IMPROVED MFCC FEATURE EXTRACTION BY PCA-OPTIMIZED FILTER-BANK FOR SPEECH RECOGNITION

Shang-Ming Lee, Shi-Hau Fang, Jeih-weih Hung and Lin-Shan Lee*

Graduate Institute of Comm. Eng., National Taiwan University.
Taipei, Taiwan, Republic of China
jwhung@iis.sinica.edu.tw

ABSTRACT

Although Mel-frequency Cepstral Coefficients (MFCC) have been proven to perform very well under most conditions, some limited efforts have been made in optimizing the shape of the filters in the filter-bank in the conventional MFCC approach. This paper presents a new feature extraction approach that designs the shapes of the filters in the filter-bank. In this new approach the filter-bank coefficients are data-driven obtained by applying the principal component analysis (PCA) on the FFT spectrum of the training data. The experimental results show that this method is robust under noisy environment and is well additive with other noise-handling techniques.

1. INTRODUCTION

Feature extraction is a very important key element in speech recognition since it is the first step of the whole recognition process and it produces the parameters on which the recognition algorithm is based. If the feature parameters used are not well extracted, the recognition performance is naturally limited. Mel-frequency cepstral coefficients (MFCC) are the most widely used feature parameters currently, while linear predictive cepstral coefficients (LPCC) were also used in some systems. Usually MFCC offers a performance better than what LPCC does, especially in noisy environment, but it is generally believed that it is highly desired to have feature parameters better than MFCC. Substantial efforts have been made in this area, and quite many new approaches to produce feature parameters with better recognition performance than MFCC at least under some test environments have been proposed [1] [2] [3] [4]. Techniques like Linear Discriminant Analysis (LDA) [5] and Heteroscedastic Linear Discriminant Analysis (HLDA) [6] have been proposed to improve the discriminating capabilities of the original features. Although different criteria were used, these methods tried to search for a transformation matrix by which the original feature representation can be reduced in its dimension while the recognition performance can be improved or at least maintained.

In the original MFCC feature extraction process there are in fact two steps also related to dimension reduction. One is the Mel-scaled filter-bank processing. In each frequency band, the frequency components are weighted according to the filter frequency response and then accumulated to a value representing the total energy of that band. The other step of dimension reduction is performed in the transformation from the log-spectral domain to the cepstral domain, where the size of the resulted cepstral features is often less than that in the log-spectral domain. Both of these two steps may probably result in some information loss from the original signal, although it is widely accepted that such steps are helpful in extracting the useful components in speech signals for recognition. Since the Mel-scaled filter-bank plays a very important role in feature extraction process, it is re-considered here in this paper. Conventionally, triangular filters are used in the filter-bank in the MFCC derivation process [7], which seems to be a reasonably good but relatively rough solution. However, it seems that not too many efforts have been reported in trying to optimize the shape of each filter in the filter-bank. In fact, the shape of the above filter also has to do with the signal-to-noise ratio of the filter output. For example, if the noise added to the clean signal is white, then different frequency components have different signal-to-noise ratios (SNRs) since the noise components are roughly the same for all frequencies while the speech components are not. The filter shape determines the weights on different signal components in the same frequency band, and thus determines the output SNR.

In the past years, Ensemble Interval Histogram (EIH) model [7][8] and Auditory Spectrum Based Features (ASBF) [4] are two examples that applied auditory based spectral analysis models and also took the shape of the filters in the filter-bank into consideration. EIH model used a model of the cochlea and the hair cell transduction. It consisted of a filter-bank that models the frequency selectivity at various points along a simulated basilar membrane, and a nonlinear processor for converting the filter-bank outputs to neural firing patterns along a simulated auditory nerve. ASBF was based on the cochlea model of the human auditory systems as well, and was able to track the formants. Both approaches had different shapes for different filters in the filter-bank. Although these approaches offered good performance under noisy environment, they significantly changed the normal process of feature extraction, and thus the conventional feature enhancement techniques developed on the MFCC may not be directly applicable on this kind of features. Moreover, they are quite complicated and require large number of decisions and computations in the feature extraction process. On the other hand, a series of work of filter-bank design based on a data-driven approach in order to obtain a more discriminative representation of speech features has been presented [9]. In this approach, the bandwidths, shapes and positions of the filters in the filter-bank can all be tuned and optimized according to the criterion of Minimum Classification Error (MCE). In this way the filter-bank is well matched to the specific task and the relevant speech corpora since it is obtained *data-driven*. Also, this approach makes the front-end feature extractor well matched to the back-end classifier since the procedure of MCE training is based on minimizing the error for the overall recognizer, which includes both the front-end feature

extractor and the back-end classifier. However, the MCE training requires relatively high computation complexity with many parameters to be decided all together for a specific task. For example, the overall filter-bank needs to be re-trained if the back-end classifier structure is slightly modified for a different task with a different speech corpus.

In this paper, we proposed that the *shape* alone of each filter in the Mel-scale filter-bank in MFCC feature extraction can be derived *data-driven* by applying the criterion of principal component analysis (PCA). This filter-bank is easy to be obtained for a given task and corpus, the improved MFCC obtained in this way can be well compatible to many feature enhancement techniques previously developed, and the feature extraction can be de-coupled from the many parameters in the back-end classifier which can be different for different tasks. It is shown in this paper the new features obtained using this PCA-derived filter-bank give comparable performance for clean speech and better performance under noisy environment, as compared with the conventional MFCC features. Besides, it's also shown that it performs very well when combined with some other basic de-noise techniques like Spectral Subtraction.

The remainder of this paper is organized as follows. Section 2 describes the proposed approach. Sections 3 and 4 are the experiments and discussions. Finally, a brief concluding remark is given in Section 5.

2. THE PROPOSED APPROACH OF PCA-OPTIMIZED FILTER-BANK

The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set that consists of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set [10]. To state PCA briefly, if $\boldsymbol{x}$ is an Nx1 random vector, the objective is then to find a set of Nx1 orthonormal vectors $\{\mathbf{w}_i | \ 1 \le i \le k, \ k \le N\}$ such that the inner product of each $\mathbf{w}_i$ and $\boldsymbol{x}$,

$$y_i = \mathbf{w}_i^T \boldsymbol{x} \tag{1}$$

has the maximum variance, where y_i is a scalar value. The above set of vectors $\{\mathbf{w}_i\}$ is in fact the eigenvectors of the covariance matrix for $\boldsymbol{x}$ corresponding to the largest k eigenvalues. The above idea of PCA can be applied in the filter shape optimization problem considered here. Each filter in the filter-bank can be viewed as a process of dimensionality reduction, where the signal components within that frequency band are weighted and then combined into a single value, whose variation is to be maximized. The detailed procedure is stated as follows.

Let $\{x_k(n), \ n = 1,2,\cdots,m_k\}$ be the random variables representing the m_k signal components belonging to the *k-th* frequency band to be handled by the *k-th* filter in the filter-bank, where m_k is the total number of components in that band, and let $\boldsymbol{x}_k$ be the vector representation for these random components. That is,

$$\boldsymbol{x}_k = [\, x_k(1), x_k(2), \ldots\ldots, x_k(m_k)]^T \tag{2}$$

For each training signal of the training database, its spectral components corresponding to the *k-th* filter of the filter-bank can be extracted, represented as a vector and then this vector can be viewed as a sample of the random vector $\boldsymbol{x}_k$ in equation (2). By collecting these sample vectors for the random vector $\boldsymbol{x}_k$, the covariance matrix $cov(\boldsymbol{x}_k)$ can be calculated and then diagonalized into $\mathbf{D}_k$,

$$cov(\boldsymbol{x}_k) = \boldsymbol{F}_k^{-1} \mathbf{D}_k \boldsymbol{F}_k \tag{3}$$

where $\boldsymbol{F}_k$ is the matrix whose column vectors are the eigenvectors of $cov(\boldsymbol{x}_k)$. The coefficients of the *k-th* filter are then simply the components of the column vector $\mathbf{w}_k$ of $\boldsymbol{F}_k$ corresponding to the largest diagonal element (or eigenvalue) of $\mathbf{D}_k$. This process is shown in figure 1.

With the filter obtained above, apparently the variance of the filter output $y_k = \mathbf{w}_k^T \boldsymbol{x}_k$ can be maximized. Furthermore, if the additive noise within that frequency band is assumed flat (white), the ratio of the variance for the signal to that for the noise (say, signal-to-noise variance ratio, which can be viewed as a different form of the signal-to-noise ratio (SNR)) can be maximized as well, which is also a desired property. This can be briefly shown in the following.

Let $\boldsymbol{x}_k = \boldsymbol{s}_k + \boldsymbol{n}_k$, where $\boldsymbol{s}_k$ and $\boldsymbol{n}_k$ are the vector representations of the clean signal and noise power spectral components within the *k-th* band, respectively. By assuming $\boldsymbol{s}_k$ and $\boldsymbol{n}_k$ are uncorrelated, the "signal-to-noise variance ratio" for the output of the *k-th* filter $\mathbf{w}_k$ is,

$$\frac{\mathrm{var}\left(\mathbf{w}_k^t \mathbf{s}_k\right)}{\mathrm{var}\left(\mathbf{w}_k^t \mathbf{n}_k\right)} = \frac{\mathrm{var}\left(\mathbf{w}_k^t \mathbf{x}_k\right) - \mathrm{var}\left(\mathbf{w}_k^t \mathbf{n}_k\right)}{\mathrm{var}\left(\mathbf{w}_k^t \mathbf{n}_k\right)} = \frac{\mathrm{var}\left(\mathbf{w}_k^t \mathbf{x}_k\right)}{\mathrm{var}\left(\mathbf{w}_k^t \mathbf{n}_k\right)} - 1 \tag{4}$$

If the noise spectrum within that frequency band is assumed flat (white), that is, the covariance of the noise vector $\boldsymbol{n}_k$ can be assumed as $\sigma_{n_k}^2 \mathbf{I}$, where $\sigma_{n_k}^2$ is the variance for each component of $\boldsymbol{n}_k$, then in equation (4) the term in the denominator is

$$\mathrm{var}\left(\mathbf{w}_k^T \mathbf{n}_k\right) = \mathbf{w}_k^T \, \mathrm{cov}\left(\mathbf{n}_k\right) \mathbf{w}_k = \mathbf{w}_k^T \left(\sigma_{n_k}^2 \mathbf{I}\right) \mathbf{w}_k = \sigma_{n_k}^2 \mathbf{w}_k^T \mathbf{w}_k = \sigma_{n_k}^2 . \tag{5}$$

Since in equation (4) the term $\mathrm{var}(\mathbf{w}_k^T \boldsymbol{n}_k)$ is a fixed value $\sigma_{n_k}^2$ while the term $\mathrm{var}(\mathbf{w}_k^T \boldsymbol{x}_k)$ is maximized respect to $\mathbf{w}_k$, the "signal-to-noise variance ratio" for the output of the *k-th* filter $\mathbf{w}_k$ in equation (4) is also maximized accordingly. It should be pointed out that what is really desired here is to maximize both the signal-to-noise variance ratio for the filter output and the variation of this output among the phone classes, but not the variation among speakers, channels or different noise types because the desired discrimination is in phone classes, not in speakers, channels or noise. This will be further discussed later on.

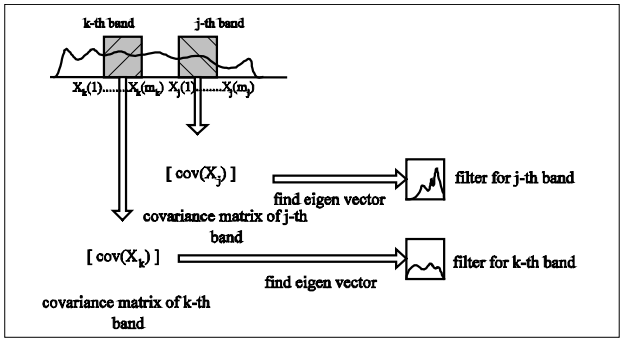

Figure 1: The process of finding PCA-optimized filter bank coefficients

3. EXPERIMENTAL SETUP

The major speech database used in the experiments was the NUM-100A digit database provided by the Association for Computational Linguistics and Chinese Language Processing at Taipei [11]. This database includes 8000 Mandarin digit strings produced by 50 males and 50 females. The speech signal was

recorded under normal laboratory environment at 8 kHz sampling rate and encoded with 16bit linear PCM. The database consisted of 1000 2, 3, 4, 5, 6, 7-digit strings and 2000 single digit utterances. This database was split into 7520 training digit strings and 480 testing utterances. Another database tested was MAT-2000 with the same encoding condition as NUM-100A, and collected through telephone networks in Taiwan. There are a total of 12149 training utterances and 500 test utterances in this database. The first set of experiments was performed on NUM-100A, in which a zero-mean white Gaussian noise was added to the test utterances at each specified signal to noise ratio. A 32ms Hamming window shifted with 10ms steps and a pre-emphasis factor of 0.97 were used. Then cepstral coefficients were generated through a filter-bank of 23 filters and IDCT, and the first 12 coefficients plus the log energy were chosen as the feature parameters. The conventionally used triangular filters in the filter-bank were applied for the baseline experiments for the further comparison. On the other hand, the modified filter-bank as shown in Figure 2 is generated using the training utterances of NUM-100A database by the PCA technique as described previously. The dimension of the baseline MFCC feature vector is 39, which include 13 coefficients as mentioned above, its 13 derivatives and 13 accelerations. We use a total of 16 3-state left-to-right sub-word HMM models each with 1, 2, 4, 8, 16 mixtures per state. The second set of experiments performed on MAT-2000 database will be discussed below. From the Figure 2, we can see that the shapes of the filters in the filter-bank are not only quite different from one another, they are also not always triangular.

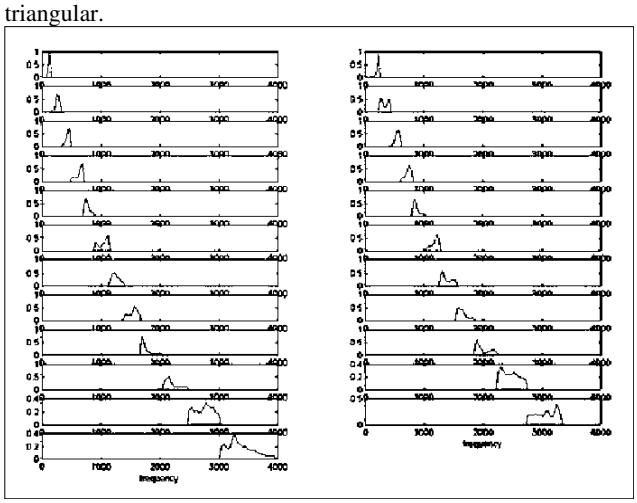

Figure 2: The shape of 23 filters in the filter-bank

4. EXPERIMENTAL RESULTS

Table 1 lists the recognition results for the first set of experiments on NUM-100A database under various noisy conditions with 4 mixtures per state. Each column is for a different SNR condition, and each row is the results for a processing approach. The first two rows (1)(2) compare the proposed approach with the MFCC baseline. It can be found from these two rows that for clean speech, the proposed approach was exactly the same, without any degradation, as compared to the MFCC baseline. However, in noisy speech the proposed approach clearly outperformed the MFCC baseline and the improvements with respect to MFCC baseline became more significant at worse noisy conditions. The next two rows (3)(4) compare the results when the spectral subtraction (SS) was added to the MFCC baseline and the proposed approach. It can be found that in this case the proposed approach was better than MFCC even for the clean speech case, and again offered more improvements at worse noisy conditions. In the next two rows (5)(6) the cepstral mean subtraction (CMS) was applied. In this case the proposed approach was slightly worse than MFCC, although very close. In the last two rows (7)(8) both cepstral mean subtraction (CMS) and spectral subtraction (SS) were applied. Again the proposed approach is better, although only slightly. When comparing the results in each column of Table 1, it can be found that the proposed approach plus the spectral subtraction (SS) provided the best results for clean speech, 30dB and 20dB, while for 10dB the proposed approach alone already gave the best result, although the addition of spectral subtraction is only slightly worse. Apparently the proposed approach is quite additive with the spectral subtraction. One explanation for it is that the spectral subtraction performs well when the noise is not too serious, while it can't help too much for serious noisy conditions. The proposed approach, on the other hand, did well for serious noise conditions too, and thus the two approaches thus complement each other. The cepstral mean subtraction (CMS) approach, on the other hand, didn't seem to work very well here in this set of experiments, probably because it is primarily for channel bias removal but the NUM-100A database didn't include channel effect yet. Also, the proposed approach performed very well especially at worse noisy conditions and equally well for clean speech. This is in good agreement with the discussions made previously in equations (4)(5), i.e., the signal-to-noise variance ratio has been maximized for each filter output. The results in Table 1 are for 4 mixtures per state. In fact, the similar trend can be observed in all different numbers of mixtures from 1 to 16. Figures 3 and 4 show the accuracy comparison between the proposed approach and the MFCC baseline, used alone and together with spectral subtraction respectively, i.e., the situation of rows (1)(2)(3)(4) in Table 1, but with different numbers of mixtures. The solid lines are for the proposed approach and the dotted lines the MFCC baseline. These figures verify the above statement for different mixture numbers.

SNR	clean	30dB	20dB	10dB
MFCC baseline (1)	96.20	88.08	73.86	34.20
PCA approach (2)	96.20	89.69	78.12	**46.23**
MFCC baseline, SS (3)	96.26	91.25	79.04	34.83
PCA approach, SS (4)	**96.43**	**92.29**	**81.52**	45.08
MFCC baseline, CMS (5)	95.22	89.35	72.54	35.92
PCA approach, CMS (6)	94.93	90.04	72.71	33.79
MFCC baseline,CMS,SS (7)	95.16	91.54	78.99	39.32
PCA approach,CMS,SS (8)	95.45	92.06	79.50	39.84

Table 1: Recognition Accuracy under various noisy conditions

To check if the proposed features is really more robust with additive noise, the average Euclidean distance was calculated between the clean feature vector $\mathbf{x}(n)$ and noisy speech feature vector $\tilde{\mathbf{x}}(n)$:

$$D = \frac{1}{N}\sum_{n}\left(\tilde{\mathbf{x}}(n)-\mathbf{x}(n)\right)^{T}\left(\tilde{\mathbf{x}}(n)-\mathbf{x}(n)\right), \qquad (6)$$

where the average is performed over all data frames n tested in the above experiments. The average distance results are listed in

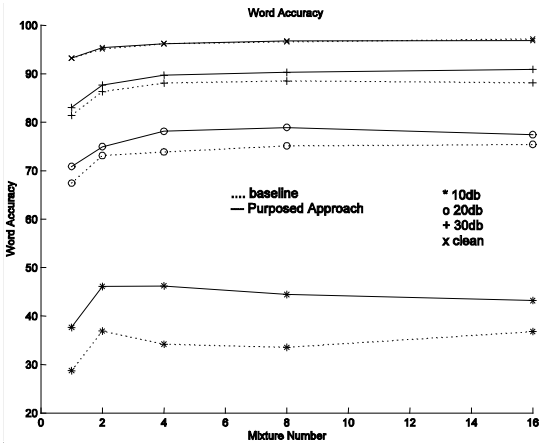

Figure 3: Proposed approach v.s. MFCC baseline for different numbers of mixtures

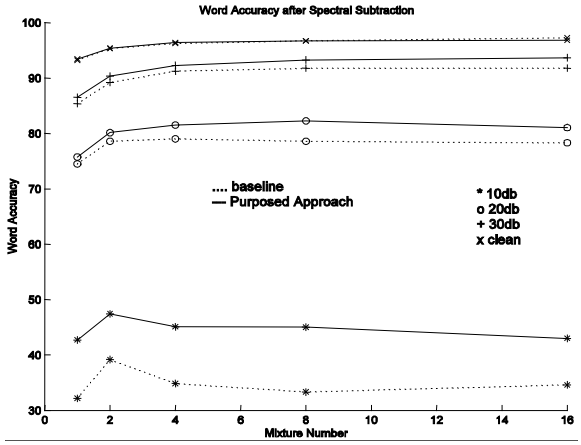

Figure 4: Proposed approach v.s. MFCC baseline under Spectral Subtraction for different numbers of mixtures

Table 2. It is clear from the table that the distance for the proposed approach is always smaller than the corresponding MFCC baseline, regardless of used alone or with spectral subtraction. It can also be found that the average distance is significant larger for worse noisy conditions, but the distance reduction by the proposed approach is also slightly more for those cases. Also, the spectral subtraction actually reasonably reduced the average distance in all cases. The next set of experiments was conducted on the MAT-2000 database, in which all speech utterances were collected via telephones, thus including the channel effect. Two tests were performed, the first used the filter-bank coefficients obtained with the NUM-100A database used previously, which were for clean speech, and the second used the MAT2000 database training data to generate the filter-bank coefficients. The results are shown in Table 3. The baseline test used the conventional MFCC features. In both cases the proposed approach gave slightly better results. The improvements were not significant, since there was essentially no additive noise here. However, the results of the first test were slightly better than the second. This is probably because in the second test PCA was performed on the speech data collected from many different speakers via many different telephone channels, thus the channel variation was mixed up with the variation among phone classes and jointly maximized when defining the filter-bank shape. Thus the channel variation somehow offset the discrimination among phone classes. Moreover, since in the second row of Table 3 the filter-bank is obtained by the clean speech database, while the test data is telephone speech, there is a mismatch between them and thus the improvements is limited. It is expected that a proper combination of the PCA approach and some other approaches such as temporal filtering may offer better results.

SNR	30dB	20dB	10dB
MFCC baseline	1.4032	2.7495	4.6993
PCA approach	1.3785	2.7174	4.6635
MFCC baseline+SS	1.2321	2.4751	4.3645
PCA approach +SS	1.1970	2.4255	4.3078

Table 2 : Average distance between the noisy speech and the clean speech

MFCC baseline	72.36
PCA approach with filter-bank by NUM-100A (without channel effects)	73.23
PCA approach with filter-bank by MAT-2000 (with channel effects)	72.53

Table 3 : Recognition accuracy on MAT2000

5. CONCLUSION

In this paper, the conventional MFCC feature is improved by PCA-optimized filter-bank. The results in the experiments show that the proposed features are robust to additive noise for speech recognition, provide exactly the same performance for clean speech, and are quite additive with some robustness techniques such as the spectral subtraction. This is because the PCA-optimized filter-bank maximizes both the signal-to-noise variance ratio and the variation of the features.

6. REFERENCES

[1] K. Demuynck, J. Duchateau, and D. V. Compernolle, "Optimal feature sub-space selection based on discriminant analysis," Eurospeech, 1999.

[2] R. A. Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," ICASSP, 1998.

[3] G. Saon and M. Padmanabhan, "Minimum Bayes error feature selection," ICSLP, 2000.

[4] C. H. Yim, "Auditory Spectrum Based features (ASBF) for Robust Speech Recognition," ICSLP, 2000.

[5] X. Aubert, R. Haeb-Umbach, and H. Ney, "Continuous mixture densities and linear discriminant analysis for improved context-dependent acoustic models," ICASSP, 1993.

[6] N. Kumar and A. G. Andreou, "Heteroscedastic discriminan analysis and reduced rank HMMs for improved speech recognition," Speech Communication, pp. 283-297, 1998.

[7] L. Rabiner and B.-H. Juang, Fundamentals of Speech Recognition. Prentice Hall Press, 1993.

[8] O.Ghitza, Auditory Nerve Representations as a Basis for Speech Recognition. Marcel Dekker, 1991.

[9] A. Biern, S. Katagiri, E. McDermott and B-H. Juang, "An Application of Discriminative Feature Extraction to Filter-Bank-Based Speech Recognition", IEEE Transactions on Speech and Audio Processing, Vol 9, No. 2, Feb. 2001

[10] I. T. Jolliffe, Principal Component Analysis. Springer-Verlag, 1986.

[11] http://rocling.iis.sinica.edu.tw/ROCLING/