Fifth International Symposium on Signal Processing and its Applications, ISSPA '99, Brisbane, Australia, 22-25 August, 1999
Organised by the Signal Processing Research Centre, QUT, Brisbane, Australia

ON THE USE OF FILTER-BANK ENERGIES AS FEATURES FOR ROBUST SPEECH RECOGNITION

K.K. Paliwal

School of Microelectronic Engineering
Griffith University
Brisbane, QLD 4111, Australia
(e-mail: K.Paliwal@me.gu.edu.au)

ABSTRACT

Though Mel frequency cepstral coefficients (MFCCs) have been very successful in speech recognition, they have the following two problems: 1) They do not have any physical interpretation, and 2) Liftering of cepstral coefficients, found to be highly useful in the earlier dynamic warping-based speech recognition systems, has no effect in the recognition process when used with continuous observation Gaussian density hidden Markov models. In this paper, we propose to use the filter-bank energies (FBEs) as features. The FBEs are physically meaningful quantities and amenable for applying human auditory processing such as masking. We describe procedures to decorrelate and lifter the FBEs and show that the FBEs perform at least as good as (and sometimes even better than) the MFCCs for robust speech recognition.

INTRODUCTION

Mel frequency cepstral coefficients (MFCCs) are perhaps the most widely used features for speech recognition [1, 2]. The MFCCs are computed from the speech signal using the following three steps [2]:

1. Compute the FFT power spectrum of the speech signal.
2. Apply a Mel-spaced filter-bank to the power spectrum to get N energies. (N is typically 20 to 60).
3. Compute discrete cosine transform (DCT) of log filter-bank energies to get MFCCs. (M is typically about 10.) The DCT is applied to filter-bank energies (FBEs) to mainly decorrelate them and get (approximately) uncorrelated MFCCs.

Though MFCCs have been very successful in speech recognition, they have the following two problems: 1) They do not have any physical interpretation, and 2) Liftering of cepstral coefficients, found to be highly useful in the earlier dynamic warping-based speech recognition systems, has no effect in the recognition process when used with continuous observation Gaussian density hidden Markov models (HMMs).

The filter-bank energies (FBEs) have a clear physical interpretation which can be used advantageously for incorporating the properties of human speech perception system (such as masking) into the automatic speech recognition process. A question arises whether one can use the FBEs as features for speech recognition. In fact, the FBEs have been used in 50's, 60's and 70's for speech recognition. However, they have been abandoned in favor of the cepstral coefficients because of the following two problems: 1) Use of 20 to 60 FBEs increases the dimensionality of feature space, and 2) The FBEs are highly correlated. The first problem can solved by restricting the number of FBEs computed from the FFT spectrum to about 10 (i.e., by making N equal to M). The second problem can be solved by filtering the FBE sequence of a given frame so as to remove correlation among them. Another advantage of FBEs is that the phenomenon of cepstral liftering can be easily applied on FBEs, which is effective in the recognition process even with continuous observation Gaussian density HMMs.

In this paper, we investigate the use of the filter-bank energies (FBEs) as features (See [10] for more details). We carry out speech recognition experiments with clean as well as noisy speech and quantify the effect of decorrelation and liftering of FBEs in terms of their speech recognition performance. A number of studies on this topic have been reported earlier by Nadeu and his collaborators [3, 4].

ROLE OF CEPSTRAL LIFTERING IN SPEECH RECOGNITION

In the 70's and early 80's, the dynamic time warping (DTW) based framework was the most popular approach for speech recognition. In this framework, the dissimilarity between the test vector $\mathbf{x}$ and the mean vector $\hat{\mathbf{x}}$ representing a given class is measured in terms of the Euclidean distance measure of the form:

$$d(\mathbf{x};\hat{\mathbf{x}}) \quad = \quad (\mathbf{x}-\hat{\mathbf{x}})^t(\mathbf{x}-\hat{\mathbf{x}}),$$

$$= \sum_{i=1}^{D}(x_i - \hat{x}_i)^2, \qquad (1)$$

where D is the dimensionality of the feature space and the superscript t denotes transpose of a vector (or matrix).

If the cepstral coefficients are used as recognition features, then it has been demonstrated through a number of studies [5, 6, 7, 8, 9] that the performance of the speech recognition system can be improved significantly by liftering the cepstral coefficients. If x_i is the i-th cepstral coefficient, then the corresponding liftered cepstral coefficient is given by

$$y_i = w_i x_i, \qquad (2)$$

where weights $w_i, i = 1, 2, ..., D$, define the lifter. The distance measure in terms of the liftered cepstral coefficients is given by

$$\begin{aligned} d(\mathbf{y};\hat{\mathbf{y}}) &= (\mathbf{y}-\hat{\mathbf{y}})^t(\mathbf{y}-\hat{\mathbf{y}}), \\ &= \sum_{i=1}^{D}(y_i - \hat{y}_i)^2, \\ &= \sum_{i=1}^{D}[w_i(x_i - \hat{x}_i)]^2, \qquad (3) \end{aligned}$$

Various types of lifters are proposed in the literature. Some of the important lifters are listed below.

1. Linear lifter [5]:

$$w_i = i. \qquad (4)$$

2. Statistical lifter [5, 6]:

$$w_i = 1/\hat{\sigma}_i, \qquad (5)$$

where $\hat{\sigma}_i$ is the standard deviation of the i-the cepstral coefficient computed from the training data.

3. Sinusoidal lifter [7]:

$$w_i = 1 + \frac{D}{2}\sin(\frac{\pi i}{D}). \qquad (6)$$

4. Exponential lifter [8]:

$$w_i = i^s \exp(-\frac{i^2}{2\tau^2}), \qquad (7)$$

where s and τ are constants. Their typical values are $s = 1.5$ and $\tau = 5$.

Note that the weights w_i in the linear and statistical lifters increase with quefrency i, thus giving more importance to higher cepstral coefficients in the computation of distance $d(\mathbf{y};\hat{\mathbf{y}})$ (Eq. (3)). In the sinusoidal and exponential lifters, the weights w_i initially increase with i, attain a maximum and then start decreasing with an increase in i. Thus, these lifters provide less weight to both lower and higher cepstral coefficients in distance computation. It has been shown [9, 8, 10] that all the four types of cepstral liftering help in recognizing clean as well as noisy speech better. Howevėr, improvement in recognition performance due to liftering is much better for noisy speech than that for clean speech.

In order to understand why cepstral liftering helps speech recognition, note that the cepstral coefficients are obtained by taking inverse Fourier transform of the log-power spectrum. In cepstral liftering, the cepstral coefficients $\{x_i\}$ are multiplied by the weights $\{w_i\}$. The multiplicative weighting is equivalent to convolution in spectral domain where the the log-power spectral sequence is convolved with a filter impulse response (determined by taking Fourier transform of the weight sequence $\{w_i\}$). It can be shown that the linear and statistical weights lead to high-pass (HP) filter in the spectral domain; while the sinusoidal and exponential weights result in a band-pass (BP) filter. All the four types of liftering give less weight to lower cepstral coefficients. This tends to suppress slow variations in the log-power spectrum, which come mainly from factors such as speaker-to-speaker variability and additive white noise. This is the reason for the cepstral weighting working so well for noisy speech. Since the higher cepstral coefficients provide less discrimination for speech recognition than the lower cepstral coefficients, the sinusoidal and exponential lifters give less weight to higher cepstral coefficients and, hence, provide better recognition performance than the other lifters for clean speech.

So far, we have discussed the role of cepstral liftering on the performance of the DTW-based speech recognition system. Currently, the DTW-based framework is rarely used for speech recognition. Instead, most of the speech recognition systems presently utilize the hidden Markov model (HMM) based framework. It can be easily shown [10] that cepstral liftering, found to be highly useful in the earlier DTW-based speech recognition systems, has no effect in the recognition process when used with continuous observation Gaussian density HMMs. This is a major drawback with the cepstral coefficients. However, note that a statistical lifter (similar to Eq. (5)) is implicit in the continuous Gaussian density HMM formulation. Also, note that when discrete density HMMS with vector quantization and Euclidean distance measure are used for speech recognition, cepstral liftering offers improvements in recognition performance similar to the DTW-based framework.

DECORRELATION OF FILTER BANK ENERGIES

In this paper, our aim is to use FBEs as features for speech recognition. However, the FBEs have the problem that they are correlated. If we want to use them with continuous Gaussian density HMMs with diagonal covariance matrices, we must decorrelate them. This is done as follows:

Let the N FBEs computed for a given frame are given by the sequence, e_n, $n = 0, 1, ..., N-1$. We model the correlation among consecutive FBEs by a linear predictor (LP)

$$P(z) = \sum_{i=1}^{p} a_i z^{-i},$$

where a_i, $i = 1, 2, ..., p$, are the LP coefficients. These can be estimated from the FBE sequence $\{e_n\}$ of the given frame using the covariance method of LP analysis. The decorrelated FBEs are computed by passing the sequence $\{e_n\}$ through the inverse filter $A(z) = 1 - P(z)$. If we are interested in M decorrelated FBE features, then we must have $N = M + p$.

LIFTERING OF FILTER BANK ENERGIES

As discussed earlier, the cepstral liftering is achieved by applying multiplicative weights $\{w_i\}$ to cepstral coefficients. Because of this, liftering of cepstral coefficients has no effect in the recognition process when used with continuous observation Gaussian density HMMs. Since the multiplicative weighting of cepstral coefficients is equivalent to convolution in spectral domain, liftering of FBEs can be used with the continuous Gaussian density HMMs with positive effect on recognition performance.

In order to lifter the FBEs, we design an L-th order finite impulse response (FIR) filter

$$H(z) = \sum_{i=0}^{L} h_i z^{-i},$$

from the lifter weights $\{w_i\}$ using the windowing method. The liftered FBEs are computed by passing the sequence $\{e_n\}$ through the filter $H(z)$. If we are interested in M liftered FBE features, then we must have $N = M + L$.

SPEECH RECOGNITION EXPERIMENTS AND RESULTS

In this paper, we investigate the use of FBEs as recognition features using a speaker independent isolated word recognition system as a test bed. For this, we use the ISOLET spoken letter database from Oregon Graduate Institute [11]. Here, the vocabulary consists of 26 English alphabets (A-Z). From this data base, we use 90 utterances for each word from 90 talkers (45 male and 45 female) for training and 30 utterances for each word from 30 talkers (15 male and 15 female, different from training talkers) for testing. In the original data base, these utterances were digitized at 16 kHz sampling rate. We down-sample the speech utterances from this data base to 8 kHz using a lowpass filter with cutoff frequency of 3.5 kHz. The speech signal is analyzed every 10 ms with a frame width of 30 ms (with Hamming window and preemphasis), and each frame is represented in terms of $M = 10$ features.

The recognition system uses a multi-mixture continuous density HMM framework. We use a 5-state continuous density HMM recognizer with probability density functions approximated by a mixture of 5 multivariate normal distributions with diagonal covariance matrices.

We investigate the decorrelated and liftered FBEs as features for recognizing noisy speech. For this, we train our continuous Gaussian density HMM-based recognizer on clean speech and test it on white noise corrupted speech at different SNRs. Table 2 lists results in terms of recognition accuracy (in %) for the following feature sets:

- Feature set 1: 10 FBEs.
- Feature set 2: 10 decorrelated FBEs (obtained by using first order linear predictor $P(z) = a_1 z^{-1}$, where a_1 is computed from the FBEs of a given frame).
- Feature set 3: 10 decorrelated FBEs (obtained by using second order linear predictor $P(z) = a_1 z^{-1} + a_2 z^{-2}$, where a_1 and a_2 are computed from the FBEs of a given frame).
- Feature set 4: 10 liftered FBEs (obtained by using $H(z) = 1 - 0.5z^{-1}$).
- Feature set 5: 10 liftered FBEs (obtained by using $H(z) = 1 - 0.75z^{-1}$).
- Feature set 6: 10 liftered FBEs (obtained by using $H(z) = 1 - z^{-1}$).
- Feature set 7: 10 liftered FBEs (obtained by using $H(z) = 1 - z^{-2}$).
- Feature set 8: 10 decorrelated and liftered FBEs (decorrelation by $P(z)$ from the feature set 3 and liftering from the feature set 7).

We can make the following observations from this table:

1. Speech recognition performance improves by decorrelating FBEs using the first order predictor. However, the performance goes down when the second order predictor is used for decorrelation.
2. Liftering of FBEs always helps in improving the speech recognition performance. Liftering with the filter $H(z) = 1 - z^{-2}$ provides the best results.

Table 1: Effect of decorrelation and liftering of FBEs on the speech recognition performance.

Feature	SNR in dB				
set	∞	30	25	20	15
Set 1	74.3	71.5	66.2	57.4	42.1
Set 2	74.5	73.5	71.5	67.6	57.7
Set 3	65.3	63.5	61.0	56.7	47.5
Set 4	76.0	75.0	71.2	65.6	55.3
Set 5	75.6	74.7	72.8	68.2	59.5
Set 6	75.8	75.6	73.8	70.1	61.7
Set 7	76.5	75.4	72.6	69.1	62.2
Set 8	69.3	69.8	69.3	66.4	58.7

Table 2: Comparison of the MFCC and the liftered FBE features for speech recognition.

Feature	SNR in dB				
set	∞	30	25	20	15
MFCC:					
c	76.4	74.3	71.4	66.2	55.5
$c + \Delta c$	86.0	83.6	81.8	78.6	71.6
$c + \Delta c + \Delta\Delta c$	89.0	86.5	85.0	82.2	76.6
FBE:					
e	76.5	75.4	72.6	69.1	62.2
$e + \Delta e$	86.6	85.6	84.2	81.3	74.7
$e + \Delta e + \Delta\Delta e$	88.9	87.7	86.5	83.7	77.3

3. When both decorrelation and liftering of FBEs are used, the speech recognition performance goes down.

Thus, we have seen that when we use $H(z) = 1 - z^{-2}$ to filter the FBE sequence $\{e_n\}$ to get the liftered FBE features, we achieve the best results. We will use hereafter in this paper these liftered FBEs as the FBE features.

Since MFCCs are currently used by most of the researchers as recognition features, we compare the liftered FBE features with the MFCC features in terms of their recognition performance. This comparison will help us in putting the FBE features in proper perspective. Since temporal derivatives of the feature-vector sequence has been found to be greatly useful for speech recognition, we include delta and delta-delta parameters of the MFCCs and FBEs in the feature set. Results are shown in Table 3. In this table, symbol c is used to denote the MFCCs, and e the FBEs. It can be observed from this table that the liftered FBE features perform as well as (and sometimes better than) the MFCC features.

CONCLUSIONS

In this paper, we have investigated the use of FBEs as features for speech recognition. We have studied the role of decorrelation and liftering of FBEs and shown that liftering provides best advantage for speech recognition. We have compared the liftered FBE features with the MFCC features for speech recognition and found them to be as good as (and sometimes better than) the MFCC features.

REFERENCES

[1] S. Young, "A review of large-vocabulary continuous-speech recognition", *IEEE Signal Processing Magazine*, pp. 45-57, Sept. 1996.

[2] L. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[3] C. Nadeu, J. Hernando and M. Gorricho, "On the decorrelation of filter-bank energies in speech recognition", *Proc. EUROSPEECH*, pp. 1381-1384, Sept. 1995.

[4] J. Hernando and C. Nadeu, "CDHMM speaker recognition by means of frequency filtering of filter-bank energies", *Proc. EUROSPEECH*, pp. 2363-2366, 1997.

[5] K.K. Paliwal, "On the performance of the quefrency-weighted cepstral coefficients in vowel recognition", *Speech Communication*, pp. 151-154, May 1982.

[6] Y. Tohkura, "A weighted cepstral distance measure for speech recognition", *IEEE Trans. Acoust., Speech and Signal Processing* Vol. ASSP-35, No. 10, pp. 1414-1422, Oct. 1987.

[7] B.H. Juang, L.R. Rabiner and J.G. Wilpon, "On the use of bandpass liftering in speech recognition", *IEEE Trans. Acoust., Speech and Signal Processing* Vol. ASSP-35, No. 7, pp. 947-954, July 1987.

[8] F. Itakura and T. Umezaki, "Distance measure for speech recognition based on the smoothed group delay spectrum", *Proc. ICASSP*, pp. 1257-1260, April 1987.

[9] B.A. Hanson and H. Wakita, "Spectral slope distance measures with linear prediction analysis for word recognition in noise", *IEEE Trans. Acoust., Speech and Signal Processing* Vol. ASSP-35, No. 7, pp. 968-973, July 1987.

[10] K.K. Paliwal, "Decorrelated and liftered filter-bank energies for robust speech recognition", *Proc. EUROSPEECH*, 1999.

[11] R. Cole, Y. Muthusamy and M. Fanty, "The ISOLET spoken letter database", Technical Report No. CSE 90-004, Dept. of Computer Science and Engineering, Oregon Institute of Science and Technology, Beaverton, OR, USA, Mar. 1990.