On the Use of Bandpass Liftering in Speech Recognition

B. H. Juang, L. R. Rabiner, J. G. Wilpon

AT&T Bell Laboratories
Murray Hill, New Jersey 07974

Abstract. In this paper, we extend the interpretation of distortion measures, based upon the observation that measurements of speech spectral envelopes (as normally obtained from analysis procedures) are prone to statistical variations due to window position fluctuations, excitation interference, measurement noise, etc. and may possess spurious characteristics because of analysis model constraints. We have found that these undesirable spectral measurement variations can be controlled (i.e. reduced in the level of variation) through proper cepstral processing and that a statistical model can be established to predict the variances of the cepstral coefficient measurements. The findings lead to the use of a bandpass "liftering" process aimed at reducing the variability of the statistical components of spectral measurements. We have applied this liftering process to various speech recognition problems; in particular, vowel recognition and isolated word recognition. With the liftering process, we have been able to achieve an average digit error rate of 1%, which is about half of the previously reported best results, with dynamic time warping in a speaker-independent isolated digit test.

1. Introduction

Speech recognition tasks involve such necessary steps as analysis, similarity calculation, time normalization and decision logic. 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 the reference patterns. The choice of dissimilarity or distortion measure is extremely important since the final recognition decision is based upon the calculated distances. Accordingly, extensive comparative studies have been conducted in order to find a good distortion measure [1-3] for best recognition accuracy.

There exist 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 between the unknown (test) and the reference patterns, and insensitive to those 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 features in the spoken input. The problem of defining a good distortion measure is then equivalent to 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, to be followed by a simple distortion calculation step in the automatic speech recognition task chain. There are two main reasons why this 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 measure that works with 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.

2. Spectral Variability and A Statistical Model

Most speech recognition systems use some type of spectral analysis to analyze the speech input waveform. Two main types of spectral analysis methods are frequently employed: filter bank and linear prediction.

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 tone noise contamination or spectral dips in the transfer function etc., linearly affect the filter bank spectral vectors and are relatively easy to deal with in spectral comparisons. On the other hand, spectral measurements from the output of filter banks, are sensitive to excitational variations such as fundamental frequency changes. Since these variations are inevitable in natural speech, they become the main factor that makes spectral measurements and comparisons unreliable, unless the bandpass filters are somehow adapted to handle these effects.

As is well known, linear prediction analysis uses an all-pole spectrum to model the short time speech segment. One of the main advantages of the linear prediction analysis method is that the resulting linear prediction polynomial coefficients are largely insensitive to the intrinsic variabilities in the speech signal due to source variation. For this reason we assume that we are using 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 constraint, although leads to an effective resolution of the source-tract interaction, creates other spurious spectral components that are not very desirable in speech recognition applications. The fact that a distortion measure consistent with the linear prediction analysis, i.e. the likelihood ratio measure, exists does not remedy this problem. These spurious spectral components can best be modelled and their effects described in the quefrency domain.

2.1 A Statistical Analysis

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

CH2243-4/86/0000-0765 $1.00 © 1986 IEEE

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

If the expected value of $F(\omega)$ is a constant A, then the expected value of the complex cepstrum is zero except for $k = 0$, where it is the value A. The second moment of c_k is

$$E\{c_k c_k^*\} = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} E\{F(\omega_1)F^*(\omega_2)\} e^{j\omega_1 k} e^{-j\omega_2 k} \frac{d\omega_1}{2\pi}\frac{d\omega_2}{2\pi}.$$

$$= \frac{1}{2\pi}\int_{-2\pi}^{2\pi} (1 - \frac{|\phi|}{2\pi})\, G(\phi) e^{jk\phi} d\phi \qquad (2)$$

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 δ is the Kronecker delta, and B is a constant, the second moment is $B/2\pi$. This is the case when the spectral components of the signal at frequencies ω_1 and ω_2 are uncorrelated. A more realistic model of the correlation function for the spectral components, G, is of the periodic form of $G(\phi) = e^{-\beta|\phi|}$, $-\pi < \phi < \pi$, which leads to

$$E\{|c_k|^2\} = \frac{2\beta}{\beta^2 + k^2}(1 - e^{-\beta\pi}\cos k\pi) \qquad (3)$$

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 $k \neq 0$.

Although the above analysis is for a general, unconstrained spectrum, it, in fact, applies to the case of all-pole model spectra obtained from linear prediction analysis. Denoting the predictor polynomial by $A(z) = 1 + a_1 z^{-1} + \cdots + a_p z^p$, we have the well-known recursion formula for the LPC cepstral coefficients [4]

$$-kc_k - ka_k = \sum_{n=1}^{k-1} (k-n) c_{k-n} a_n \quad \text{for } k > 0 \qquad (4)$$

The measured variance of each of the first 8 LPC cepstral coefficients from a collection of more than 10,000 frames of speech signals is found to be well approximated by the above statistical analysis model for an appropriate value of β [5].

2.2 Variability of LPC Cepstrum Components

The above analysis provides a means of investigating the variability of different components of the LPC cepstrum. Clearly, less correlation between log spectral components results in higher variances at high quefrencies. When the spectral components are totally uncorrelated, the variances become a constant. It has been observed that the cepstral coefficients computed from the LPC spectrum have relatively large variances at high quefrencies.

To verify the above analysis, a Gaussian i.i.d. noise signal was used as the excitation for a fixed 8th order all-pole filter with reflection coefficients typical of those of 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 using the recursion formula (4). It is found, as shown in Fig. 1, that the variances for the higher quefrency terms are relatively large compared to those predicted from the general statistical model of the previous section. Since the signal was generated from a fixed 8th order all-pole filter, one would expect high correlation between spectral components if the spectral measurements were made with an 8th order analysis model. This suggests that higher quefrency terms are the inherent artifact of the LPC spectrum and may not be desirable in the spectral similarity comparisons.

Low quefrency cepstral terms do not have high discrimination power either. The variability of these low quefrency terms is primarily due to type of transmission, speaker characteristics, and vocal efforts, etc. These variabilities diminish the discriminating capability of the corresponding cepstral terms.

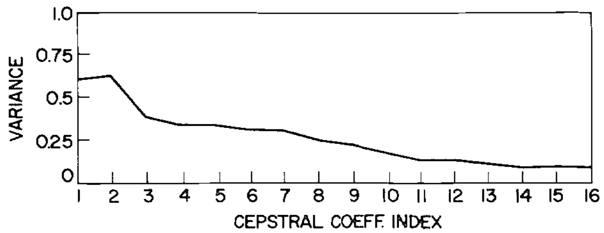

Fig. 1 Variances of the first 16 LPC cepstral coefficients measured from a signal, obtained by driving an 8th order fixed all-pole filter with a Gaussian i.i.d. sequence.

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 pre-emphasis. When spectral notches or zeroes 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 those undesirable variations present in the LPC cepstral coefficients.

3. Liftering Procedure for Speech Recognition

The liftering procedure we propose here is very straightforward. It is simply windowing in the cepstral (quefrency) domain. Fig. 2 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 (4), up to the desired number of terms. A window $w(k)$ is then multiplicatively applied to the cepstral vector. The resultant windowed cepstral vector is used in the recognizer, with a simple Euclidean distance as the distortion or dissimilarity measure.

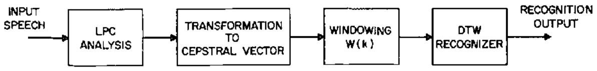

Fig. 2 A liftering procedure.

The effect of this liftering process can be visualized by inverse transforming the windowed cepstral vector back to the log spectrum. Fig. 3 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 $k = 1,2,\ldots,L$ 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 8dB/octave, as shown in the figure, is effectively removed. 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. 4a 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 distortion sensitivity are clearly seen. A liftering window of the form $w(k) = 1 + 6\sin(\pi k/12)$

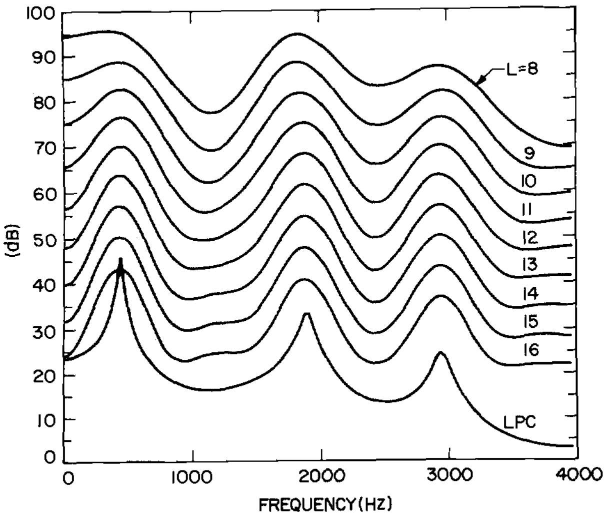

Fig. 3 Effects of liftering on LPC log spectra.

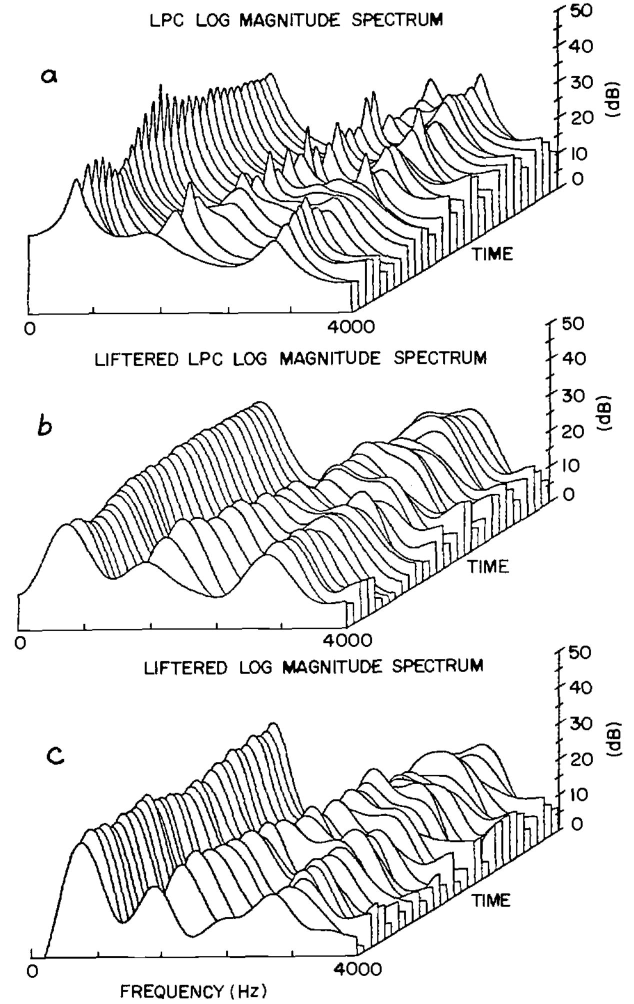

Fig. 4 a) A consecutive sequence of log LPC spectra; b) the result of applying a liftering window to the LPC spectral sequence; c) the corresponding spectral sequence obtained by direct cepstral smoothing on the signal without the intermediate LPC modelling stage.

for $k = 1,2,\ldots,12$ and $w(k) = 0$ otherwise, is then applied, and the corresponding smoothed spectral sequence is plotted in Fig. 4b. 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 and the result is plotted in Fig. 4c. 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, the goal of obtaining reliable spectral measurements, i.e. with low variability, has essentially been demonstrated.

4. 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.

4.1 Choice of Liftering Window

We considered only the following three types of liftering windows: Type 1) $w_1(k) = 1$, for $k = 1,2,\ldots,L$, $= 0$, otherwise; Type 2) $w_2(k) = 1 + h\cdot(k-1)/(L-1)$, for $k = 1,2,\ldots,L$, $= 0$, otherwise; Type 3) $w_3(k) = 1 + h\sin(k\pi/L)$, for $k = 1,2,\ldots,L$, $= 0$, otherwise. These liftering windows were used to recognize a particular isolated digit set, consisting of a total of 1000 utterances (100 utterances for each digit). The linear prediction analysis was 8th order.

The first window is rectangular. For $L = 8$ and 12, the number of misrecognized digits was 38 and 35, respectively out of 1000 trials. The second 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 1. 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 third 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 half of the window length. We then investigated this particular form of liftering window: $w(k) = 1 + 0.5\,L\sin(\pi k/L)$, $1 \leqslant k \leqslant L$. 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

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

TABLE I. Total number of recognition errors (out of 1000 trials) with the triangular lifter

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

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

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

TABLE III. Total number of recognition errors with the $1 + 0.5\, L \sin(\pi k / L)$ lifter

suggest that a liftering window of the form

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

is a good choice for recognition experiments.

4.2 Single Frame Vowel Recognition

The data base 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. Half the vowel frames were used as a training sequence to design vector quantization (VQ) codebooks with 1, 2, 4, and 8 vectors per vowel [6]. The other half of the vowel frames were used as an independent test set. Seven talkers (4 male, 3 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 Ref. [6]. The key result was that, on average, the cepstral lifter provided the best recognition performance for all size codebooks that were tested.

4.3 Speaker Independent Isolated Digit Recognition

The data base 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 data base has been studied previously, and a more detailed description can be found in [7].

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%, 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 Ref. [7] with 12 templates per digit. Finally, the effect of going from 12 templates per digit to a single template per digit only increased the error rate by 2.68%; in fact, the error rate for 3 templates per digit is comparable to the error rate with 12 templates per digit reported earlier [7] 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% higher than that obtained previously with 12 templates per digit [7], and the computation rate is reduced by a factor of 12.

	DATA SETS				Errors	
# of Templates per digit	DAT-1*	DAT-2	DAT-3	DAT-4	Total	%
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

* Training Set

TABLE IV. Total number of recognition errors using the liftering window in a DTW recognizer for speaker independent isolated digit recognition

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 to a more reliable one, it can be used in other recognition schemes, such as the hidden Markov model [7] as well. Further improvements may still be possible when other parameters such as the energy term or a durational model [7] are incorporated into the scheme.

5. 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.

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. on Acoustics, Speech, and 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," *ICASSP-85 Proceedings,* pp. 25-28, Tampa, Fl., March 1985.

[4] J. D. Markel and A. H. Gray, Jr. *Linear Prediction of Speech,* Springer-Verlag, New York, 1976.

[5] Y. Tohkura, "A Weighted Cepstral Distance Measure for Speech Recognition," *ICASSP-86 Proceedings,* Tokyo, Japan, Apr. 1985.

[6] L. R. Rabiner and F. K. Soong, "Single Frame Vowel Recognition Using Vector Quantization with Several Distance Measures," *AT&T Technical Journal,* Vol. 64, No. 10, Dec. 1985.

[7] 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," *ICASSP-85 Proceedings,* pp. 9-12, Tampa, Fl., March 1985.