Exploiting Principal Component Analysis in Modulation Spectrum Enhancement for Robust Speech Recognition

Jan-Yee Lee
Department of Environmental Engineering
Kun Shan University
Tainan, Taiwan

Jeih-weih Hung
Department of Electrical Engineering
National Chi Nan University
Nantou, Taiwan

Abstract

In this paper, we present a novel method to improve the noise robustness of speech features based on principal component analysis (PCA). The PCA process is employed to extract a set of basis spectral vectors for the modulation spectra of clean training speech features. The new modulation spectra of the speech features, constructed by mapping the original modulation spectra into the space spanned by these PCA-derived basis vectors, have shown robustness against the noise distortion. The experiments conducted on the Aurora-2 digit string database revealed that the proposed PCA-based approach, together with mean and variance normalization (MVN), can provide average error reduction rates of over 65% and 12% relative as compared with the baseline MFCC system and that using the MVN method alone, respectively.

Keywords: robust speech recognition, modulation spectrum, principal component analysis

1. Introduction

A speech recognition system is often degraded seriously in performance due to additive noise and/or channel distortion. Various robustness methods have been proposed to reduce this mismatch, and one class of methods focus on extracting the speech-dominant modulation frequency components within the temporal-domain speech features. It has been shown in [1] that different modulation frequency components have unequal importance for speech recognition, and the components within the range [1 Hz, 16 Hz] contain a great amount of useful linguistic information, with the component around 4 Hz being the most significant. As a result, many temporal processing methods have been presented to emphasize these important frequency components, either explicitly or implicitly, for robust speech recognition. Some of these methods are cepstral mean subtraction (CMS) [2], mean and variance normalization (MVN) [3], RASTA [4] and a series of data-driven filter design schemes [5][6].

In this paper, we investigate a novel use of principal component analysis (PCA) [7] to learn a more "expressive" representation of the magnitude modulation spectrum of speech features. PCA is a well-known subspace method for finding a linear combination scheme to represent the original data. With PCA, a set of orthogonal basis vectors can be found, which makes the mappings of the original data to different basis vectors uncorrelated. Besides, PCA performs dimension reduction well since the PCA-reserved data approximate the original data optimally in the minimum-mean-squared sense (MMSE).

Consequently, it seems that PCA is suitable for the purpose of analysis and dimension reduction for the magnitude modulation spectrum of speech features, which often presents a relatively narrow bandwidth characteristic.

By viewing the magnitude modulation spectra of various clean feature sequences as the observations (samples), we apply PCA and obtain the dominant basis spectral vectors to span a subspace for the magnitude modulation spectrum. Then any clean or noise-corrupted feature sequence is updated in modulation spectrum, in which the magnitude part is replaced by its mapping on the PCA subspace mentioned above. Experiments conducted on the Aurora-2 database show that the new features with the PCA-updated modulation spectrum can maintain high recognition accuracy for the matched clean condition, and they provide significant accuracy improvement relative to the original features under mismatched noisy conditions. Accordingly, the presented new PCA-based approach can give rise to a more noise-robust speech feature representation.

The remainder of the paper is organized as follows: Section 2 briefly introduces the principal component analysis. Next, we describe the proposed PCA-based modulation spectrum update procedure in Section 3. The experimental setup is described in Section 4, followed by a series of experiments and discussions in Section 5. Finally, Section 6 concludes this paper and discusses the future work.

2. Principal Component Analysis

Principal component analysis (PCA) has been widely used in data analysis and dimensionality reduction in order to obtain the most "expressive" representation of the data [7]. For a random vector $\mathbf{x} \in \mathbb{R}^N$ with zero mean, the PCA process finds r ($r \le N$) orthogonal vectors on which the projection of $\mathbf{x}$ has the maximum variance. These r orthogonal vectors are in fact the eigenvectors of the covariance matrix for $\mathbf{x}$ corresponding to the largest r eigenvalues. It can be shown that, the projection of $\mathbf{x}$ the subspace spanned by these r eigenvectors is closest to $\mathbf{x}$ in the mean-squared sense compared with its projection to any other subspace with dimension r . That is, the cost function defined by

$$L = E\left(\left\|\mathbf{x} - \sum_{i=1}^{r} \langle \mathbf{x}, \mathbf{e}_i \rangle \mathbf{e}_i \right\|^2\right), \quad (1)$$

978-1-61284-181-6/11/$26.00 ©2011 IEEE

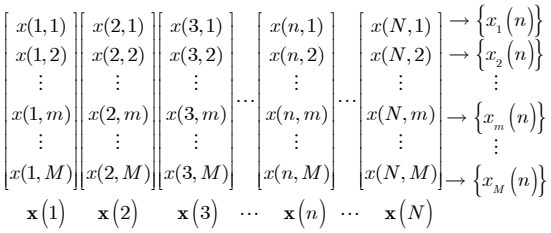


Figure 1. The representation of the time trajectories of feature parameters.

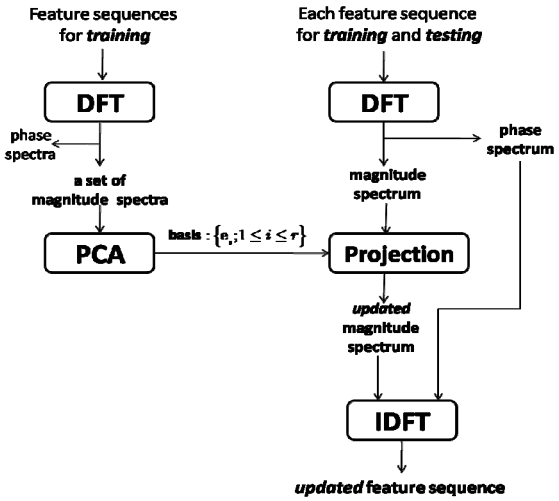


Figure 2. The flowchart of the proposed PCA-based approach for updating the modulation spectrum of features.

subject to the orthonormality constraint

$$\left\langle \mathbf{e}_i, \mathbf{e}_j \right\rangle = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}, \tag{2}$$

can be minimized by choosing $\{\mathbf{e}_i; 1 \le i \le r\}$ to be the eigenvectors of the covariance of $\mathbf{x}$ with the largest r eigenvalues. The symbol "$\langle .,. \rangle$" in Eqs. (1) and (2) represents the inner-product operation.

3. Updating the Modulation Spectrum via PCA

An ordered sequence of M-dimensional feature vectors $\{\mathbf{x}(n), n = 1, 2, \cdots, N\}$, where n is the time index, is illustrated in Figure 1. Each vector $\mathbf{x}(n)$ is represented as an $M \times 1$ column in the matrix shown in Figure 1,

$$\mathbf{x}(n) = [x(n,1), x(n,2), \cdots, x(n,M)]^T, \; n = 1, 2, \cdots, N, \tag{3}$$

where $x(n,m)$ is the m-th component of the feature vector $\mathbf{x}(n)$ at time n. Therefore, the time trajectory for the m-th feature parameter is the m-th row in the matrix shown in Figure 1:

$$[x(1,m), x(2,m), \cdots, x(N,m)], \; m = 1, 2, \cdots, M,$$

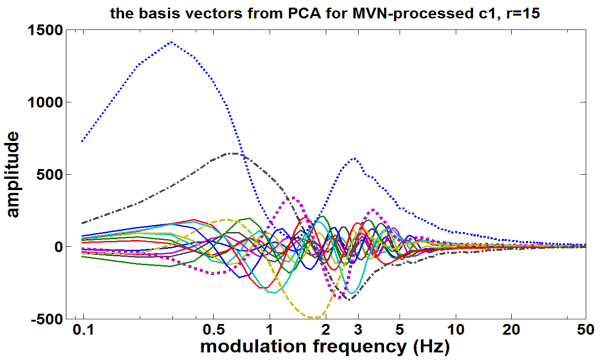


Figure 3. 15 basis spectra by PCA (each weighted by the corresponding eigenvalue) from the MVN-processed *c*1 features.

which is denoted here as a sequence $\{x_m(n), n = 1, 2, \cdots, N\}$, where

$$x_m(n) = x(n,m), \; n = 1, 2, \cdots, N, \;\; m = 1, 2, \cdots, M, \tag{4}$$

noting that n is the time index and m is the feature index. Here, each feature stream $\{x^{(m)}[n]\}$ is processed with the objective that the resulting new feature stream, denoted as $\{\tilde{x}^{(m)}[n]\}$, can be more noise-robust or more discriminative, and thus the recognition accuracy can be enhanced. For the sake of compact notation, we hereafter omit the superscript "$^{(m)}$" in $\{x^{(m)}[n]\}$ and $\{\tilde{x}^{(m)}[n]\}$. The procedures of our proposed method are depicted in Figure 2, and described as follows:

Step I: Obtain the PCA subspace for the magnitude modulation spectrum of speech features.

The time sequence $\{x[n]\}$ for each utterance in the clean training set is converted to its spectrum $\{X[k]\}$ via a 2L-point DFT. Since the property of conjugate symmetry, only the first L+1 points of $\{X[k]\}$ is reserved. The corresponding magnitude parts, $\{|X[k]|\}$ are represented in a (L+1)-by-1 vector, which is viewed as a sample (instant) of a random vector $\mathbf{x}$. With these samples from the clean utterances in the training set, we calculate the sample covariance $\mathbf{\Sigma}_x$ and then find the eigenvectors corresponding to the largest r eigenvalues of $\mathbf{\Sigma}_x$ (assuming these eigenvalues are distinct). Thus, the resulting r eigenvectors, denoted by $\{\mathbf{e}_i; 1 \le i \le r\}$ form a basis for the PCA subspace of the magnitude modulation spectrum.

Step II: Map the magnitude modulation spectrum for each utterance in the training and testing sets to the PCA subspace.

The (L+1)-point magnitude modulation spectrum of each utterance in the training and testing sets, denoted by a vector $\mathbf{v}$, is projected to the PCA subspace spanned by the basis $\{\mathbf{e}_i; 1 \le i \le r\}$, and thus the projection is:

$$\tilde{\mathbf{v}} = \sum_{i=1}^{N} a_i \mathbf{e}_i \tag{5}$$

where the coefficient for each eigenvector is

$$a_i = \langle \mathbf{v}, \mathbf{e}_i \rangle , \quad (6)$$

in which the operation $\langle .,. \rangle$ is the inner product.

Accordingly, the vector $\tilde{\mathbf{v}}$ is the mapping of $\mathbf{v}$ to the PCA subspace, which is created from the modulation spectra of clean utterances. We thus expect that the vector $\tilde{\mathbf{v}}$, representing the new magnitude spectrum, can emphasize the important information for speech recognition and reduce the effect of noise from the original $\mathbf{v}$.

Step III: Create the new time sequence via the updated modulation spectrum

A 2*L*-point inverse DFT is performed on the new modulation spectrum (with the conjugate symmetric last-half part being appended), which consists of the *updated* magnitude parts and the original phase parts, to obtain the new time sequence. Note that only the first *N* points of the obtained 2*L*-point sequence are reserved, where *N* is the length of the original time sequence. In fact, this *N*-point truncation process is equivalent to the *N*-point least-squared estimation of the signal from a 2*L*-point discrete-time spectrum.

Figure 3 depicts the PCA basis spectra (each weighted by the eigenvalue) for the modulation spectrum of the MVN-processed *c*1 (the first MFCC feature), corresponding to the clean training set of the Aurora-2 database [8]. From this figure we have three findings. First, these basis spectra are primarily located in the frequency region below 15 Hz, and thus they can preserve the lower modulation frequency components of the speech features, which correspond to important information for speech recognition [1]. Second, the eigenvalues (each corresponding to the degree of oscillation for the basis spectrum) decreases rapidly, implying the magnitude modulation spectrum can be approximated well with a small number of basis spectra. Finally, the observation that the basis spectra overlap one another along the frequency axis indicates the components with different frequency are mutually correlated.

4. Experimental Setup

The proposed PCA-based method was tested with the Aurora-2 database [8], which is widely used for evaluating robustness algorithms under noisy conditions. For the experiments, three subsets are defined: Test Sets A and B are each affected by four types of noise, and Test Set C is affected by two types. Each noise instance is added to the clean speech at six SNR levels. In addition, compared with the clean training set, the signals in both Sets A and B are distorted by additive noise, while those in Set C are distorted by additive noise and a channel mismatch.

Each utterance in the clean training set and three noise-corrupted testing sets was first converted into a sequence of 13-dimensional feature vectors (MFCC, *c*0-*c*12), which was then processed by mean-and-variance normalization (MVN) as to alleviate the effect of noise. Next, following the procedures described in Section III, we obtained the basis spectral vectors from PCA and then updated the modulation spectrum of each feature sequence of each utterance in both the training and testing sets. The DFT size 2*L* was set to 1,024, and the number of basis vectors, r, is varied from 5 to 15, with an interval of 5. The obtained new features plus their delta and delta-delta were the components of the final 39-dimensional feature vectors. With these feature vectors, the HMMs for each digit and silence were trained following the Microsoft complex back-end training scripts [9]. Each digit HMM had 16 states and 20 mixtures per state.

Table 1. Recognition accuracy (%) achieved by MFCC baseline, MVN and the various methods processed on the MVN-processed MFCC features, averaged across the ten noise types. r is the number of basis spectra used in PCA. RR_1 (%) and RR_2 (%) are the relative error rate reductions over the MFCC baseline and MVN, respectively.

Method		clean	Noisy (SNR: 20 dB~0 dB)	RR_1	RR_2
MFCC baseline		99.79	71.92	—	—
MVN		99.80	85.38	47.93	—
PCA	$r = 5$	99.55	**89.40**	**62.25**	**27.49**
	$r = 10$	99.65	88.65	59.58	22.37
	$r = 15$	99.69	88.01	57.30	17.99
HEQ		99.76	87.36	54.99	13.54
MVA		99.80	88.55	59.22	21.68

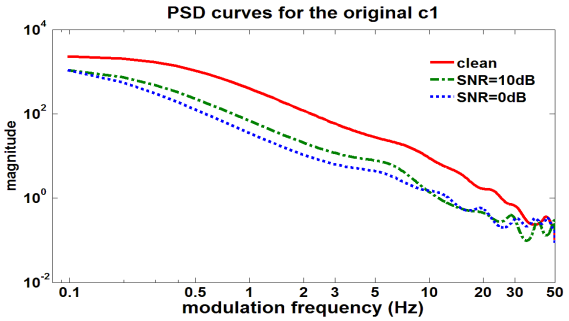
5. Experimental Results and Discussions

We compare the recognition performance of the MFCC baseline, MVN baseline and the proposed PCA-based approach conducted on MVN-processed MFCC. For comparison purpose, two well-known temporal processing approaches, histogram equalization (HEQ) [10] and MVN plus ARMA filtering (MVA)[11], were conducted here. The corresponding results are shown in Table 1, from which several observations can be made:

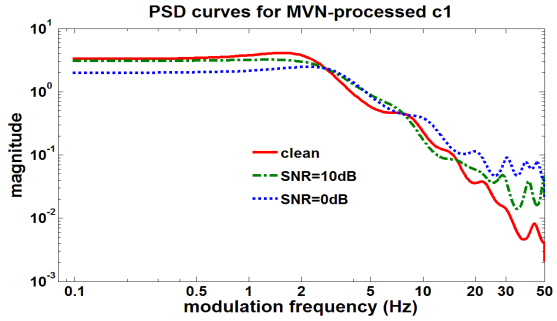
(1) Under the matched clean condition, PCA slightly worsens the recognition accuracy compared to the MFCC baseline and MVN. However, the accuracy degradation is relatively insignificant, which supports our previous statement that using only a small number of basis vectors from PCA can preserve most essential information in modulation spectrum for speech recognition.

(2) For mismatched noisy conditions, PCA enhances the MVN-processed MFCC to promote the recognition accuracy. Compared with MVN, the optimal accuracy improvement provided by PCA with $r = 5$ is 4.02%, which corresponds to 27.49% in relative error reduction. Observing the PCA basis spectra shown in Figure 3, the higher modulation spectral components that probably correspond to noise can be alleviated, the MVN-processed MFCC is further enhanced in noise robustness by PCA.

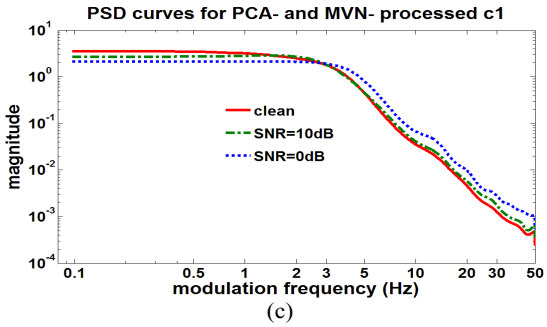
(3) Increasing the number of basis spectra, r, in PCA brings a slight recognition improvement for the clean case (from 99.55% with $r = 5$ to 99.69% with $r = 15$). However, the situation is converse for the noisy case. The probable explanation for the above phenomenon is: A greater number of basis spectra somewhat reserve the speech information better, but it may introduce the



(a)



(b)



(c)

Figure 4. The $c1$ PSD curves of an utterance ("FAK_48Z66ZZA.08" in the Aurora-2 database) after various processing methods with three SNR levels, clean, 10 dB and 0 dB: (a) no processing, (b) MVN, (c) PCA+MVN.

noise-affected components simultaneously and thus degrade the recognition performance.

(4) Both HEQ and MVA perform well and are additive to MVN. However, PCA with $r = 5$ behaves better than HEQ and MVA in recognition accuracy for mismatched noisy conditions, showing the proposed PCA method compares favorably with these two well-known robustness approaches in this recognition task.

In addition to the recognition accuracy, the PCA method is examined by its capability of reducing the modulation spectrum distortion. Fig. 4(a)-(c) show the power spectral density (PSD) curves of the first MFCC feature $c1$ of an utterance (the file "FAK_48Z66ZZA.08" in the Aurora-2 database) for three SNR levels, clean, 10 dB and 0 dB (with airport noise), before and after various processes, respectively.

First, for the unprocessed case as in Fig. 4(a), it shows that the additive noise results in a significant PSD mismatch over the entire frequency range [0 50 Hz]. Second, Fig. 4(b) shows that MVN reduces the lower-frequency PSD mismatch significantly, whereas it is less capable of dealing with the PSD mismatch higher than 10 Hz. Finally, from Fig. 4(c) we find that integrating PCA with MVN can further alleviate the PSD distortion over the entire band compared to MVN alone. As a result, the above findings again imply that, the presented PCA-based approach can reduce the effect of noise in the MVN-processed MFCC features and result in a more noise-robust feature representation.

6. Conclusion and Future Work

We presented a novel approach to improve the noise robustness of speech features based on PCA. We show that the basis spectra from PCA correspond to important modulation frequency components for speech recognition. This PCA-based method performing on the MVN-processed MFCC yields an accuracy improvement of up to 18% and 4% absolute compared with MFCC baseline and MVN alone, respectively, on average over all test conditions of the Aurora-2 task.

Future work will investigate whether further normalizing the projection coefficients a_i in Eq. (4) can bring better recognition accuracy. Also, we will examine if some other factor analysis techniques, such as nonnegative matrix factorization (NMF) [12], linear discriminant analysis (LDA) [7] and independent component analysis (ICA) [13], can provide a more compact and noise-robust representation of the modulation spectrum for speech features.

References

[1] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the importance of various modulation frequencies for speech recognition," *in European Conf. Speech Communication and Technology* (*Eurospeech*), 1997

[2] S. Furui, "Cepstral analysis technique for automatic speaker verification", IEEE Trans. Acoust., Speech, Signal Process., vol. 29, no. 2, Apr. 1981

[3] R. Haeb-Umbach, "Investigations on inter-speaker variability in the feature space", *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* (*ICASSP*), 1999

[4] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 2, no. 4, 1994

[5] J-W. Hung and W-Y. Tsai, "Constructing modulation frequency domain based features for robust speech recognition", *IEEE Trans. Acoust., Speech, Lang. Process.*, 2008

[6] J-W. Hung and L-S. Lee, "Optimization of temporal filters for constructing robust features in speech recognition", *IEEE Trans. Acoust., Speech, Lang. Process.*, 2006

[7] Richard O. Duda, Peter E. Hart and David G. Stork, "Pattern Classification", 2nd Edition, *Wiley*, 2000

[8] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech

recognition systems under noisy conditions," *ISCA ITRW ASR* 2000

[9] J. Droppo, L. Deng, and A. Acero, "Evaluation of SPLICE on the AURORA 2 and 3 tasks," *in Proc. Int. Conf. Spoken Lang. Process*. (*ICSLP*), 2002

[10] F. Hilger and H. Ney, "Quantile based histogram equalization for noise robust speech recognition", *in European Conf. Speech Communication and Technology* (*Eurospeech*), 2001

[11] C-P. Chen and J. Bilmes, "MVA processing of speech features", *IEEE Trans. Acoust., Speech, Lang. Process.*, 2007

[12] D. D. Lee and H. S. Seung, "Algorithms for Non-negative Matrix Factorization," *Advances in Neural Information Processing Systems* 13, 2000

[13] P. Comon, "Independent Component Analysis, a new concept?" *Signal Processing*, *Elsevier*, 1994