FFT is taken, and the complex modulation factors are applied. There is a consideration that needs to be discussed to complete the implementation. If we define the length of the prototype low-pass filters as L samples, then when $L > N$ special care must be exercised to create the N-point signal needed for the DFT implementation. In these cases Schafer and Rabiner [17] have shown how the L-point signal $x(m)\, w(n-m)$ can be time aliased onto itself to give the N-point signal $\tilde{x}(m)$, by forming the sequence

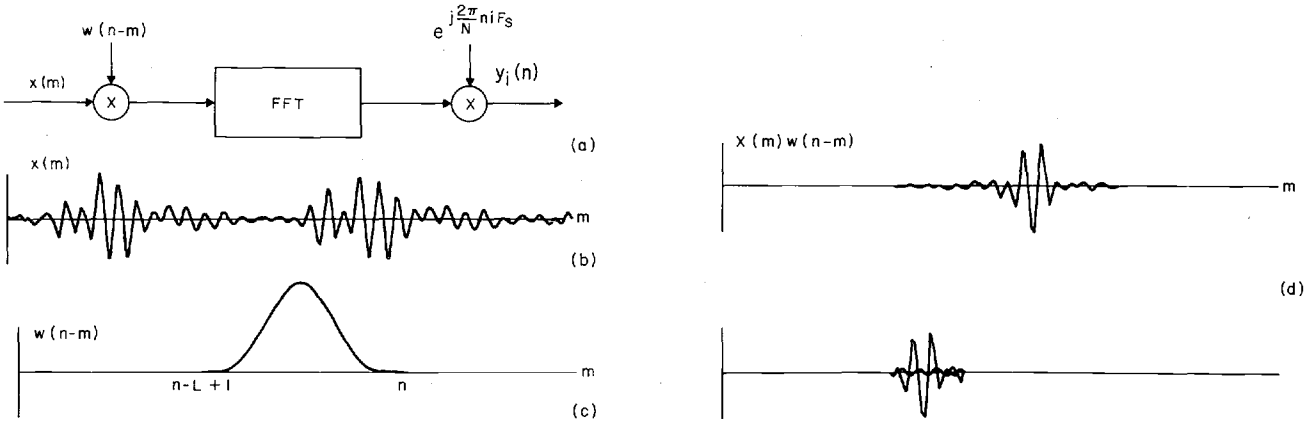

Fig. 7. Summary of processing necessary to obtain N-point windowed sequence to be used to calculate filter bank outputs.

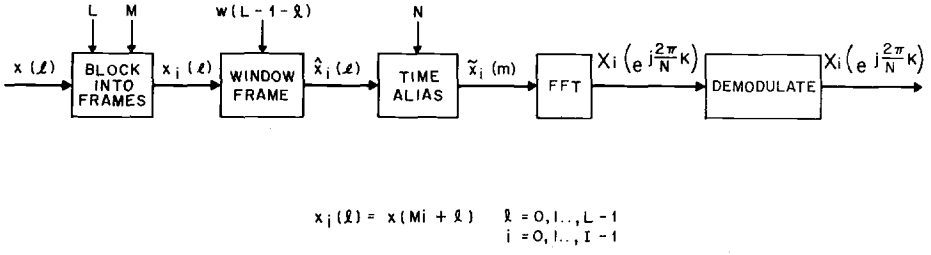

Fig. 8. Block diagram of DFT implementation of a uniform filter bank.

$$\tilde{x}(m) = \sum_{r=-\infty}^{\infty} x(m+rN)\, w(n-m-rN)$$

$$m = 0, 1, \cdots, N-1. \tag{18}$$

This aliasing process is illustrated in Fig. 7. Fig. 7(a) shows $x(m)$, Fig. 7(b) shows the low-pass response $w(n-m)$, and Fig. 7(c) shows the product which is nonzero for $m = n-L+1, \cdots, n$. Fig. 7(d) shows the principle components of (18).

Fig. 8 shows a block diagram of the uniform filter bank feature analysis system. This system is a block processing system in which a frame of L speech samples is processed to yield a single feature vector. This is done by taking advantage of the fact that the window is nonzero for only L time samples. Thus, to obtain a feature vector, the speech signal is blocked into L sample sections (frames) for feature measurement. Consecutive frames are spaced M samples apart. Clearly, the choice of M determines the sampling rate at the output of the filter bank.

If we denote the ith frame of speech as $x_i(l)$, we have

$$x_i(l) = x(Mi+l) \qquad \begin{matrix} l = 0, 1, \cdots, L-1 \\ i = 0, 1, \cdots, I-1 \end{matrix} \tag{19}$$

where $i = 0$ is the first frame and $i = I-1$ is the Ith frame of speech. Next, these frames of speech are multiplied by the time-reversed low-pass window resulting in the windowed signal $\hat{x}_i(l)$

$$\hat{x}_i(l) = x_i(l) \cdot w(L-1-l). \tag{20}$$

The windowed signal is then time-aliased as shown in (18) to obtain the desired N-point signal $\tilde{x}_i(m)$. The next step in the analysis is to calculate the DFT of this signal resulting in the filter bank outputs. These filter bank outputs are then demodulated to give $\tilde{X}_i(e^{j(2\pi/N)k})$ as

$$\tilde{X}_i[e^{j(2\pi/N)k}] = \mathrm{Re}\,\{X_i[e^{j(2\pi/N)k}]\} \cdot \cos(\theta_i) + \mathrm{Im}\,\{X_i[e^{j(2\pi/N)k}]\} \cdot \sin(\theta_i) \tag{21}$$

where

$$\theta_i = 2\pi \cdot \mathrm{mod}_N((i-1) \cdot M \cdot K)/N. \tag{22}$$

From the above discussion it can readily be seen that the design of a uniform filter bank reduces to choosing a value for Q (the number of filters) and designing an appropriate low-pass window.

IV. Design of the Filter Banks

A. Uniform Filter Bank Designs

For the uniform filter banks, the design involves choosing the number of filters and then designing an appropriate window (low-pass filter). In this study we chose to look at four different values of Q, namely, 3, 7, 15, and 31 (filters). These values correspond to the calculation of 8, 16, 32, and 64 point FFT's, in the structure of Fig. 7, respectively. With these choices for Q, M (the frame shift) was chosen to be 10 samples for the first three filter banks and 25 samples for the 31-channel filter bank. This results in sampling rates of 667 Hz (for $M = 10$) and 267 Hz (for $M = 25$) at the output of the filter banks. The lengths of the windows used to implement the filter banks were 51, 51, 101, and 201 samples, respectively.

To design the low-pass window function it was decided that a Kaiser window should be used [23]. This window type has the property that it is the finite duration sequence that has the maximum spectral energy contained in the main lobe. This window was used in two different ways. The first was to use

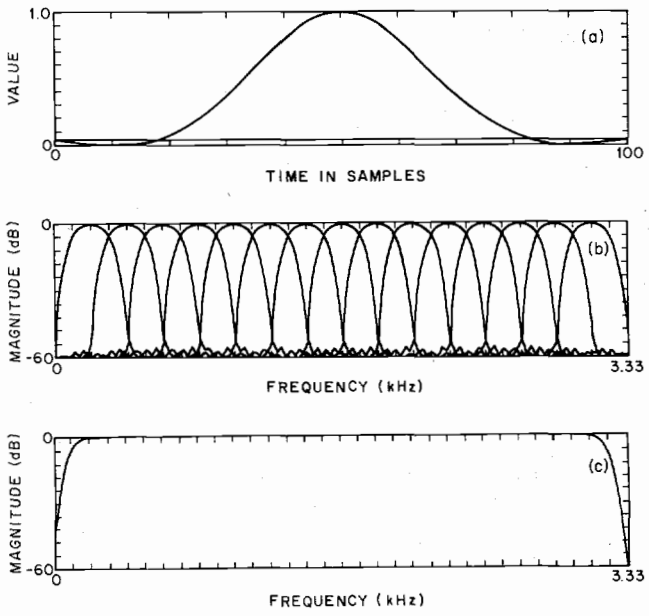

Fig. 9. Example of uniform filter bank using DFT implementation. This example shows the results obtained for a 15-channel filter bank using the window design technique.

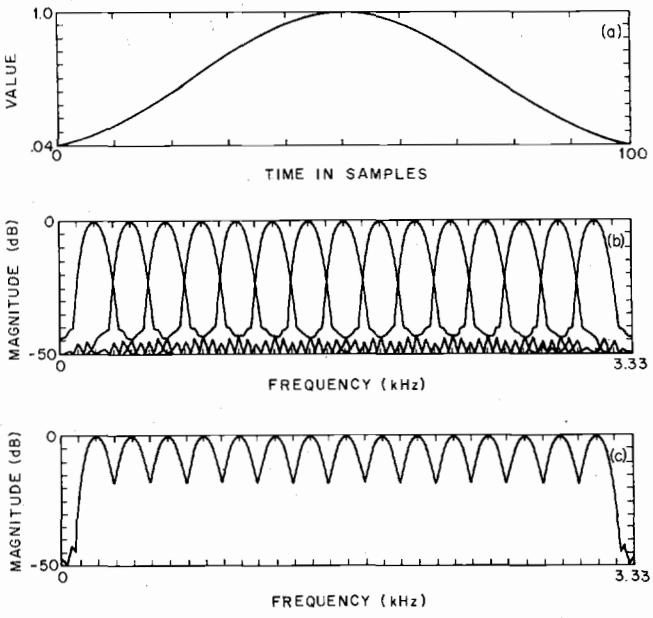

Fig. 10. Example of uniform filter bank using DFT implementation. This example shows the results obtained for a 15-channel filter bank using the Kaiser window directly.

the Kaiser window to design an appropriate low-pass filter by using the well known window design technique [24]. When the Kaiser window is used in this manner it has the desirable property that the composite spectrum of the filter bank is extremely flat [25]. The second was to use the window directly, since the window is essentially a low-pass filter with poor frequency characteristics. When used in this manner the composite spectrum is not flat, but contains valleys between adjacent filters in the filter bank. The premise that a flat composite spectrum is necessary to obtain good recognition accuracy could be tested in this way. The results of the window design for $Q = 15$ are shown in Figs. 9 and 10. In Fig. 9(a) the time response of the low-pass window is plotted. Fig. 9(b) shows the frequency responses (log magnitudes) of the individual filters in the 15-channel filter bank. Fig. 9(c) shows the composite spectrum of the overall 15-channel filter bank. It can be seen that the composite spectrum is essentially flat over the frequency range of the filter bank.

Fig. 10 shows a similar set of plots for the 15-channel filter bank where the Kaiser window is used directly. In this case, the individual filters are narrower in bandwidth and the overall filter bank shows 18 dB gaps at the boundaries between each filter.

In this study a set of eight uniform banks were designed—four using the Kaiser window directly, and four using the Kaiser window to aid in designing a window-based low-pass filter. The specifications of each of the four basic Kaiser windows are given in Table I which shows values of Q (number of filters), L (impulse response duration), M (frame shift), β (normalized bandwidth of the Kaiser window), and α (peak sidelobe attenuation of the resulting low-pass filter). Values of α from 53 to 60 dB were attained for the four designs.

B. Nonuniform Filter Banks

For the nonuniform filter banks we chose to investigate three different filter bank spacings; octave spacing, critical bands, and $\frac{1}{3}$ octave spacing. In particular we considered 4-, 7-, and 12-channel filter banks for the octave, critical band, and $\frac{1}{3}$ octave filter banks. The ideal filter characteristics of a 4-channel octave-band filter bank were shown in Fig. 4(a) when the band of interest was from 200 to 3200 Hz. Similarly, the ideal filter characteristics of a 12-channel $\frac{1}{3}$ octave filter bank were shown in Fig. 4(b). Plots of the actual filter characteristics for both the octave and $\frac{1}{3}$ octave filter banks are given in Figs. 11 and 12. These figures show plots of the log magnitude responses of each of the channels and also the composite frequency response of the filter bank. The ideal filter characteristics of the filters in the critical band filter bank are based on the articulation index [22] and were shown in Fig. 5. The filters in this filter bank were spaced to incorporate two critical bands in each filter. A plot of the actual filter characteristics and the composite spectrum for the critical band filter bank is given in Fig. 13.

In addition to the above set of uniform and nonuniform filter banks, two specially designed nonuniform filter banks were studied. The first was a 5-channel filter bank designed for use in the IBM speech terminal by Silverman and Dixon [26]. For use in a recognition system based on telephone quality speech, the cutoff frequencies of the lowest and highest frequency bands were suitably changed to 200 and 3200 Hz, respectively. As a result of the changes made to these low and high frequency cutoffs, the performance of this modified IBM filter bank need not reflect the performance of the original 5-channel filter bank as designed by Silverman and Dixon. Fig. 14 shows plots of the log magnitude responses of the five channels of this filter bank, and the composite frequency response. The composite frequency response is seen to be essentially flat from 200 to 3200 Hz, and each individual channel provides about 70 dB of out-of-band signal rejection.

The second specially designed filter bank was based on the system by Martin [27]. The filters used were spaced along critical bands; however, the frequency selectivity of these fil-

TABLE I
UNIFORM FILTER BANK DESIGN PARAMETERS

Q	L	M	β	α (dB)
3	51	10	5.65	60.00
7	51	10	4.961	53.72
15	101	10	4.864	52.84
31	201	25	4.864	52.84

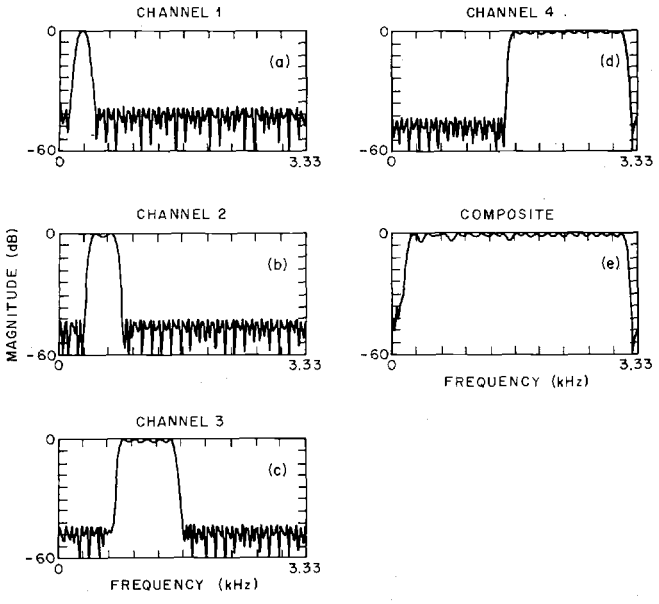

Fig. 11. Results of octave filter bank design. The response of the individual channels and the composite are shown.

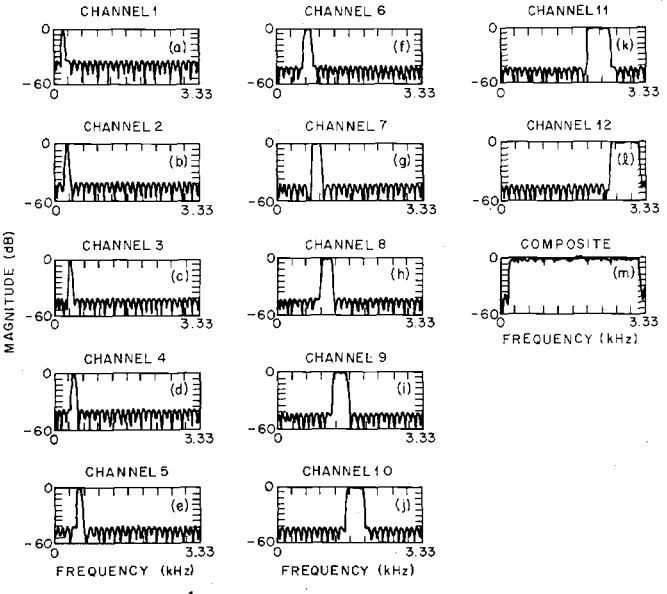

Fig. 12. Results of $\frac{1}{3}$-octave filter bank design. The responses of the individual channels and the composite are shown.

ters was very poor (the ratios of center frequency to bandwidth were about 8). This poor frequency selectivity was chosen to provide good time resolution. Plots of the individual channel frequency responses as well as the composite frequency response are given in Fig. 15. A slight nonflat overall frequency response is seen in this figure. The filters in this

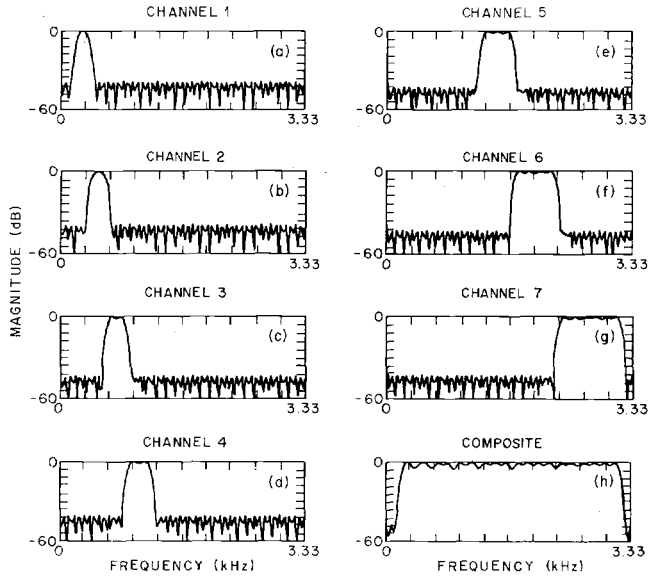

Fig. 13. Results of critical-band spaced filter bank design. The responses of the individual channels and the composite are shown.

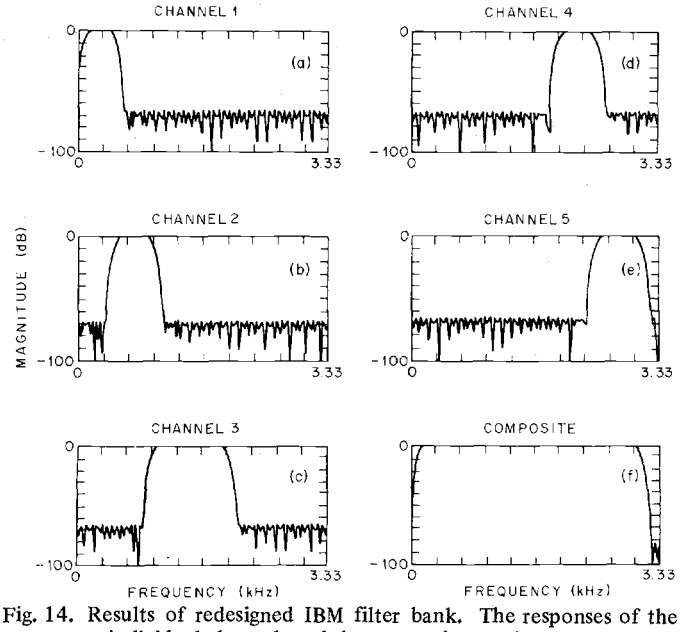

Fig. 14. Results of redesigned IBM filter bank. The responses of the individual channels and the composite are shown.

filter bank are highly overlapping in contrast to all previous cases where there was little or no filter overlap.

Table II shows the actual bandpass filter cutoff frequencies, the resulting passband and stopband ripples, and the actual FIR filter lengths for the 4-channel octave band design, the 5-channel IBM design, the 7-channel critical band design, and the 12-channel $\frac{1}{3}$ octave design. Table III gives the filter center frequencies (f_c) and filter durations for the 13-channel filter bank.

C. Summary of Filter Banks

A total of 13 filter banks were implemented and studied. Eight of the filter banks were uniformly distributed in frequency with as few as 3 and as many as 31 channels covering

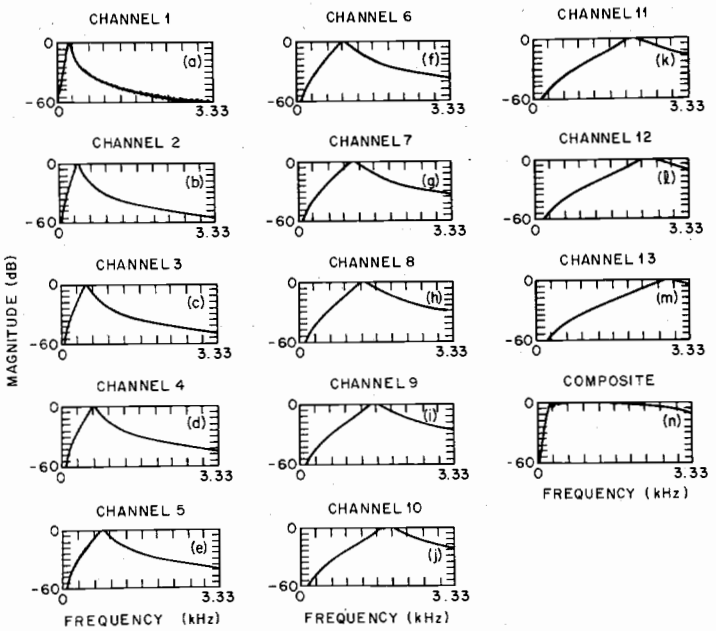

Fig. 15. Result of digital simulation of the Martin filter bank. The responses of the individual channels and the composite are shown.

TABLE II
DESIGN PARAMETERS FOR NONUNIFORM FILTER BANKS
(Q = 4, 5, 7, 12), F_s = 6.67 kHz

Channel No.	Lower Stopband Frequency (Hz)	Lower Passband Frequency (Hz)	Upper Passband Frequency (Hz)	Upper Stopband Frequency (Hz)	δ_p	δ_s
1	150	250	350	450	.133	.0133
2	350	450	750	850	.0810	.0081
3	750	850	1550	1650	.0743	.0074
4	1550	1650	3150	3250	.0722	.0072
(a) Bandedge frequencies for Q = 4 filter bank with L = 101						
1	0	154	451	605	.0039	.0011
2	381	586	917	1122	.0015	.0004
3	881	1086	1915	2120	.0019	.0005
4	1279	2084	2416	2621	.0016	.0005
5	2279	2584	3122	3333	.0020	.0005
(b) Bandedge frequencies for Q = 5 filter bank with L = 101						
1	150	250	350	450	.1334	.0133
2	350	450	580	680	.1253	.0125
3	580	680	870	970	.0737	.0074
4	870	970	1220	1320	.1009	.0101
5	1220	1320	1670	1770	.0790	.0079
6	1670	1770	2270	2370	.0788	.0079
7	2270	2370	3150	3250	.0803	.0080
(c) Bandedge frequencies for Q = 7 filter bank with L = 101						
1	175	225	225	275	.1744	.0174
2	225	275	290	390	.0982	.0098
3	290	340	375	425	.0918	.0092
4	375	425	475	525	.1243	.0124
5	475	525	605	655	.0876	.0088
6	605	655	775	825	.0980	.0098
7	775	825	975	1025	.0792	.0079
8	975	1025	1235	1285	.0827	.0083
9	1235	1285	1575	1625	.0735	.0074
10	1575	1625	1975	2025	.0745	.0074
11	1975	2025	2495	2545	.0751	.0075
12	2495	2545	3175	3225	.0660	.0066
(d) Bandedge frequencies for Q = 12 filter bank with L = 201						

the baseband of interest (200-3200 Hz). The remaining five filter banks were implementations of octave band, $\frac{1}{3}$ octave band, critical band, and two specially-designed filter banks. In the next section we describe the experiments used to measure the performance of each of the filter banks in a speaker-trained isolated word recognition test.

TABLE III
DESIGN PARAMETERS FOR 13-CHANNEL NONUNIFORM FILTER BANK
(L = 201)

Channel No.	Center Frequency (Hz)
1	260
2	395
3	535
4	683
5	841
6	1011
7	1198
8	1405
9	1635
10	1892
11	2179
12	2505
13	2885

V. DESCRIPTION OF EXPERIMENTS AND RESULTS

In order to evaluate the effects of filter bank parameters on speaker-trained isolated word recognition accuracy, a 39 word vocabulary which consisted of the alphabet, the digits, and three command words (STOP, ERROR, and REPEAT) was chosen. This vocabulary was selected for its high degree of complexity and moderate size [28]. The measured recognition accuracy for this vocabulary has been shown to be relatively low in previous tests [10], [29]. Thus, small differences in system performance can often be reliably measured with a reasonable size set for this vocabulary.

To evaluate the recognition performance of the filter banks, a set of reference patterns was collected for several talkers over a several week period. These reference patterns consisted of a set of 39 robust tokens, one for each of the words in the vocabulary [30]. This was done for each of four talkers (two male, two female) for all thirteen filter banks. Each of the four talkers had participated in a wide variety of offline tests of isolated word recognition systems, i.e., with no feedback as to how the recognizer performed on their spoken inputs. Next, an independent test set, consisting of ten recordings of the 39 word vocabulary spoken by each of the four talkers, was recorded several weeks later. In this manner a total of 390 isolated-word inputs for each of the four speakers was obtained.

Each of the isolated words (for both the reference and test recordings) was obtained by flashing the word on a video monitor and asking the talker to speak the word after hearing an appropriate starting cue (a beep). An automatic word endpoint detector was used to locate word boundaries [31]. About 2 percent of the words (all of these occurred for one of the four talkers) had endpoint errors which were corrected manually.

A modified form of the robust training procedure [30] was used to provide a single reference template from the set of reference tokens for each talker. The modification consisted of using filter bank channel signals in place of the LPC vectors of the standard robust training algorithm. The philosophy of the robust training procedure was preserved in the modified algorithm.

For each test set three experiments were performed. The first consisted of testing each filter bank with the entire test set for each of the four talkers. The measure of performance for this experiment was the error rate as a function of the

candidate position C. This error measure is defined to be the percentage of correct words which are not in the top C word choices. For most applications the results for $C = 1$ (top candidate) provide the best measure of recognizer performance. For some tasks, however, it is reasonable to compare error rates for $C > 1$ since task syntax can be used to detect and correct errors in the top C recognition choices [14]. Using standard statistical tables, it can be shown that at the 99 percent confidence level, a difference of about 1.5-2 percent in error rate is statistically significant.

The second experiment consisted of measuring the performance of a standard LPC system [7]-[10] on each test set used in the first experiment. The performance measure for this experiment was again the error rate as a function of the candidate position. The purpose of this experiment was to determine the relative performance of the LPC and filter bank systems.

The third experiment evaluated the performance of both LPC and filter bank systems with a subset of the original test set. This subset consisted of only the digits vocabulary. This vocabulary was chosen because of its low complexity and small size, and because the digits are widely used in many applications. In this way, differences in performance on a simple recognition task could be measured. This experiment was carried out by using the 100 digits of the 390 isolated word inputs in the original test set. The measure of performance for this experiment was the error rate of the first candidate only.

A. Results for Experiment 1

The results of experiment 1 are given as a series of plots of average word error rate for each talker as a function of either best candidate position or the number of channels in the filter bank.

The results for the uniform filter banks are given in Figs. 16 and 17. Fig. 16 shows plots of average word error rate versus candidate position for each of the filter banks for each talker. Fig. 16(a) and (b) corresponds to male talkers while Fig. 16(c) and (d) corresponds to female talkers. The solid curves are for filter banks using window designed low-pass filters and the dashed curves are for filter banks using the window itself as the low-pass filter. The reader should note the different scales on the ordinates for different talkers.

Two general trends emerge from the curves of Fig. 16. First we see that the word error rates differ greatly among talkers, i.e., the first talker had about an 8.5 percent word error rate ($C = 1$, $Q = 15$, window design), whereas the fourth talker had an 18.5 percent word error rate with the same conditions. This variation in error scores is typical for the alphadigits vocabulary [25], and the scores of the four talkers fall within the normally expected range.

The second trend seen in the curves of Fig. 16 is that the uniform filter banks with a flat composite spectrum (solid curves) generally do better than those having the same number of channels with a composite spectrum which contains valleys (dashed curves). This result is independent of the number of channels of the filter bank with only one exception. This exception is for the first talker and the 3-channel uniform

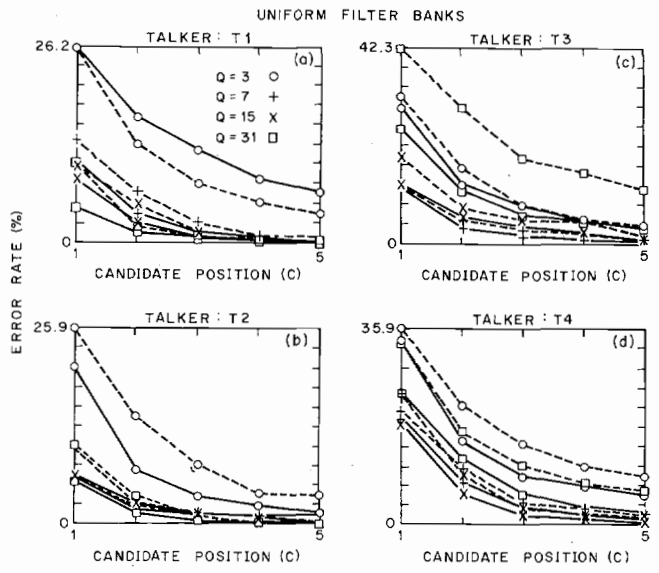

Fig. 16. Plots of word error rates versus candidate position for the uniform filter banks for each of the four talkers.

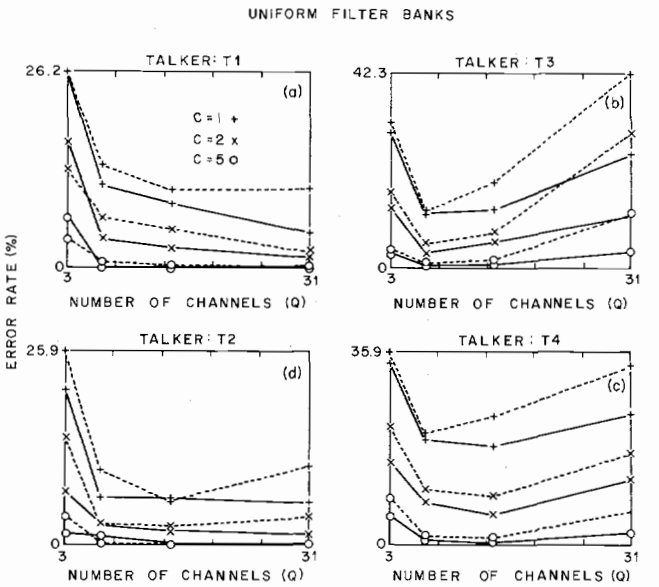

Fig. 17. Plots of word error rates versus number of channels for the uniform filter banks for each of the four talkers.

filter bank. Because of the poor performance of the 3-channel uniform filter bank this deviation has little significance.

Fig. 17 shows plots of the average word error rate versus Q, the number of channels in the filter bank, for three values of C ($C = 1, 2, 5$), and for each talker. Here we see two different trends depending upon the sex of the talkers. For the males it can be seen that as Q increases to 31 the word error rate has a tendency to steadily decrease. Conversely, for females as Q increases beyond 15 the error rate has the tendency to increase substantially. This effect occurs because as the number of channels increases, the bandwidth of the filters decreases, and at some point the bandwidth becomes small enough to that there is a high probability that there will be no energy from the speech signal present in a particular band. When this occurs the pattern matching algorithm effectively attempts to

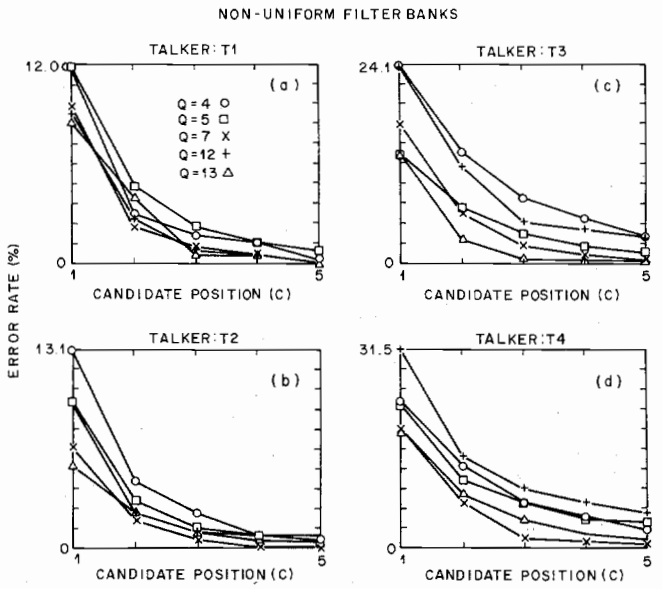

Fig. 18. Plots of word error rates versus candidate position for the nonuniform filter banks for each of the four talkers.

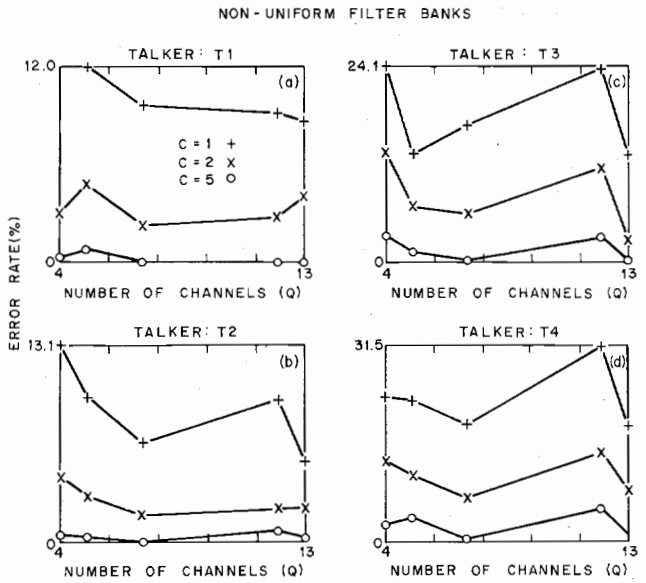

Fig. 19. Plots of word error rates versus number of channels for the nonuniform filter banks for each of the four talkers.

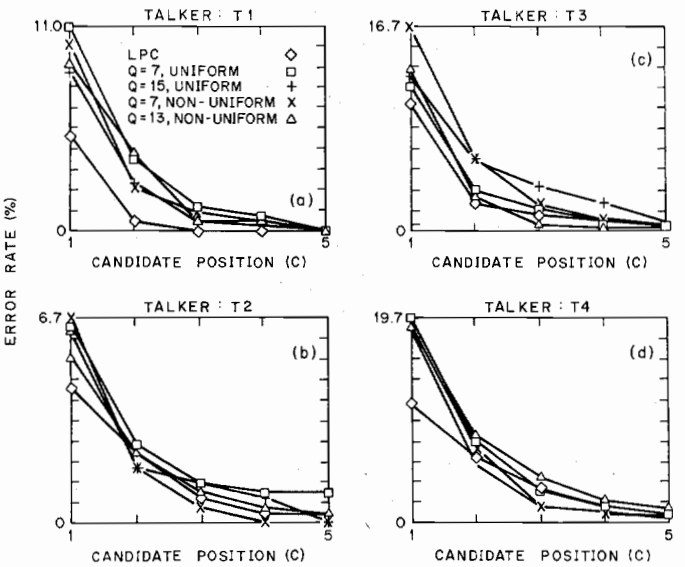

Fig. 20. Plots of word error rates versus candidate position for standard LPC system and selected filter bank systems for each of the four talkers.

TABLE IV
WORD ERROR RATE (PERCENT) AS A FUNCTION OF CANDIDATE POSITION C FOR LPC AND SELECTED FILTER BANKS

Recognition System	Talker	C = 1	2	3	4	5
LPC	1	5.1	0.5	0.0	0.0	0.0
	2	4.1	2.3	0.8	0.3	0.3
	3	10.3	2.3	1.3	1.0	0.5
	4	11.8	6.7	3.3	1.3	0.8
	MEAN	7.8	3.0	1.4	0.7	0.4
7-channel uniform filter bank	1	11.0	3.8	1.3	0.8	0.0
	2	6.4	2.6	1.3	1.0	1.0
	3	11.8	3.3	1.8	0.8	0.5
	4	19.7	7.9	3.1	1.5	0.8
	MEAN	12.2	4.4	1.9	1.0	0.6
15-channel uniform filter bank	1	8.5	2.6	0.5	0.3	0.0
	2	6.2	1.8	1.3	0.8	0.0
	3	12.6	5.6	3.6	2.3	0.8
	4	18.5	5.6	1.5	1.0	0.3
	MEAN	11.5	3.9	1.7	1.1	0.3
7-channel non-uniform filter bank	1	10.0	2.3	1.0	0.5	0.0
	2	6.7	1.8	0.5	0.0	0.0
	3	16.7	5.9	2.1	1.0	0.3
	4	19.0	7.2	1.5	0.8	0.5
	MEAN	13.1	4.3	1.3	0.6	0.2
13-channel non-uniform filter bank	1	9.0	4.1	0.5	0.5	0.0
	2	5.4	2.3	1.0	0.5	0.3
	3	13.1	2.8	0.5	0.3	0.3
	4	18.7	8.5	4.4	2.1	1.3
	MEAN	11.6	4.4	1.6	0.9	0.5

match random background level energies. For such cases substantial variability in channel distances leads to increased word error rate. This effect occurs predominantly for female talkers because for voiced sounds only a few harmonics are present in the speech due to the high female pitch. Therefore, the likelihood of a fixed bandwidth channel measuring only background noise is higher for females than for males.

The results for the nonuniform filter banks are given in Figs. 18 and 19. Fig. 18 shows plots of average word error rates versus candidate position for each of the filter banks for each talker. As was the case with the uniform filter banks, we see that the word error rate curves differ greatly among talkers. In Fig. 19 it is shown that the 13-channel filter bank does not follow the trend (for the females) of increasing error rate as the number of filters increases. This is due to the fact that the 13-channel filter bank consisted of poor frequency resolution, good time resolution filters. Because of this, the probability of a single filter measuring only background noise for high-pitched female talkers is greatly reduced.

B. Results of Experiment 2

The results of the recognition experiment using a standard LPC system are given in Fig. 20 which shows comparisons between average word error rates versus candidate position for both the LPC system and the four best filter systems.[2] As with the filter bank systems, a great deal of variation of perfor-

[2]These were the 7- and 15-channel uniform filter banks, and the 7- and 13-channel nonuniform filter banks.

TABLE V
WORD ERROR RATE (PERCENT) FOR DIGITS VOCABULARY

	Talker Number				
Q (Uniform)	1	2	3	4	MEAN
3	7.0	5.0	1.0	7.0	5.0
7	0.0	0.0	1.0	0.0	0.3
15	0.0	0.0	1.0	0.0	0.3
31	0.0	0.0	3.0	3.0	1.5
Q (Non-Uniform)					
4	0.0	0.0	1.0	0.0	0.3
5	1.0	0.0	1.0	1.0	0.8
7	0.0	0.0	1.0	0.0	0.3
12	0.0	0.0	6.0	8.0	3.8
13	0.0	0.0	1.0	0.0	0.3
LPC	0.0	0.0	1.0	1.0	0.5

mance is observed between talkers. The results for the standard LPC system are given in Table IV along with those of the four best filter banks. The data in Table IV show that the LPC system has, on average, a 4 percent lower error rate for the first candidate position than the best of the filter bank recognizers. It can also be observed that as C increases to 5, the performance of both types of systems are equal, to within statistical variations.

C. Results of Experiment 3

The results of the third experiment, in which the vocabulary was limited to include only the digits, are given in Table V which shows the number of errors made for each talker on this vocabulary for several filter banks and for the LPC recognizer. The data given in this table show that, for the digits vocabulary, the performance of the filter bank systems is nearly identical to that of the LPC system.

VI. DISCUSSIONS

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

1) Filter bank recognizer performance degrades for filter banks with too few filters (Q in the range of 3) or too many filters (Q in the range of 31) for nonoverlapping filter banks. The reasons for this degradation in performance are that for small values of Q the system is giving very poor frequency resolution leading to an inability to discriminate between words, and for large values of Q the individual filters become so narrow in bandwidth that they are often measuring noise rather than speech. This effect is especially pronounced for female talkers (with high pitch) since the speech harmonics are widely spaced; and for large values of Q (e.g., $Q = 31$) a number of the bands are usually measuring only background noise.

2) For all filter banks (both uniform and nonuniform) the composite spectrum should be essentially without sharp valleys (i.e., flat or slowly changing as from a mild preemphasis) so as to retain all the information about the speech spectrum in the analysis.

3) For nonuniform filter banks, the recognizer performance obtained when the filters were spaced along a critical band frequency scale was significantly better than when the filters were spaced along octave bands, $\frac{1}{3}$ octave bands, or arbitrary spacings. The critical band scale is essentially a linear frequency scale in the range 100–1500 Hz and becomes highly nonlinear above this frequency range. Hence, the critical band scale can be considered a modified uniform scale so this result indicates that a uniform frequency spacing up to 1500 Hz is desirable for filter bank systems.

4) The performance of 7-band and 13-band critical band filter banks was statistically the same as for 7-band and 15-band uniform filter banks. Again this result reflects the similarities between both types of filter banks in the important frequency range from 100 to 1500 Hz.

5) The performance of the LPC-based word recognizer was statistically better than that of any of the filter bank recognizers (for the conditions studied) for the 39 word alphadigits vocabulary. In particular, the average error rate for the LPC recognizer was about 4 percent lower than that of the *best* filter bank recognizer. (The reader will recall that a 1.5–2 percent improvement in performance is statistically significant at the 99 percent confidence level). For the digits vocabulary, the performances of both the LPC and the best filter bank recognizers were comparable with error rates close to 0 percent.

A key question raised by the above results is why does the LPC-based recognizer perform better than any of the filter bank recognizers? A related question is why did White and Neely, in a classical comparison of LPC and filter bank systems [13], find comparable performances for both systems using the same alphadigits vocabulary? The answers to these questions are fundamental and are related to the basic ideas behind parametric models of LPC and filter banks. A Q-band filter bank analysis, as seen in Fig. 1, is a *fixed* quantization of the frequency axis into Q regions. The designer of the filter bank has freedom to choose the value of Q and the way in which the frequency scale is subdivided; however, once chosen, this fixed frequency quantization is applied to all talkers, words, etc. Problems arise when relevant information in a pattern (e.g., for a word) lies at the edge of a band (and hence is present in one band during training, and a different band during testing) because large differences in band energy between test and reference are often obtained. Other problems arise when, during the course of a word, little or no energy of the speech signal occurs in a given band. In such cases the measured level of the pattern (either reference or test) is highly variable, and again leads to large random distance components in comparing test and reference patterns.

LPC analysis, on the other hand, is an *adaptive* frequency quantization analysis procedure in that the Q poles of approximation (for a Qth-order analysis) distribute themselves to occur where the speech spectrum has the most energy. Hence, differences in frequency locations of speech energy between reference and test are reflected as movements of the speech poles (resonances) and generally lead to distances proportional to the magnitude of the difference in frequency location. Thus, the distance function for LPC recognizers is far better behaved than the distance function for filter bank recognizers when differences occur between test and reference patterns.

One simple way of illustrating the relative sensitivity of filter bank and LPC-based recognizers is to show plots of the differential distance histograms of both systems for correct recognitions and for errors. The differential distance is defined as the

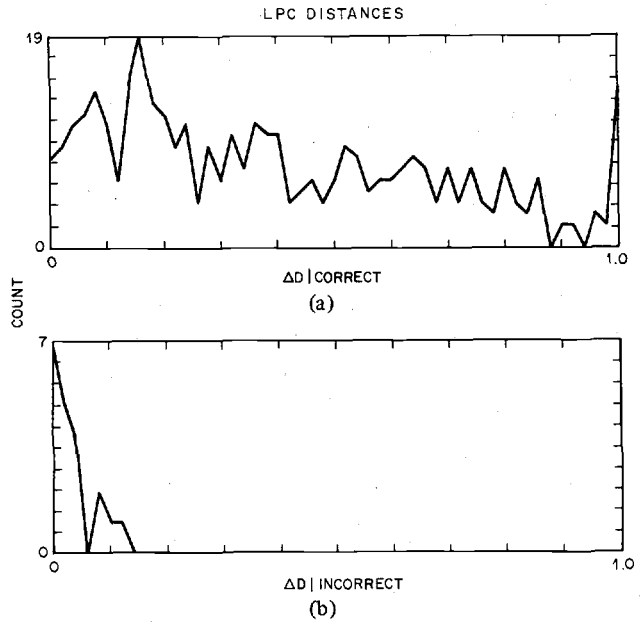

Fig. 21. Histograms of the differential distance for standard LPC system for talker *T*1. Histogram (a) shows differential distance given correct recognition, while (b) shows a similar histogram given incorrect recognition.

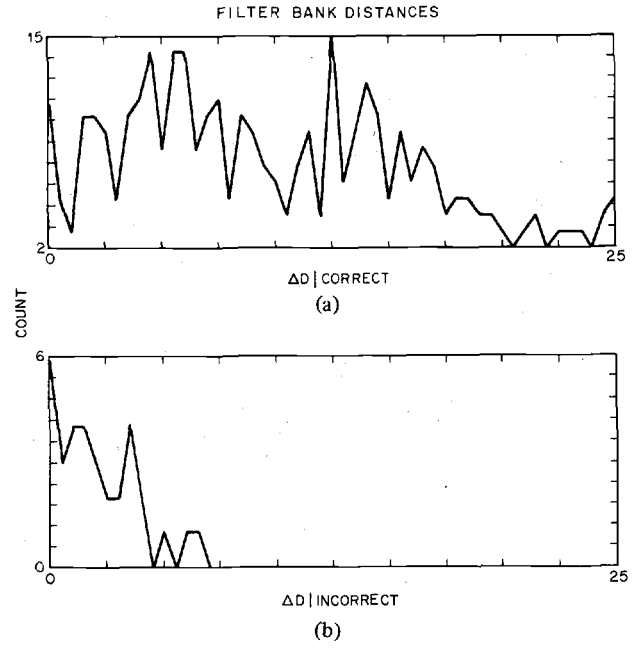

Fig. 22. Histograms of the differential distance for 15-channel uniform filter bank system for talker *T*1. Histogram (a) shows differential distance given correct recognition, while (b) shows a similar histogram given incorrect recognition.

magnitude of the *difference* in distance between the best reference which is not the spoken word, and the reference which represents the spoken word, i.e.,

$$\Delta D = \left| d(T, R_{i*}) - d(T, R_{j*}) \right|$$

where $d(T, R_{i*})$ is the distance between the test and the reference corresponding to the test word, and $d(T, R_{j*})$ is the distance between the test and the best reference which does not correspond to the test word. Figs. 21 and 22 show histogram plots of ΔD conditioned in correct recognition [Fig. 21(a)] and incorrect recognition [Fig. 21(b)] for the LPC recognizer and the 15-channel uniform filter bank (Fig. 22) for one of the four talkers. Table VI gives statistics on average distance and average separations for the best filter banks and for the LPC system. For the LPC system the differential distance tends to be large for correct recognition and small for incorrect recognition, indicating high confidence in the recognition decision when correct and low confidence (high uncertainty) when incorrect. For the filter bank system the differential distance (suitably normalized) is somewhat smaller for correct recognition than for the LPC system, and somewhat larger for incorrect recognition. This result indicates poorer confidence in the recognition system, when correct, and lower uncertainty in the decision when incorrect.

TABLE VI
SEPARATION STATISTICS OF SEVERAL OF THE WORD RECOGNITION SYSTEMS

System	Talker	Average Distance	Average Differential Distance For Correct Recognition	Average Differential Distance For Incorrect Recognition	Normalized Average Differential Distance For Correct Recognition	Normalized Average Differential Distance For Incorrect Recognition
7-channel uniform filter bank	1	14.80	5.530	1.101	2.595	.517
	2	14.78	6.734	1.634	3.347	.812
	3	16.68	5.657	1.258	2.684	.597
	4	17.24	5.125	1.729	2.040	.688
15-channel uniform filter bank	1	44.77	12.95	3.046	2.838	.667
	2	45.44	14.73	2.693	3.036	.555
	3	48.75	11.08	3.466	2.567	.803
	4	52.59	10.84	3.299	1.891	.567
7-channel non-uniform filter bank	1	13.72	72.79	1.166	2.897	.464
	2	13.64	8.171	1.205	3.602	.531
	3	16.73	6.265	1.856	2.560	.758
	4	17.61	5.743	1.730	1.969	.593
13-channel non-uniform	1	22.63	11.91	1.764	3.285	.486
	2	22.48	13.31	3.798	3.934	1.122
	3	27.10	10.18	2.478	2.562	.623
	4	28.90	10.66	4.401	2.098	.866
LPC	1	.1832	.2098	.0233	3.592	.398
	2	.1755	.2602	.0285	4.163	.601
	3	.2112	.2126	.0405	3.624	.691
	4	.2550	.1742	.0762	2.101	.919

The question then remains as to why White and Neely found comparable performance for LPC and filter bank recognizers, and why most commercial systems use filter banks. The answers to this question are related to the differences in implementation and application of the recognizers. In the White and Neely study (as well as for most commerical systems) the input speech is wide-band and includes frequencies up to about 7 kHz. (White and Neely sampled the speech for the filter bank system at a 20 kHz rate.) The channels from 3–7 kHz are generally broad bandwidth, high time resolution channels that provide accurate and reliable information about fricative sounds and speech transients (e.g., bursts) and these channels provide the margin of improvement over LPC which makes the overall performance of both systems comparable. Furthermore, commerical systems generally only use very simple, highly structured word vocabularies (e.g., the digits) for which differences in performance between LPC and filter bank systems are negligible.

Another reason for preferring filter bank recognizers over LPC-based recognizers is their robustness to channel noise and

other forms of channel distortion. It is well known [32] that LPC analysis systems tend to become error prone with high background noise levels and other transmission distortions, whereas filter bank systems seem to be far less sensitive to noise.

Cost is also an important consideration in choosing between LPC and filter bank analyzers. Filter bank analyzers are lower in cost than LPC-based systems, although the future availability of improved integrated circuit signal processors may soon minimize this cost difference.

VII. Summary

Performance of a wide variety of designs of filter bank word recognizers has been measured for a standard vocabulary of alphadigit terms. Results indicate that the highest word accuracies are obtained with either 15 filters spaced uniformly in frequency or 13 filters spaced along a critical band frequency scale. In general, better performance was obtained for male talkers than for female talkers because of the known interactions between filter bandwidths and pitch harmonic spacings. In comparison to a standard LPC-based word recognizer, the performance of the best filter bank system was significantly poorer than the LPC system for the alphadigits vocabulary. When the vocabulary complexity was reduced to that of a digits-only vocabulary, both systems performed equally well. A discussion of the strengths and weaknesses of the filter bank processing model for isolated word recognition was given.

References

[1] T. B. Martin, "Practical applications of voice input to machines," *Proc. IEEE*, vol. 64, pp. 487–501, Apr. 1976.

[2] S. Moshier, "Talker independent speech recognition in commercial environments," in *Speech Commun. Papers, 97th ASA Meet.*, June 1979, pp. 551–553.

[3] H. Sakoe, "Two-level DP-matching–A dynamic programming based pattern matching for connected word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 588–595, Dec. 1979.

[4] Interstate Electronics Corp., Voice Data Entry System, unpublished tech. descriptions.

[5] Centigram Corp., Mike, unpublished tech. descriptions.

[6] Heuristics Corp., Speechlab, unpublished tech. descriptions.

[7] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 67–72, Feb. 1975.

[8] A. E. Rosenberg and F. Itakura, "Evaluation of an automatic word recognition system over dialed-up telephone lines," *J. Acoust. Soc. Amer.*, vol. 60, suppl. 1, p. S12, 1976.

[9] L. R. Rabiner, "On creating reference templates for speaker independent recognition of isolated words," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 34–42, Feb. 1978.

[10] L. R. Rabiner, S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon, "Speaker-independent recognition of isolated words using clustering techniques," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 336–349, Aug. 1979.

[11] Reticon, literature on filter bank integrated circuit.

[12] J. B. Allen, "Cochlear micro mechanics–A physical model of transduction," *J. Acoust. Soc. Amer.*, vol. 68, pp. 1660–1670, Dec. 1980.

[13] G. M. White and R. B. Neely, "Speech recognition experiments with linear prediction, bandpass filtering, and dynamic programming," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 183–188, Apr. 1976.

[14] S. Seneff, D. Klatt, and V. Zue, "Design considerations for optimizing the intelligibility of DFT-based, pitched-excited, critical-band spectrum speech analysis/resynthesis system," *J. Acoust. Soc. Amer.*, vol. 69, suppl. 1, p. S17, May 1981.

[15] N. R. Dixon and H. F. Silverman, "What are the significant variables in dynamic programming for discrete utterance recognition," in *Proc. ICASSP '81*, Mar. 1981, pp. 728–731.

[16] L. R. Rabiner and S. E. Levinson, "Isolated and connected word recognition–Theory and selected applications," *IEEE Trans. Commun.*, vol. COM-29, pp. 621–659, May 1981.

[17] J. D. Markel and A. H. Gray Jr., "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 380–391, Oct. 1976.

[18] H. F. Silverman and N. R. Dixon, "A comparison of several speech-spectra classification methods," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 289–295, Aug. 1976.

[19] D. H. Klatt, "Prediction of perceived phonetic distance for critical-band spectra: A first step," in *Proc. ICASSP'82*, May 1982, pp. 1278–1281.

[20] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[21] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*, 2nd ed. New York: Springer-Verlag, 1972.

[22] E. Zwicker, "Subdivision of the audible frequency range into critical bands, (Frequenzgruppen)," *J. Acoust. Soc. Amer.*, vol. 23, p. 248, 1961.

[23] J. F. Kaiser, "Nonrecursive digital filter design using the I_0-sinh window function," *Proc. IEEE*, vol. 65, pp. 1558–1564, Nov. 1977.

[24] L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

[25] R. W. Schafer, L. R. Rabiner, and O. Herrmann, "FIR digital-filter banks for speech analysis," *Bell Syst. Tech. J.*, vol. 54, pp. 531–544, Mar. 1975.

[26] H. F. Silverman and N. R. Dixon, "State constrained dynamic programming (SCDP) for discrete utterance recognition," in *Proc. ICASSP 80*, vol. 1, pp. 169–172.

[27] T. B. Martin, "Acoustic recognition of a limited vocabulary in continuous speech," Ph.D. dissertation, Univ. Pennsylvania, Philadelphia, 1970; Ann Arbor, MI: University Microfilms Ltd., 1970.

[28] L. R. Rabiner, A. E. Rosenberg, J. G. Wilpon, and W. J. Keilin, "Isolated word recognition for large vocabularies," *Bell Syst. Tech. J.*, vol. 61, pp. 2989–3005, Dec. 1982.

[29] L. R. Rabiner, J. G. Wilpon, and J. G. Ackenhusen, "On the effects of varying analysis parameters on LPC based isolated word recognizer," *Bell Syst. Tech. J.*, vol. 60, pp. 893–911, July–Aug. 1981.

[30] L. R. Rabiner and J. G. Wilpon, "A simplified robust training procedure for speaker trained, isolated word recognition systems," *J. Acoust. Soc. Amer.*, vol. 68, pp. 1271–1276, Nov. 1980.

[31] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 777–785, Aug. 1981.

[32] J. S. Lim, "Estimation of LPC coefficients from speech waveforms degraded by additive random noise," in *Proc. ICASSP 78*, pp. 599–601.

Bruce A. Dautrich received the B.A. degree in electrical engineering from Pennsylvania State University, University Park, in 1980, and the M.S. degree in electrical engineering from Purdue University, West Lafayette, IN, in 1981.

Since 1981 he has been with Bell Laboratories, Murray Hill, NJ, where he has engaged in speech communication research and development and is presently concentrating on problems of speech recognition.

Mr. Dautrich is a member of Eta Kappa Nu, Tau Beta Pi, and Phi Kappa Phi.

Lawrence R. Rabiner (S'62-M'67-SM'75-F'75), for a photograph and biography, see p. 726 of the June 1983 issue of the TRANSACTIONS.

Thomas B. Martin was born in Burlington, IA, on June 23, 1935. He received the B.S.E.E. degree from the University of Notre Dame, South Bend, IN, in June 1957 and the M.S.E.E. and Ph.D. degrees from the University of Pennsylvania, Philadelphia, in 1960 and 1970, respectively.

In 1970 he cofounded Threshold Technology Inc., which developed the first practical automatic speech recognition systems to be used in industry. In March 1982, he joined Bell Laboratories, Murray Hill, NJ, and has continued to work in speech recognition.

On Temporal Alignment of Sentences of Natural and Synthetic Speech

HANS D. HÖHNE, CECIL COKER, STEPHEN E. LEVINSON, SENIOR MEMBER, IEEE, AND LAWRENCE R. RABINER, FELLOW, IEEE

Abstract—One way to improve the quality of synthetic speech, and to learn about temporal aspects of speech recognition, is to study the problem of time aligning pairs of spoken sentences. For example, one could evaluate various sets of duration rules for synthesis by comparing the time alignments of speech sounds within synthetic sentences to those of naturally spoken sentences. In this manner, an improved set of sound duration rules could be obtained by applying some objective measure to the alignment scores. For speech recognition applications, one could obtain automatic labeling of continuous speech from a hand-marked prototype to obtain models and/or statistical data on sounds within sentences. A key question in the use of automatic alignment of sentence length utterances is whether the time warping methods, developed for isolated word recognition, could be extended to the problem of time aligning sentence length utterances (up to several seconds long). A second key question is the reliability and accuracy of such an alignment. In this paper we investigate these questions.

It is shown that, with some simple modifications, the dynamic time warping procedures used for isolated word recognition apply almost as well to alignment of sentence length utterances. It is also shown that, on the average, the uncertainty in the location of significant events within the sentence is much smaller than the event durations although the largest errors are longer than some event durations. Hence, one must apply caution in using the time alignment contour for synthesis or recognition applications.

Manuscript received May 5, 1982; revised November 24, 1982.

H. D. Höhne was on leave at Bell Laboratories, Murray Hill, NJ 07974. He is with the Technische Universitat, Berlin, West Germany.

C. Coker, S. E. Levinson and L. R. Rabiner are with Bell Laboratories, Murray Hill, NJ 07974.

I. INTRODUCTION

THE state of the art in speech synthesis by rule is that one can synthesize (from either printed text or from a phonetically based set of input symbols) speech whose intelligibility is quite high but whose naturalness is often poor. One reason for the unnatural quality is the rudimentary state of knowledge as to how to properly control pitch and duration of sounds within a sentence. In order to make improvements in the pitch and duration rules used for synthesis, it would be helpful to be able to compare rule generated synthesis of sentence length material to natural productions of the same sentences. By time aligning events within the sentence, one could modify the duration rules of the synthesizer to improve the quality of the match. By experimenting with a number of sentences (and talkers), one could, hopefully, make major improvements in the duration rules of the synthesizer.

Another area that would benefit from the ability to time align a spoken sentence with another spoken version of the same sentence is speech recognition. One of the most difficult and time consuming problems in building a speech recognizer is collecting data for modeling (statistically or otherwise) the properties of speech sounds. By carefully hand labeling a set of test sentences, one could, in theory, automatically obtain a good set of labels on repetitions of the test sentences by using a time alignment procedure. The time aligned events of

0096-3518/83/0800-0807$01.00