

# On the Use of Bandpass Liftering in Speech Recognition

BIING-HWANG JUANG, MEMBER, IEEE, LAWRENCE R. RABINER, FELLOW, IEEE,  
AND JAY G. WILPON, MEMBER, IEEE

**Abstract**—In a template-based speech recognition system, distortion measures that compute the distance or dissimilarity between two spectral representations have a strong influence on the performance of the recognizer. Accordingly, extensive comparative studies have been conducted to determine good distortion measures for improved recognition accuracy. Previous studies have shown that the log likelihood ratio measure, the likelihood ratio measure, and the truncated cepstral measures all gave good recognition performance (comparable accuracy) for isolated word recognition tasks.

In this paper we extend the interpretation of distortion measures, based upon the observation that measurements of speech spectral envelopes (as normally obtained from standard analysis procedures such as LPC or filter banks) are prone to statistical variations due to window position fluctuations, excitation interference, measurement noise, etc., and may not accurately characterize the true speech spectrum because of analysis model constraints. We have found that these undesirable spectral measurement variations can be partially controlled (i.e., reduced in the level of variation) by appropriate signal processing techniques. In particular, we have found that a bandpass "liftering" process reduces the variability of the statistical components of LPC-based spectral measurements and hence it is desirable to use such a liftering process in a speech recognizer. We have applied this liftering process to several speech recognition tasks: in particular, single frame vowel recognition and isolated word recognition. Using the liftering process, we have been able to achieve an average digit error rate of 1 percent in a speaker-independent isolated digit test. This error rate is about one-half that obtained without the liftering process.

## I. INTRODUCTION

**S**PEECH recognition tasks involve such necessary steps as analysis, similarity calculation, time normalization, and decision logic, as depicted in Fig. 1. The analysis procedure, performed on the raw input speech waveform, results in some representation of the signal which characterizes the relevant features of the spoken speech. It can be regarded as a data reduction procedure that retains the vital characteristics of the signal and eliminates undesirable interference from irrelevant characteristics of the speech, thus easing the inference or decision making process in the later stages. The automatic speech recognition task, strictly speaking, starts from the calculation that measures the distance or dissimilarity between the unknown and a set of stored reference patterns. The choice of dissimilarity or distortion measure is extremely important since the final recognition decision is generally based

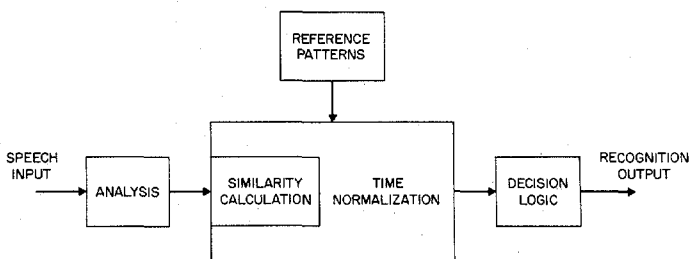


Fig. 1. A block diagram of a general speech recognizer.

entirely upon the calculated distances. Accordingly, extensive comparative studies have been conducted in order to find a distortion measure [1]–[3] that gives good recognition accuracy.

There exists an almost infinite number of distortion measures. An exhaustive comparison is clearly impossible. The key question is what makes a distortion measure at least a good, if not the best, choice. Qualitatively, as with the analysis procedure, a good distortion measure should be sensitive to differences in the vital characteristics of the unknown (test) and the reference patterns, and insensitive to the irrelevant variations among observations of the unknown or reference patterns. In this regard, the analysis procedure and the similarity calculation have the same objective, i.e., to provide a robust and reliable measurement of the information bearing features in the spoken input. The problem of finding a good distortion measure is then equivalent to that of finding a good data reduction/measurement procedure, and the above objective can be accomplished, to some extent, by either of these two steps in the recognition algorithm.

In this paper we propose and discuss a data reduction or measurement procedure that is slightly more complicated than the standard procedure to be followed by a simple distortion calculation step in the automatic speech recognition task chain. There are two reasons why this type of recognition structure is more desirable than trying to find a good distance measure that works for a given analysis or data reduction process. First, the analysis method that is used might not be effective in eliminating the irrelevant signal variability, and might actually inadvertently remove some desired speech information. This loss of information is generally not correctable in the similarity calculation step. Second, the required distortion mea-

sure that works well for a given analysis procedure could be very complicated to evaluate. Since distortion calculation, which is usually embedded in the time normalization procedure, requires most of the computing resources in a speech recognizer, a small increase in distortion measure complexity often corresponds to a large increase in the computational requirements of the overall system.

In the next section, we describe the possible sources of signal variability in some typical spectral analysis procedures (e.g., filter banks and LPC analysis). A simple simulation of some statistical variabilities of the measured (analyzed) parameters is given to facilitate the presentation. In Section III, we describe a bandpass liftering procedure which effectively reduces the undesired variability in the components of the spectral measurements. We then report results of experiments using the liftering procedure applied to the tasks of vowel and word recognition in Section IV. Finally, we discuss and summarize our findings in Section V.

## II. SOURCES OF SPECTRAL VARIABILITY

Most speech recognition systems use some type of spectral analysis on the raw speech input waveform. The two types of spectral analysis methods most frequently employed are filter bank analysis and linear prediction.

Filter bank approaches typically use a bank of from 8 to 32 bandpass filters, either uniformly spaced or critical-band-spaced (generally highly overlapped) to cover the relevant frequency range of the input signal (typically 0–3 kHz for telephone input, 0–8 kHz for broad-band input [10]). Time and frequency resolution is an important factor in the filter bank design [10]. The output signal of each bandpass filter is generally passed through a magnitude nonlinearity and a low-pass filter. The output signal essentially represents the time varying energy of the speech signal over the band of the bandpass filter. The resulting low-pass signals are then typically sampled at a rate of 50–100 Hz so that a spectral measurement vector is obtained every 10–20 ms.

One advantage of the filter bank approach, or other DFT-based approaches, is that each bandpass channel is treated essentially independently, i.e., there are no global spectral constraints on the filter bank outputs. Artifacts of the speech channel, over which the speech is transmitted, such as noise contamination or spectral dips in the transfer function, etc., linearly affect the spectral vectors and are relatively easy to deal with in comparing spectral vectors. On the other hand, spectral measurements from the output of filter banks are sensitive to variations in the speech excitation, such as changes in fundamental frequency from utterance to utterance. Since these variations are inevitable and unavoidable in natural speech, they become the main factor that makes spectral measurements and comparisons unreliable, unless the bandpass filters are somehow adapted to compensate for these effects.

The usual alternative to filter bank analysis is linear prediction analysis. As is well known, linear prediction

analysis uses an all-pole spectrum to model the short time speech spectrum. The resulting linear prediction coefficients form an all-zero polynomial which characterizes the spectral pattern being measured. One of the main advantages of the linear prediction analysis method is that it leads to a consistent and meaningful resolution of the source-tract interaction. For this reason many recognition systems use linear prediction analysis as the first step in the analysis or data reduction procedure.

The linear prediction method has its own drawbacks, however. The all-pole representation of the speech spectrum is a kind of constraint on spectral continuity that is lacking in the filter bank approach. This all-pole constraint, although it alleviates excessive sensitivity to fundamental frequency variations in the speech excitation, can create other spurious spectral components that are not very desirable in speech recognition applications. These spurious components will be further explained in the latter sections. Also, the LPC spectrum includes components that are related to the talkers' glottal shape and vocal cord duty cycles. Inclusion of these components in the spectral measurement often reduces the effectiveness of the spectral representation for *speaker-independent* speech recognition problems. The fact that a distortion measure, consistent with the linear prediction analysis, i.e., the likelihood ratio measure, exists does not easily remedy this problem. We discuss the control of these undesirable spectral components through the consideration of LPC cepstrum variabilities in the following.

### A. Global Variability of LPC Cepstrum

Let  $S(\omega)$  be the Fourier transform of the speech signal,  $s(n)$ , and  $F(\omega) = \log S(\omega)$ . The complex cepstrum coefficients [4] are defined by

$$c_k = \int_{-\pi}^{\pi} F(\omega) e^{j\omega k} \frac{d\omega}{2\pi}, \quad -\infty < k < \infty. \quad (1)$$

If  $S(\omega)$  is modeled by a  $p$ th-order all-pole spectrum  $\sigma/A(\omega)$  where  $A(z) = 1 + a_1 z^{-1} + \dots + a_p z^{-p}$ , then the following well-known recursion formula relates the (real) LPC cepstrum coefficients of the speech signal to the above polynomial coefficients:

$$-kc_k - ka_k = \sum_{n=1}^{k-1} (k-n) c_{k-n} a_n \quad \text{for } k > 0. \quad (2)$$

It can be shown that under certain common conditions, the cepstral coefficients, except  $c_0$ , have 1) zero means and 2) variances essentially inversely proportional to the square of  $k$ , the coefficient index, i.e.,

$$E\{|c_k|^2\} \sim \frac{1}{k^2} \quad (3)$$

(see Appendix). To verify the above result based upon simple statistical estimate, the variance of each of the first 16 LPC cepstral coefficients (with the order  $p = 8$ ) from a collection of more than 10 000 frames (200 s) of different speech signals including a wide range of sounds

from 7 speakers was measured. The measured results, normalized by the variance of  $c_1$ , are plotted in Fig. 2 (mixed data). The relationship of (3), or more closely, (A7) for some value of  $\beta$  in the Appendix, can be observed from the plot. Similar measurement results were also reported in [5].

We can incorporate this  $k^2$  factor into the traditional cepstral measure so as to normalize the contributions from each cepstral term, thereby giving, as a weighted distance, the measure

$$\begin{aligned} d(C, C') &= \sum_k k^2 (c_k - c'_k)^2 \\ &= \sum_k (kc_k - kc'_k)^2 \end{aligned} \quad (4)$$

where  $C$  and  $C'$  are two cepstral sequences,  $C = [c_1 c_2 \dots]$  and  $C' = [c'_1 c'_2 \dots]$ . The  $kc_k$  sequence is often referred to as root power sums and (4) thus coincides with the root power sums measure proposed by Paliwal [6] in the study of vowel recognition. Note that if the summation in (4) is over the entire cepstral sequence, i.e.,  $k = 1, 2, \dots, \infty$ , the distance measure of (4) becomes the  $L_2$  measure based upon the differences between slopes (first order derivatives) of the corresponding log LPC spectra pair.

### B. Variability Components

In this section, we try to identify the cepstral variability components that are primarily artifacts of the analysis procedure and/or those that are simply not useful for speech recognition. This is in part accomplished by contrasting the above global variabilities to the coefficient variabilities due to only a fixed filter simulated signal.

A zero mean Gaussian independently, identically distributed noise signal was used as the excitation for a fixed, 8th-order, all-pole filter, with reflection coefficients typical of those of a vowel sound, i.e.,  $[-0.3301, 0.2251, -0.3992, 0.2806, 0.3038, 0.6082, -0.1013, 0.1799]$ . The output signal was then analyzed using an 8th-order linear prediction analysis with a 160-point Hamming window applied. The analysis results were then converted to the cepstrum domain using the recursion formula (2). A total of 767 vectors was obtained and the variances were accordingly computed. These variances for  $c_1$  through  $c_{16}$  normalized by that of  $c_1$  are also plotted in Fig. 2 (fixed filter simulated data). For better contrast, we show in Fig. 3 the ratio of the variances of the simulated fixed filter data to the variances of the mixed data. It is clearly seen from the figures that the variances for the higher quefrency terms are relatively large compared to those from the general overall statistical model of the mixed data. The increase in variance ratio with increasing coefficient index indicates the *diminishing* discriminating power of the higher quefrency terms. This shows that the variability of higher quefrency terms are inherent artifacts of the analysis procedure, and hence less desirable in spectral similarity comparisons than lower quefrency terms. Therefore, one should not depend on a simple spectral

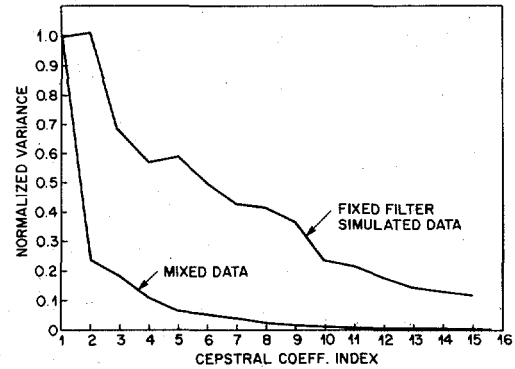


Fig. 2. Normalized variances of the first 16 LPC cepstral coefficients measured from 1) real, mixed speech data, and 2) simulated data, obtained by driving an 8th-order fixed all-pole filter with a Gaussian i.i.d. sequence.

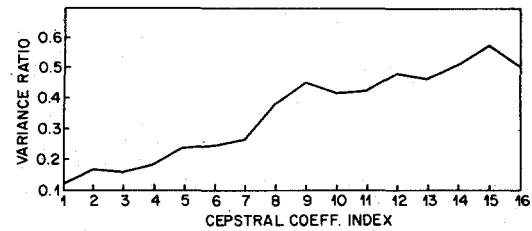


Fig. 3. Variance ratio between the LPC cepstral variances of mixed speech data and fixed filter simulated data.

slope measure or equivalently the untruncated root power sums measure as they tend to overemphasize high quefrency terms.

It is important not to confuse this diminishing discriminating power of the higher quefrency cepstral terms with the dependence of higher cepstral terms upon lower cepstral terms due to the recursion of (2). On the one hand, one can use a truncated cepstral measure [2] with the number of terms equal to the LPC analysis order, arguing that only these terms are needed to reconstruct the LPC spectrum. On the other hand, an untruncated cepstral measure is equivalent to the  $L_2$  measure of the log LPC spectral differences. It is not yet fully understood whether or not there exists, between the two extremes, some measure, which could give improved performance.

Next we discuss the origin of the variability of the low quefrency terms of the cepstrum. The variability of low quefrency terms is primarily due to variations in transmission, speaker characteristics, and vocal efforts, etc., of the speech. Different transmission channels usually have different frequency responses and these differences generally affect the low quefrency terms much more than the high quefrency terms. For example, the effect of differences in channel frequency response rolloff is usually most prominent in the first couple of cepstral coefficients. Additionally, as mentioned before, the LPC spectrum also includes components that are strong functions of the speaker's glottal shape and vocal cord duty cycles. To first order, these components appear as the overall spectral tilt which, in turn, affects mainly the first few cepstral

coefficients. In a speaker-independent speech recognition environment, variations in these components are thus responsible for a significant portion of the variability of the first few cepstral coefficients. Such variability diminishes the discriminating capability of the corresponding cepstral terms.

Another undesirable component of spectral variability, particularly associated with LPC analysis, occurs when the signal spectrum has spectral notches or zeros. These spectral zeros may be the result of transmission, filtering, or even improper preemphasis. Linear prediction analysis uses the criterion of minimizing the output residual energy of the inverse filter; it is equivalent to choosing the all-pole filter that maximizes the spectral flatness of its output [4]. When spectral notches or zeros are present, the analysis results vary significantly for different signals, particularly around the regions of spectral zeros, due to the overall, fixed-order, all-pole model constraints. This type of variation often results in excessively high variability in low quefrency cepstral terms.

The above discussion points to the necessity of applying some type of cepstral liftering window to remove or suppress the undesirable variations present in the high and low quefrency LPC cepstral coefficients.

### III. PROPOSED LIFTERING PROCEDURE FOR SPEECH RECOGNITION

The liftering procedure we propose here is very straightforward. It is simply windowing in the cepstral (quefrency) domain. Fig. 4 depicts the procedure, in the form of a modified front-end processor for the recognizer. The speech signal is first analyzed with the linear prediction method. The predictor coefficients are transformed into the cepstral coefficients, using the recursion formula (2), up to the desired number of terms. A prechosen window  $w(k)$  is then applied to the cepstral vector. The resulting windowed cepstral vector is used in the recognizer, with a simple Euclidean distance as the distortion or dissimilarity measure.

The effect of this liftering process can be visualized by inverse transforming the windowed cepstral vector back to the log spectrum domain. Fig. 5 shows a series of liftered log spectra, with their original LPC all-pole spectrum plotted at the bottom. The window used in liftering is of the form  $w(k) = 1 + h \sin(\pi k/L)$  where  $h = L/2$ , for  $k = 1, 2, \dots, L$  and  $w(k) = 0$  for other  $k$ . We varied  $L$  from 8 (the top curve) to 16 (the curve above the LPC log spectrum). As is clearly shown, the sharp spectral peaks in the LPC log spectrum are smoothed. The shape of these peaks is characteristic of the LPC log spectrum. While these peaks essentially represent the "formants" of the signal and are important in characterizing the sound, their shapes create unnecessary sensitivity in the spectral comparison. The liftering process tends to reduce the unnecessary sensitivity by smoothing these peaks without distorting the fundamental formant structure. Furthermore, the LPC log spectral tilt of approximately 8 dB/octave, as shown in the figure, is effectively removed.

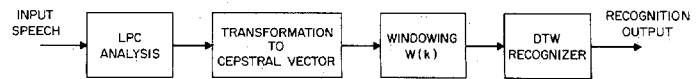


Fig. 4. A liftering procedure.

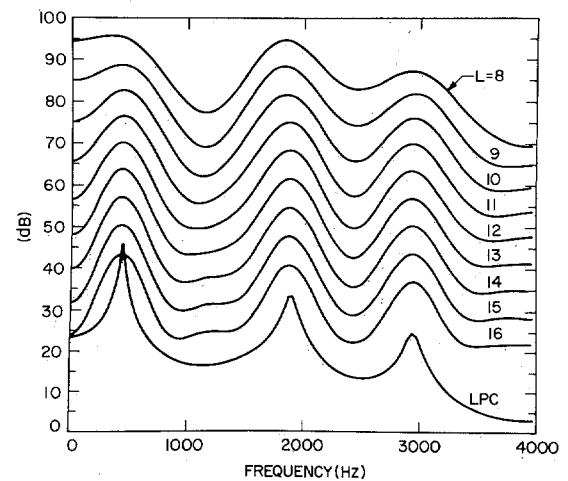


Fig. 5. Effects of liftering on LPC log spectra.

This is, of course, a result of the deemphasis of the low quefrency cepstral terms.

The effects of the liftering process, on recognition, are demonstrated through a sequence of spectral plots. Fig. 6(a) is a hidden line plot of a series of 30 consecutive LPC log spectra, corresponding to a vowel-like sound. The randomness of the spectral components and the sharp spectral peaks that lead to excessive spectral sensitivity are clearly seen. A liftering window of the form  $w(k) = 1 + 6 \sin(\pi k/12)$  for  $k = 1, 2, \dots, 12$ , and  $w(k) = 0$  otherwise, is then applied, and the corresponding smoothed spectral sequence is plotted in Fig. 6(b). It is seen that the undesirable (noiselike) components of the LPC spectral measurements are reduced or removed and the essential characteristics of the "formants" are retained. Applying liftering to the LPC spectra is certainly different from direct cepstral smoothing on the signal. To show how the two may differ, the corresponding segment of signal is cepstrally smoothed with the same window  $w(k) = 1 + 6 \sin(\pi k/12)$  and the result is plotted in Fig. 6(c). As can be seen, this spectral sequence is not as smooth as the liftered LPC spectra. Although these figures do not directly indicate the contribution of liftering to the recognition results, they do show that the undesired variability of the spectral measurements is greatly reduced.

### IV. EXPERIMENTAL RESULTS

We applied the liftering process to the tasks of recognizing vowels from single frame spectra, and isolated digits in a speaker-independent environment. Before we performed the actual recognition tests, we first studied the effects of various types of liftering windows.

#### A. Choice of Liftering Window

We considered only the following three types of liftering windows.

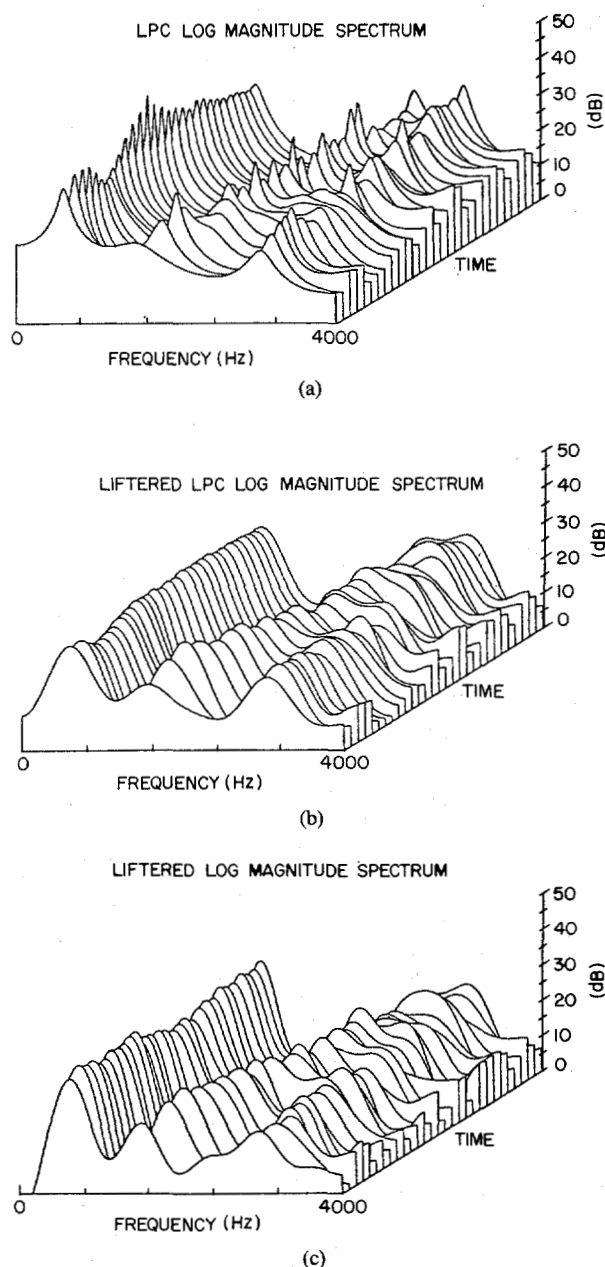


Fig. 6. (a) A consecutive sequence of log LPC spectra; (b) the result of applying a liftering window to the LPC spectral sequence; (c) the corresponding spectral sequence obtained by direct cepstral smoothing on the signal without the intermediate LPC modeling stage.

$$\text{Type 1) } w_1(k) = 1, \quad k = 1, 2, \dots, L \\ = 0, \quad \text{otherwise.}$$

$$\text{Type 2) } w_2(k) = 1 + h \cdot (k - 1) / (L - 1), \\ k = 1, 2, \dots, L \\ = 0, \quad \text{otherwise.}$$

$$\text{Type 3) } w_3(k) = 1 + h \sin(k\pi/L), \\ k = 1, 2, \dots, L \\ = 0, \quad \text{otherwise.}$$

These liftering windows were used to recognize a partic-

ular isolated digit set, consisting of a total of 1000 utterances (100 utterances for each digit). The linear prediction analysis was 8th order. It is important to note that the above windows are not the only possible choices. Our purpose here is mainly to demonstrate the possibility of controlling undesirable spectral variabilities through liftering for better recognition accuracy.

The Type 1 window is rectangular. For  $L = 8$  and 12, the number of misrecognized digits was 38 and 35, respectively, out of 1000 trials. The Type 2 window is triangular. Two window lengths were studied, namely,  $L = 10$  and 12. For each fixed length window, we also examined the effects of the height  $h$  upon the recognition accuracy. The results, in terms of number of errors, in 1000 recognition trials, are summarized in Table I. It can be seen that although the  $L = 12, h = 10$  window gave the fewest digit errors (11), the sensitivity of the results to different values of  $L$  and  $h$  was small. The Type 3 window,  $w_3(k)$ , is a raised sine. Two cases,  $L = 12$  and 14, with varying height, were studied. Table II summarizes the recognition results, also in terms of number of recognition errors in 1000 trials. As can be seen from the table, the best result occurred when the height was approximately one-half of the window length. We then investigated this particular form of liftering window:  $w(k) = 1 + 0.5L \sin(\pi k/L)$ ,  $1 \leq k \leq L$ . We varied  $L$  from 8 (the original LPC order) to 16. The number of recognition errors for this case is given in Table III. As shown, the recognition accuracy essentially increases with the liftering window length. Beyond  $L = 12$ , however, the number of digit errors stays the same up to the tested maximum of 14. These results suggest that a liftering window of the form

$$w(k) = 1 + 6 \sin(\pi k/12) \quad (5)$$

is a good choice for recognition experiments.

### B. Single Frame Vowel Recognition

The database used for this speaker-trained recognition test consisted of all vowel frames that occurred in 10 occurrences of 10 carrier words, each one with a single characteristic vowel. One-half of the vowel frames were used as a training sequence to design vector quantization (VQ) codebooks with 1, 2, 4, and 8 vectors per vowel [7]. The other one-half of the vowel frames were used as an independent test set. Seven talkers (four male, three female) were used in the test. Recognition was performed on single frames by finding the vowel codebook whose distance to the test vector was minimum.

Five distinct types of distortion measures were used in the test including: the likelihood ratio, a weighted likelihood ratio, a cepstral distance, a weighted cepstral distance, and a bandpass lifter of the type given in (5). Further details of the individual distance measures are given in [7]. The key result was that, on average, the cepstral lifter provided the best recognition performance for all size codebooks that were tested.

TABLE I  
TOTAL NUMBER OF RECOGNITION ERRORS (OUT OF 1000 TRIALS) WITH THE TRIANGULAR LIFTER

$h$	6	8	10	12	14	16
$L = 10$	13	15	14	15	—	—
$L = 12$	13	14	11	12	13	15

TABLE II  
TOTAL NUMBER OF RECOGNITION ERRORS WITH THE RAISED SINE LIFTER AS A FUNCTION OF THE WINDOW HEIGHT  $h$  AND THE WINDOW LENGTH  $L$

$h$	2	3	4	5	6	7	8	10
$L = 12$	13	—	14	—	10	—	12	11
$L = 14$	—	11	10	11	10	10	—	—

TABLE III  
TOTAL NUMBER OF RECOGNITION ERRORS WITH THE  $1 + 0.5L \sin(\pi k/L)$  LIFTER

$L$	8	9	10	11	12	13	14
errors	22	16	13	14	10	10	10

### C. Speaker-Independent Isolated Digit Recognition

The database used for the speaker-independent isolated digit recognition test consisted of 4 sets of isolated utterances of digits. Each set of data contained 1000 utterances, 100 for each digit, spoken by 100 different speakers, 50 male and 50 female. Different data sets were from different sets of speakers. This database has been studied previously, and a more detailed description can be found in [8].

We used the liftering window of (5) throughout the test, including the generation of the reference templates. The recognizer was a DTW-based system using a Euclidean distance measure. Furthermore, the energy or gain term was *not* included in the spectral representation and comparison. We studied the recognition accuracy as a function of the number of reference templates per digit. The results are summarized in Table IV. As seen from the table, using 12 reference templates per digit, the average error rate for speaker-independent recognition of isolated digits was only 1 percent, i.e., it was about one-half of the error rate under the same DTW framework, but instead using a standard LPC analysis and a log likelihood distance *with energy terms* incorporated. Even with only 6 templates per digit, the recognition accuracy was higher than that reported in [8] with 12 templates per digit. Finally, the effect of going from 12 templates per digit to a single template per digit increased the error rate by 2.68 percent. This increase, although significant compared to the 1 percent error rate currently obtained at 12 templates per word, is not much larger than error rates obtained from earlier studies [9]. In fact, the error rate for 3 templates

TABLE IV  
TOTAL NUMBER OF RECOGNITION ERRORS USING THE LIFTERING WINDOW IN A DTW RECOGNIZER FOR SPEAKER-INDEPENDENT ISOLATED DIGIT RECOGNITION

Number of Templates per Digit	Data Sets				Errors	
	DAT-1 <sup>a</sup>	DAT-2	DAT-3	DAT-4	Total	Percent
1	29	30	38	50	147	3.68
3	22	25	35	35	117	2.93
6	7	8	31	18	64	1.60
9	5	9	24	12	50	1.25
12	1	7	21	11	40	1.00

<sup>a</sup> Training Set

per digit is comparable to the error rate with 12 templates per digit reported earlier [8] when energy was not used in the recognition scheme (which is the case here). Furthermore, for the single template per digit case, the error rate is only 1 percent higher than that obtained previously with 12 templates per digit [8], and the computation rate is reduced by a factor of 12.

The improved performance of the isolated digit DTW recognizer is primarily due to the increased reliability of the spectral measurements via the described liftering procedure. The small degradation in going from 12 templates per digit to a single template per digit gives strong evidence of this result. Since the liftering process simply filters out undesirable variability, and transforms the original measurement vector to a more reliable one, it can be used in other recognition schemes, such as the hidden Markov model system [8] as well. Further improvements may still be possible when other parameters such as the energy term or a durational model [8] are incorporated into the scheme.

### V. SUMMARY

We have presented a discussion of how highly variable spectral measurement components can be identified and suppressed using a liftering procedure. It has been shown that the liftering procedure enhances the reliability of the transformed spectral measurements, making the spectral comparison more appropriate for the recognition task. The increased recognition accuracy has been demonstrated in both a single frame vowel recognition task and in an isolated digit recognition task. For single frame vowel recognition, the proposed liftering scheme achieved the highest accuracy of all the tested systems. For speaker-independent, isolated digit recognition, the system with liftering achieved an average digit error rate of about 1 percent, using 12 templates per digit. Furthermore, the test results showed that the number of reference templates per digit could be significantly reduced (from 12 to 3) without strongly degrading the recognition performance. With a single reference template per digit, the recognition accuracy was comparable to that of previous schemes where as many as 12 templates per digit were used. Since the liftering process can be viewed as a transformation of

the spectral parameter vectors to a more reliable one, it can be applied as well in other recognition systems, such as the probabilistic scheme of hidden Markov models.

#### APPENDIX

The goal of the statistical analysis in this appendix is to provide insight into the range of variation of the undesirable components of the spectral measurements. (It is *not* our intention here to advocate a new statistical analysis method for as complicated a signal as speech.)

Let  $S(\omega)$  be the Fourier transform of the speech signal  $s(n)$ , which we consider to be a (short-time) stationary process, and  $F(\omega) = \log S(\omega)$ . Then the complex cepstrum coefficients [4] are

$$c_k = \int_{-\pi}^{\pi} F(\omega) e^{j\omega k} \frac{d\omega}{2\pi}, \quad -\infty < k < \infty. \quad (A1)$$

The expected value of the complex cepstrum is

$$E\{c_k\} = \int_{-\pi}^{\pi} E\{F(\omega)\} e^{j\omega k} \frac{d\omega}{2\pi}. \quad (A2)$$

For illustration purposes, we assume that  $E\{F(\omega)\} = A$ , a constant. (This is clearly incorrect for speech, but it enables us to estimate the variability of the cepstral coefficients.) Then

$$\begin{aligned} E\{c_k\} &= A \int_{-\pi}^{\pi} e^{j\omega k} \frac{d\omega}{2\pi} \\ &= A \frac{\sin k\pi}{k\pi} \end{aligned} \quad (A3)$$

which is zero except for  $k = 0$ , where it is the value  $A$ .

The second moment of  $c_k$  is

$$\begin{aligned} E\{c_k c_k^*\} &= \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} E\{F(\omega_1) F^*(\omega_2)\} \\ &\quad \cdot e^{j\omega_1 k} e^{-j\omega_2 k} \frac{d\omega_1}{2\pi} \frac{d\omega_2}{2\pi} \\ &= \frac{1}{2\pi} \int_{-2\pi}^{2\pi} \left(1 - \frac{|\phi|}{2\pi}\right) G(\phi) e^{jk\phi} d\phi \end{aligned} \quad (A4)$$

where  $\phi = \omega_1 - \omega_2$ , and  $E\{F(\omega_1) F^*(\omega_2)\}$  is assumed to be a function of  $\omega_1 - \omega_2$ , denoted by  $G(\omega_1 - \omega_2) = G(\phi)$ .

Obviously, if  $G(\phi) = B\delta(\phi)$ , where  $\delta$  is the Kronecker delta, and  $B$  is a constant, the second moment is

$$E\{|c_k|^2\} = \frac{1}{2\pi} \int_{-2\pi}^{2\pi} \left(1 - \frac{|\phi|}{2\pi}\right) B\delta(\phi) e^{jk\phi} d\phi = \frac{B}{2\pi}. \quad (A5)$$

This is the case when the spectral components of the signal at frequencies  $\omega_1$  and  $\omega_2$  are uncorrelated.

A somewhat more realistic model of the correlation

function for the spectral components  $G$  is of the form

$$G(\phi) = e^{-\beta|\phi|}, \quad -\pi < \phi < \pi. \quad (A6)$$

Since  $s(n)$  is a discrete time signal,  $G(\phi)$  is periodic, i.e.,

$$G(\phi) = G(\phi + 2\pi) = G(2\pi - \phi).$$

Under these conditions, the second moment of  $c_k$  can be shown to have the form

$$E\{|c_k|^2\} = \frac{2\beta}{\beta^2 + k^2} (1 - e^{-\beta\pi} \cos k\pi). \quad (A7)$$

For  $k \neq 0$ , these second moment terms become the variances of the cepstral quefrency components since the expected value of each component is zero for  $k \neq 0$ .

#### REFERENCES

- [1] K. Shikano and M. Sugiyama, "Evaluation of LPC spectral matching measures for spoken word recognition," *Trans. IECE*, vol. J 65-D, no. 5, pp. 535-544, May 1982.
- [2] A. H. Gray, Jr. and J. D. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 380-391, Oct. 1976.
- [3] N. Nocerino, F. K. Soong, L. R. Rabiner, and D. H. Klatt, "Comparative study of several distortion measures for speech recognition," in *ICASSP-85 Proc.*, Tampa, FL, Mar. 1985, pp. 25-28.
- [4] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. New York: Springer-Verlag, 1976.
- [5] Y. Tohkura, "A weighted cepstral distance measure for speech recognition," in *ICASSP-86 Proc.*, Tokyo, Japan, Apr. 1986.
- [6] K. K. Paliwal, "On the performance of the quefrency-weighted cepstral coefficients in vowel recognition," *Speech Commun.*, pp. 151-154, May 1982.
- [7] L. R. Rabiner and F. K. Soong, "Single frame vowel recognition using vector quantization with several distance measures," *AT&T Tech. J.*, vol. 64, no. 10, Dec. 1985.
- [8] B. H. Juang, L. R. Rabiner, S. E. Levinson, and M. M. Sondhi, "Recent developments in the applications of hidden Markov models to speaker-independent isolated word recognition," in *ICASSP-85 Proc.*, Tampa, FL, Mar. 1985, pp. 9-12.
- [9] 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.
- [10] B. A. Dautrich, L. R. Rabiner, and T. B. Martin, "On the effects of varying filter bank parameters on isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, pp. 793-806, Aug. 1983.



**Biing-Hwang Juang** (S'79-M'80-S'80-M'81) was born in December 1951. He received the B.Sc. degree in electrical engineering from the National Taiwan University, Taipei, in 1973, and the M.Sc. and Ph.D. degrees in electrical and computer engineering from the University of California, Santa Barbara, in 1979 and 1981, respectively.

In 1978 he joined the Speech Communications Research Laboratory, Santa Barbara, and was involved in research work on vocal tract modeling. In 1979 he became affiliated with Signal Technology, Inc., Santa Barbara, where his research work was in the areas of speech coding and speech interference suppression. Since 1982 he has been with AT&T Bell Laboratories, Murray Hill, NJ. His current research interests include speech recognition, coding, and stochastic processes.

Dr. Juang is a member of the IEEE DSP Technical Committee of the Acoustics, Speech, and Signal Processing Society, and an Associate Editor of the IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING.





**Lawrence R. Rabiner** (S'62-M'67-SM'75-F'75) was born in Brooklyn, NY, on September 28, 1943. He received the S.B. and S.M. degrees simultaneously in June 1964, and the Ph.D. degree in electrical engineering in June 1967, all from the Massachusetts Institute of Technology, Cambridge.

From 1962 through 1964 he participated in the Cooperative Plan in Electrical Engineering at Bell Laboratories, Whippany, and Murray Hill, NJ. He worked on digital circuitry, military communications problems, and problems in binaural hearing. Presently he is engaged in research on speech recognition and digital signal processing techniques at Bell Laboratories, Murray Hill. He is coauthor of the books *Theory and Application of Digital Signal Processing* (Englewood Cliffs, NJ: Prentice-Hall, 1975), *Digital Processing of Speech Signals* (Englewood Cliffs, NJ: Prentice-Hall, 1978), and *Multirate Digital Signal Processing* (Englewood Cliffs, NJ: Prentice-Hall, 1983).

Dr. Rabiner is a member of Eta Kappa Nu, Sigma Xi, Tau Beta Pi, the National Academy of Engineering, and is a Fellow of the Acoustical Society of America.



**Jay G. Wilpon** (M'84) was born in Newark, NJ, on February 28, 1955. He received the B.S. and A.B. degrees (cum laude) in mathematics and economics, respectively, from Lafayette College, Easton, PA, in 1977. He obtained the M.S. degree in electrical engineering/computer science from Stevens Institute of Technology, Hoboken, NJ, in 1982.

Since June 1977 he has been with the Speech Research Department at AT&T Bell Laboratories, Murray Hill, NJ, where he is a member of the Technical Staff. He has been engaged in speech communications research and is presently concentrating on problems in isolated and connected word speech recognition. He has published extensively in this field and has been awarded several patents. His current interests lie in training procedures for both speaker-dependent and speaker-independent recognition systems, speech detection algorithms, and determining the viability of implementing speech recognition systems for general usage.

Mr. Wilpon received the IEEE Acoustics, Speech, and Signal Processing Society's Paper Award, in 1987, for his work on clustering algorithms for use in automatic speech recognition systems.