# On the Use of Filter Bank Features for Isolated Word Recognition

*B. A. Dautrich*
*L. R. Rabiner*
*T. B. Martin*

Bell Laboratories
Murray Hill, N.J. 07974

*ABSTRACT.* The vast majority of commercially available isolated word recognizers use a filter bank analysis as the front end processing for recognition. To the designers of such recognizers there exists a myriad of design choices. It is not well understood how these designs choices (e.g. the number of filters, filter spacing, and post-processing of the filter bank outputs) affect recognizer performance. In this paper we present results of a performance evaluation of filter bank recognizers in a speaker-trained, isolated word recognition test using dialed-up telephone line recordings. First we studied the effects of various filter bank parameters on system performance. We designed a total of 13 filter banks including 8 uniform and 5 non-uniform filter banks. Within these 13 filter banks we also considered both slightly and highly overlapping filters. The results indicate that the best performance (highest word accuracy) on the 39 word alpha digits vocabulary, with 4 talkers, is obtained by both a 15 channel uniform filter bank and a 13 channel highly overlapping non-uniform critical band filter bank. Next we studied the effects of selected preprocessing and post-processing techniques on system performance. For this we used the non-uniform 13 channel filter bank. The results indicate that almost none of the processing techniques improved system performance; however, some techniques can potentially reduce hardware cost (computation and storage) without adversely affecting system performance. We also compared the results of the best filter banks recognizers with a conventional LPC based word recognizer on the same vocabulary. The performance of the best filter bank was approximately 4% worse than that of an 8th order LPC-based recognizer. We also studied the effect of additive wideband Gaussian noise on system performance of both the filter bank and the LPC recognizers. Gaussian white noise was added to the speech recordings at signal-to-noises ratios of from 0 to 30 dB. Recognition tests were then performed which indicated that the LPC system performance degraded faster than that of the filter banks; however, the point at which both systems have identical performance is at a signal-to-noise ratio of 6 dB.

## I. Introduction

The vast majority of commercial systems for word recognition use filter bank structures in the front end. Although a number of different filter bank structures have been proposed, there is no simple guideline for choosing a good filter bank for a particular application. By this we mean that there is no comparison (to our knowledge) of the effects on performance (word error rate) of different filter bank structures in an automatic speech recognizer. Even simple questions such as the type of filter bank (FIR or IIR filters), the filter spacing (uniform or non-uniform, 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 post-processed for use in conventional dynamic time warping (DTW) structures [1]. It is also of interest to study the performance of filter bank recognizers in the presence of noise, and to compare them to conventional LPC-based recognizers both with and without additive background noise.

## 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, bandlimited 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 spectral shaping operation (e.g. linear filtering) for increased immunity to finite word length processing in the remainder of the system. The preprocessed signal, $\hat{s}(n)$, is then sent to a filter bank analyzer whose structure is shown in Figure 2. The filter bank contains a set of $Q$ parallel bandpass filters which cover the speech band of interest (100-3200 Hz for telephone speech). Each bandpass filter is followed by a nonlinearity (NL), a lowpass filter (LP), a sampler, and a logarithmic compressor.

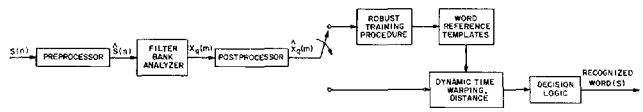


Fig. 1 Block diagram of overall filter bank word recognizer.

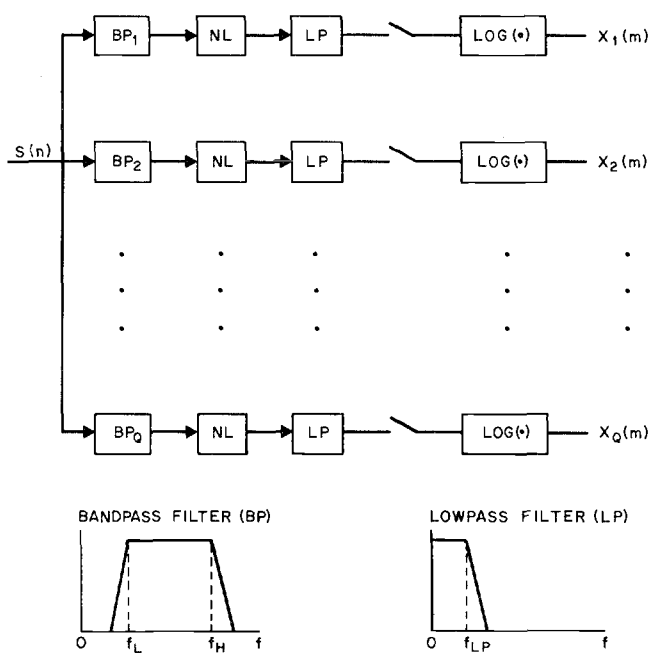


Fig. 2 Block diagram of $Q$-channel filter bank.



ICASSP 83, BOSTON  

## III. Signal Processing Choices in the Recognizer

### 3.1 Preprocessor

The function of the preprocessor is to spectrally shape the speech signal so as 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.

### 3.2 Postprocessing

The forms of postprocessing, which we have studied, include

#### 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.

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 subtract mean energy from the individual channel signals for that frame. For peak normalization we subtract the peak frame energy from the individual channel signals.

#### 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.

#### 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.

#### 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.

## IV. General Design of the Analysis Filter Banks

A variety of considerations go into the choice of filters for the filter bank of Figure 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 infinite impulse response (IIR) filters. Because of their linear phase properties, and because simple implementations are possible, FIR filters were chosen for the bandpass filters [2].

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 on the intended application of the spectrum; values of $Q$ from 10 to 32 have typically been used in vocoder applications [3].

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 channels are equally spaced across the band of interest.

An alternative filter bank is to choose channel bandwidths equally spaced on a logarithmic frequency scale. Alternative ways of choosing filter spacings are available including the so-called critical band [4] filter banks (with channels uniform until 1000 Hz and then logarithmic above 1000 Hz), and arbitrarily spaced filter banks where other considerations are used in designing the individual filters.

Once we have designed the necessary bandpass filters, the next step is to choose the nonlinearity and design the required lowpass filter. The nonlinearity chosen for this study was a full-wave rectifier. This is standard for most filter bank applications. For the lowpass 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 lowpass filter was chosen to be 30 Hz to allow for sampling the channel outputs at a rate of 67 Hz. The desired lowpass filter was realized using a third-order Bessel IIR filter.

## V. Design of the Filter Banks

### 5.1 Uniform Filter Bank Designs

For the uniform filter banks, we chose to look at four different values of $Q$, namely 3, 7, 15, and 31 (filters). With these choices for $Q$, $M$ (the frame shift) was chosen to be 10 samples for the first 3 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 lowpass window function it was decided that a Kaiser window should be used [5]. 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 the Kaiser window to design an appropriate lowpass filter by using the well-known window design technique [6]. 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 [7]. The second was to use the window directly since the window is essentially a lowpass 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. In this way the premise that a flat composite spectrum is necessary to obtain good recognition accuracy could be tested.

### 5.2 Non-uniform Filter Banks

For the non-uniform filter banks we chose to investigate three different filter bank spacings; octave spacing, critical bands, and 1/3 octave spacing. In particular we considered 4, 7 and 12 channel filter banks for the octave, critical band, and 1/3 octave filter banks. The ideal filter characteristics of the filters in the critical band filter bank is based on the Articulation Index [4]. The filters in this filter bank were spaced to incorporate two critical bands in each filter.

In addition to the above set of uniform and non-uniform filter banks, two specially designed non-uniform filter banks were studied. The first was a 5-channel filter bank designed for use in the IBM speech terminal by Silverman and Dixon [8]. 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 Hz and 3200 Hz respectively.

The second specially designed filter bank was based on the system by Martin [9]. The filters used were spaced along critical bands; however the frequency selectivity of these filters was very poor (the ratios of center frequency to bandwidth were about 8). This poor frequency selectivity was chosen to provide good time resolution. The filters in this filter bank are highly overlapping in contrast to all previous cases where there was little or no filter overlap.

## VI. 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 [10]. The measured recognition accuracy for this vocabulary has been shown to be relatively low in previous tests [11]. Thus small differences in system performance can often be reliably measured with a reasonable size test 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 [12]. This was done for each of four talkers (2 male, 2 female) for all thirteen filter banks. Next an independent test set, consisting of 10 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.

Error rates were determined for five separate experiments. The first experiment consisted of testing each filter bank with the entire test set for each of the four talkers. Using standard statistical tables, it can be shown that at the 99% confidence level, a difference of about 1.5-2% in error rate is statistically significant.

The second experiment consisted of measuring the performance of a standard LPC system [11] on each test set used in the first experiment. The purpose of this experiment was to determine the relative error rates of the LPC and filter bank system.

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 fourth experiment used one of the four top filter banks (the 13-channel Martin filter bank) and varied the preprocessing and postprocessing parameters (as mentioned in Sections 3.1 and 3.2) in an effort to improve overall recognizer performance.

In the fifth experiment broad band white noise was added to the speech signal, at specified signal-to-noise ratios, prior to recognition, and the performance of both the 13-channel filter bank and the LPC recognizer were measured.

All filter bank systems were simulated on a general purpose minicomputer. The simulations took about 9 months (12 hours/day) of computer time to run.

### 6.1 Results for Experiments 1 and 2

The results for experiments 1 and 2 are given in Table I which shows error rates for each talker (and the average) for each of the 13 filter banks and the conventional LPC recognizer.

Several general trends emerge from data of Table I. First we see that the word error rates differ greatly among talkers — i.e. talker 1 had about an 8.5% word error rate ($Q$=15, window design) whereas talker 4 had an 18.5% word error rate with the same conditions. This variation in error scores is typical for the alphadigits vocabulary [10], and the scores of the four talkers fall within the normally expected range.

The second trend is that the uniform filter banks with a flat composite spectrum generally do better than those having the same

| | $M_1$ | $M_2$ | $F_1$ | $F_2$ | Avg |
|---|---|---|---|---|---|
| $Q$=3, flat | 26.2 | 20.8 | 29.0 | 33.8 | 27.5 |
| $Q$=7, flat | 11.0 | 6.4 | 11.8 | 19.7 | 12.2 |
| $Q$=15, flat | 8.5 | 6.2 | 12.6 | 18.5 | 11.5 |
| $Q$=31, flat | 4.6 | 5.6 | 24.6 | 24.4 | 14.8 |
| $Q$=3, non-flat | 25.9 | 25.9 | 31.5 | 35.9 | 29.8 |
| $Q$=7, non-flat | 13.6 | 10.0 | 12.3 | 21.0 | 14.2 |
| $Q$=15, non-flat | 10.3 | 5.9 | 18.5 | 24.1 | 14.7 |
| $Q$=31, non-flat | 10.5 | 10.5 | 42.3 | 33.3 | 24.1 |
| $Q$=4, octave | 12.6 | 13.1 | 24.1 | 23.3 | 18.3 |
| $Q$=5, IBM | 12.6 | 9.6 | 13.1 | 22.8 | 14.5 |
| $Q$=7, critical | 10.0 | 6.7 | 16.7 | 19.0 | 13.1 |
| $Q$=12, 1/3 octave | 9.5 | 9.5 | 23.8 | 31.5 | 19.6 |
| $Q$=13, Martin | 9.0 | 5.4 | 13.1 | 18.7 | 11.6 |
| LPC | 5.1 | 4.1 | 10.3 | 11.8 | 7.8 |

Results for Experiment 1 and 2
Word Error Rate (%)

**Table I**

number of channels with a composite spectrum which contains valleys. This result is essentially independent of the number of channels of the filter bank. For the males for the uniform filter bank 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 due to interactions between filter bandwidths and pitch for female talkers.

The results for the non-uniform filter banks show that the word error rates again differ greatly among talkers. It is also seen that the 13-channel filter bank does not follow the trend (for the females) of increasing error rates as the number of filters increases. This is due to the fact that the thirteen-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. The results for the standard LPC system (as seen in the last line in Table 1) show that the LPC system has, on average, a 4% lower error rate than the best of the filter bank recognizers.

### 6.2 Results for Experiment 3

The results of the third experiment, in which the vocabulary was limited to include only the digits, showed that essentially no errors were made for either type of recognizer.

### 6.3 Results for Experiments 4 and 5

The results for experiments 4 and 5 were:

1. Essentially none of the proposed pre and postprocessing techniques for use in the filter bank word recognizer led to an improvement in performance of the system (i.e. reduced word error rate) over that obtained with simple processing. 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.
2. The filter bank coefficients (for telephone inputs) needed only about 6 uniform bits for a representation with no increase in word error rate from that with no quantization. Hence the storage requirements on the $Q$ = 13 channel recognizer were about 78 bits per frame using this 6-bit coding scheme.
3. The use of a 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 didn't increase word error rate significantly for either the LPC or FB recognizers.

4. 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.
5. Recognizer performance began to degrade for signal-to-noise ratios less than 24 dB.
6. The LPC word recognizer gave error rates the same or lower than the filter bank (FB) word recognizer for SNR $\geq$ 6 dB.

## VII. Discussions

Our general conclusions are as follows.

1. For all filter banks studied here performance degrades for filter banks with too few filters or too many filters. 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 non-uniform) 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 non-uniform filter banks, system performance is best when the filters are spaced along a critical band frequency scale (as opposed to octave bands, 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 highly desirable for filter bank systems.
4. The performance of 7-band and 13-band critical band filter banks was essentially the same as for 7-band and for 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 uniformly 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% lower than that of the *best* filter bank recognizer. For the digits vocabulary, the performances of both the LPC and the best filter bank recognizers were comparable with error rates close to 0%.
6. Neither time nor frequency smoothing of channel vectors aided performance of the filter bank recognizer.
7. 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.

## VIII. 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 somewhat 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 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 ratios less than about 24 dB.

## References

[1] C. S. Myers, L. R. Rabiner, and A. E. Rosenberg, "Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing,* Vol. ASSP-28, pp. 623-635, Dec. 1980.

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

[3] J. L. Flanagan, *Speech Analysis, Synthesis and Perception,* Second Edition, Springer-Verlag, New York, 1972.

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

[5] J. F. Kaiser, "Nonrecursive digital filter design using the $I_0$-sinh window function," *Proc. IEEE,* Vol. 65, No. 11, pp. 1558-1564, Nov. 1977.

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

[7] R. W. Schafer, L. R. Rabiner, and O. Herrmann, "FIR digital-filter banks for speech analysis," *Bell System Tech. J.,* Vol. 54, No. 3, pp. 531-544, March 1975.

[8] H. F. Silverman, and N. R. Dixon, "State Constrained Dynamic Programming (SCDP) for discrete utterance recognition," Proc. ICASSP 80, Vol. 1, pp. 169-172.

[9] T. B. Martin, *Acoustic Recognition of a Limited Vocabulary in Continuous Speech,* Ph.D Dissertation, Univ. of Penn., University Microfilms Limited, Ann Arbor, Michigan, 1970.

[10] L. R. Rabiner, A. E. Rosenberg, J. G. Wilpon, and W. J. Keilin, "Isolated word recognition for large vocabularies," submitted for publication.

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

[12] 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.