INTERNATIONAL JOURNAL OF ENGINEERING AND MANAGEMENT SCIENCES
© 2004 -14 Society For Science and Nature (SFSN). All Rights Reserved
www.scienceandnature.org

SPEECH ENCHANTMENT ALGORITHM USING KERNEL PCA

Tola Saimir, Nikolla Ligor
Department of Mathematical Engineering, Polytechnic University of Tirana,
Corresponding Author Email: saimir_tola@yahoo.com

ABSTRACT

We are analyzing a power-full method using kernel PCA (principal component analysis)for distorted speech recognition. Research for robust speech feature extraction has been done, but it is difficult to completely remove convolution noise. The noise removal techniques are based on the spectral analysis field, and then for speech recognition, the MFCC (Mel frequency cepstral coefficient) is computed, where DCT (Discrete Cousine Transform) is applied to the Mel-scale filter bank output. This paper describes a new PCA-based speech enhancement algorithm using kernel PCA instead of DCT, where the main speech element is projected onto low-order features, while the noise or the distortion element is projected onto high-order features. The effectiveness of this method is confirmed by word recognition experiments on distorted speech.

KEYWORDS: PCA (Principal Component Analysis), DCT (Discrete Cousine Transform), MFCC (Mel frequency cepstral coefficient), Kernel.

INTRODUCTION

Current speech recognition systems are capable of achieving impressive performance in clean acoustic environments. However if the speaker speaks at a distance from the microphone, the recognition accuracy is seriously degraded by the influence of additive and convolution noise. The noise is usually caused by telephone channels, reverberation ect. Its effect on the input speech appears as a convolution in the wave domain and is represented as a multiplication in the linear-spectral domain. Conventional normalization techniques, such as CMS (Cepstral Mean Subtraction) and Rasta, have been proposed, and their effectiveness has been confirmed for the telephone channel or microphone characteristics, which have a short impulse response. When the length of the impulse response is shorter than the analysis window used for the spectral analysis of speech, those methods are effective. However, as the length of the impulse response of the room reverberation (acoustic transfer function) becomes longer than the analysis window, the performance degrades. To solve problems caused by additive and convolution noise, many methods have been presented in robust speech recognition, but it is difficult to completely remove non-stationary or unknow noise. In current speech recognition technology, the MFFC(Mel Frequency cepstral coefficient) has been widely used. The feature is derived from the mel-scale filter bank output using DCT. The low-order MFFCs account for the slowly changing spectral envelope, while the high-order ones describe the fast variations of the spectrum.

Ref.[9] has investigated a suitable transformation based on PCA that can reflect the statistics of speech data better than DCT to compute the MFCCs. In [10], a PCA-based approach for speech enhancement is proposed, where PCA is applied to the wave domain instead of the Fourier Transform. In [11], the filter bank coefficients are estimated by applying PCA to the FFT spectrum. In [12], the effect of a PCA was applied only to the low-order MFCCs that account for the spectral envelope.

In this paper, we investigate robust feature extraction using kernel PCA instead of DCT provides better performance for reverberant speech.

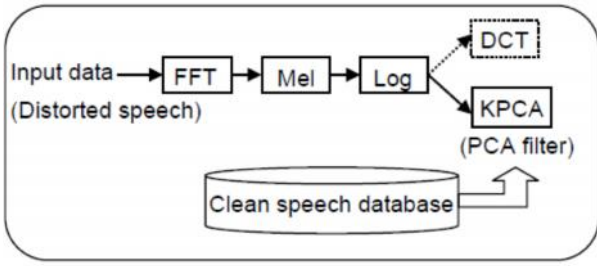

Figure 1. Feature extraction using kernel PCA.PCA filter represents the statistics of clean speech data.

FEATURE EXTRACTION USING KERNEL PCA

A. Speech Enhancement

The distorted speech, $X_n(\omega)$, is generally considered as the multiplication of the clean speech and the convolution noise:

$$X_n(\omega)=S_n(\omega) * H_n(\omega) \quad (1)$$

Where $S_n(\omega)$ and $H_n(\omega)$ are the short term linear spectrum for the clean speech and the convolution noise (acoustic transfer function) of the frequency ω at the n-th frame, respectively.

The length of the acoustic transfer function is generally longer than that of the window. Therefore, the observed distorted spectrum is approximately represented by

$$X_n(\omega)\approx S_n(\omega) * H_n(\omega) \quad (2)$$

The multiplication can be converted to addition in the log-spectral domain as follows:

$X_{log\_n}(\omega) \approx S_{log\_n}(\omega) + H_{log\_n}(\omega)$ (3)

Where $X_{log\_n}(\omega)$, $H_{log\_n}(\omega)$, and $S_{log\_n}(\omega)$ are the log spectra for the observed signal, acoustic transfer function (convolution noise), and speech signal, respectively.

Next, we consider the following filtering based on PCA in order to extract the feature of clean speech only,

$\hat{S}$=V X_{log} (4)

The filter (eigenvector matrix), V, is derived by the eigenvalue decomposition of the centered covariance matrix of a clean speech data set, in which the filter consists of the eigenvectors corresponding to the L domaminant eigenvectors corresponding to the L dominant eigenvalues (L eigenvectors corresponding to the biggest L eigenvalues).

$V=[v^{(1)}, v^{(2)}, \ldots\ldots, v^{(L)}]$ (5)

Due to the orthogonality, the component of the convolution noise belonging to the subspace $[v^{(L+1)}, \ldots.., v^{(M)}]$ is canceled by this filtering operation. However, as shown in (3), the observed signal is approximately represented under the assumption of non-correlation between the clan speech and the convolution noise. In this paper, we focus on non-linear PCA (kernel PCA) in order to deal with the influence of the approximation. Kernel PCA first maps the function and then performs linear PCA on the mapped data. We can expect that noise will be canceled in the high-dimensional space.

B. Kernel PCA

PCA is a powerful technique for extracting structure from possibly high-dimensional data sets. But it is not effective for data with non-linear structure. In kernel PCA, the input data with nonlinear structure is transformed into a higher-dimensional feature space with linear structure, and then linear PCA is performed in the high-dimensional spece [15].

Given the mel-scale filter bank output (log spectrum) x_j at j-frame, the covariance matrix is defined as

$c = \frac{1}{N}\sum_{j=1}^{N} \overline{\Phi}(x_j)^T$ (6)

$\overline{\Phi}(x_j) = \Phi(x_j) - \frac{1}{N}\sum_{j=1}^{N} \Phi(x_j)$ (7)

Where the total number of frames is N, and Φ is a nonlinear map.

$\Phi: R^d \rightarrow R^{\infty}$ (8)

Note that the data in the high-dimensional space could have an arbitrarily large, possibly infinite, dimensionality, and d is the dimension of x.

We now have to find eigenvalues ^{a} and eigenvectors v satisfying

λ v=cv, (9)

$\lambda(\Phi(x_k)\, v) = (\Phi(x_k)\, cv)$, k=1….,N (10)

Also there exist coefficients α_i such that

$v = \sum_{i=1}^{N} \alpha_i \overline{\Phi}(x_i)$ (11)

Substituting (6) and (11) in (10), we get for the left side of the equations

$\lambda(\overline{\Phi}(x_k)v) = \lambda \sum_i \alpha_i \overline{\Phi}(x_k)\overline{\Phi}(x_i) = \lambda \sum_i \alpha_i \overline{K}_{ki}$ (12)

Where

$\overline{K}_{ki} = \overline{\Phi}(x_k)\overline{\Phi}(x_i)$ (13)

Also for the right side of the equation

$\overline{\Phi}(x_k)Cv = \overline{\Phi}(x_k)\frac{1}{N}\sum_j \overline{\Phi}(x_j)\,\overline{\Phi}(x_j)^T =$

$\overline{\Phi}(x_k)\frac{1}{N}\sum_i \alpha_i\{\sum_j \overline{\Phi}(x_j)\,\overline{\Phi}(x_j)^T\overline{\Phi}(x_i)\}$

$= \frac{1}{N}\sum_i \alpha_i \left[\overline{\Phi}(x_k)\left\{\sum_j \overline{\Phi}(x_j)\overline{\Phi}(x_j)^T\overline{\Phi}(x_i)\right\}\right]$

$= \frac{1}{N}\sum_i \alpha_i \sum_j\{\overline{\Phi}(x_k)\overline{\Phi}(x_j)\}\{\overline{\Phi}(x_j)\overline{\Phi}(x_i)\}$

$= \frac{1}{N}\sum_i \alpha_i \sum_j \overline{K}_{kj}\,\overline{K}_{ji}$ (14)

Thus we get

$N\lambda\alpha = \overline{K}\alpha$

$\hat{\lambda}\alpha = \overline{K}\alpha$ (15)

Consequently, we only need to diagonalize $\overline{K}$which is computed as follows.

$\overline{K}_{ij} = \overline{\Phi}(x_i)\overline{\Phi}(x_j)$

$= (\Phi(x_i) - \frac{1}{N}\sum_{m=1}^{N}\Phi(x_m)(\Phi(x_j) - \frac{1}{N}\sum_{n=1}^{N}\Phi(x_n))$

$= \Phi(x_i)\Phi(x_j) - \frac{1}{N}\sum_{m=1}\Phi(x_m)\Phi(x_j) -$

$\frac{1}{N}\sum_{n=1}\Phi(x_n)\Phi(x_i) + \frac{1}{N^2}\sum_{m,n=1}\Phi(x_m)\Phi(x_n)$

$= K_{ij} - \frac{1}{N}\sum_{m=1} 1_{im}K_{mj} - \frac{1}{N}\sum_{n=1} K_{in}1_{nj}$ (16)

$K_{ij} = \Phi(x_i)\Phi(x_j)$ (17)

$1_{ij} = 1\ for\ all\ i,j$ (18)

Using N x N matrix $(1_N)_{ij}$:=1/N, we get the more compact expression

$\overline{K} = K - 1_N K - K1_N + 1_N K 1_N$ (19)

We thus can compute $\overline{K}$ from K, and then solve the eigenvalue problem (15).

Let $\lambda_1 \leq \lambda_2 \leq \ldots.. \leq \lambda_N$ denote the eigenvalues, and $\alpha^{(1)}, \ldots, \alpha^{(N)}$ the corresponding complete set of eigenvectors, with λ_p being the first nonzero eigenvalue. We normalize $\alpha^{(p)}, \ldots., \alpha^{(N)}$ by requiring that the corresponding vectors are normalized:

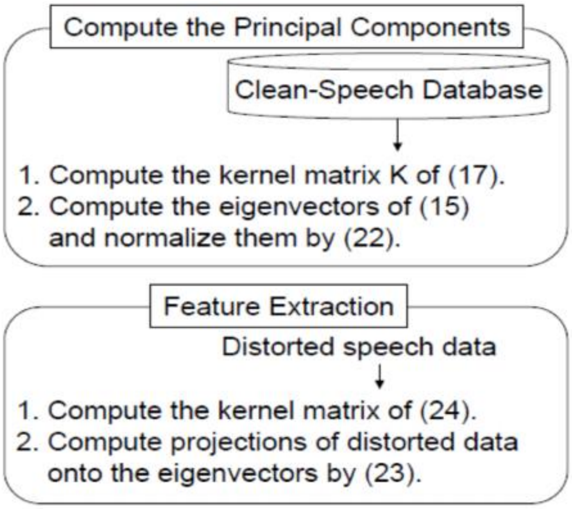
Compute the Principal Components
Clean-Speech Database
1. Compute the kernel matrix K of (17).
2. Compute the eigenvectors of (15) and normalize them by (22).
Feature Extraction
Distorted speech data
1. Compute the kernel matrix of (24).
2. Compute projections of distorted data onto the eigenvectors by (23).

Figure 2. Procedure of feature extraction

$v^{(l)}\, v^{(l)}=1$, for all l=p,…..,N (20)

from (11) and (15) we get

$1 = \sum_{i,j}^{N} \alpha_i^{(l)}\, \alpha_j^{(l)} \left(\Phi(x_i)\Phi(x_j)\right) = \sum_{i,j}^{N} \alpha_i^{(l)}\alpha_j^{(l)}\, K_{ij} =$

$\left(\alpha^{(l)}\overline{K}\alpha^{(l)}\right) = \hat{\lambda}_l(\alpha^{(l)}\alpha^{(l)})$ (21)

Next, for feature extraction, we project test data y onto eigenvectors $v^{(l)}$ in the high-dimensional space.

$$\left(v^{(i)}\bar{\Phi}(y)\right) = \sum_{i=1}^{N} \hat{\alpha}_i^{(l)} (\bar{\Phi}(x_i)\,\bar{\Phi}(y)) = \sum_{i=1}^{N} \hat{\alpha}_i^{(l)} \bar{K}^{test}(x_i, y) \quad (23)$$

Similar to (16) we can compute $\bar{K}^{test}$ from K^{test}.

$$\bar{K}_{ij}^{test} = (\Phi(y_i) - \frac{1}{N}\sum_{m=1}^{N}\Phi(x_m))(\Phi(x_j) - \frac{1}{N}\sum_{n=1}^{N}\Phi(x_n)) \quad (24)$$

$$\bar{K}^{test} = K^{test} - 1_N' K - K^{test} 1_N + 1_N' K 1_N \quad (25)$$

Here $1_N'$ is the L x N matrix with all entries equal to 1/N, and the total number of frames for the test data is L. the procedure of the feature extraction is summarized in Fig.2.

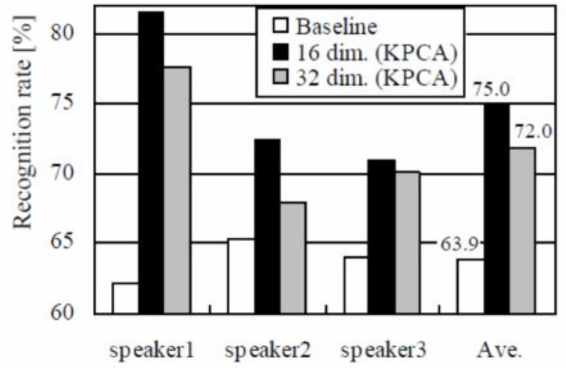

Figure 3. Recognition rates for the reverberant speech (reverbation time: 470 msec) by the proposed method (p=1 in polynomial function)

RECOGNITION EXPERIMENT

The new feature extraction method was evaluated on reverberant speech recognition task. Reverberant speech was simulated using a linear convolution of clean speech and impulse response. The impulse response was taken from the RWCP sound scene database [16]. The reverberation time was 470 msec. the distance to the microphone was about 2 meters, and the size of recording room was about 6.7 m x 4.2m (width x depth). In order to compute the matrix, K, it would be necessary to use all the training data, but it is not realistic in terms of the cost of the computation. Therefore, in this experiment, N=2,500 frames were ramdomly picked from the training data, and we used the polynomial kernel function.

$K(x, y) = (xy + 1)^p$ (26)

The speech signal was sampled at 12 kHz and windowd with a 32-msec Hamming window every 8 msec. the models of 54 context-independent phonemes were trained by using 2,600 words in the ATR Albanian speech database for the speakers-dependent HMM. Each HHM has three states and three self-loops, and each state has four Gaussian mixture components. The tests were carried out on 1,000 words. The baseline recognition rate was 63.9%, were 16-oreder MFCCs and their delta coefficients were used as feature vectors.

Experimental results

Figure 3 shows the recognition rates using kernel PCA (p=1 in polynomial functions). As can be seen from figure 3, the use of kernel PCA instead of DCT improves the recognition rates from 63.9% to 75.0%. here, in the new feature extraction, kernel PCA was applied to 32-dimension mel-scale filter bank output, and then the delta coefficients were also computed. Figure 4 shows the recognition rates using kernel PCA (p=2 in polynomial function). These results clearly show that the performance is better when using kernel PCA instead of DCT.

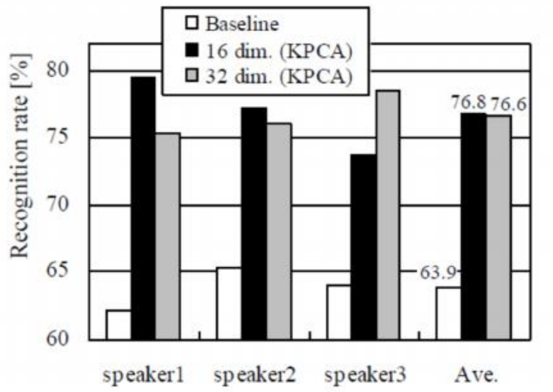

Figure 4. Rrecognition rates for the reverberant speech (reverberation time: 470 msec) by the proposed method (p=2 in polynomial function)

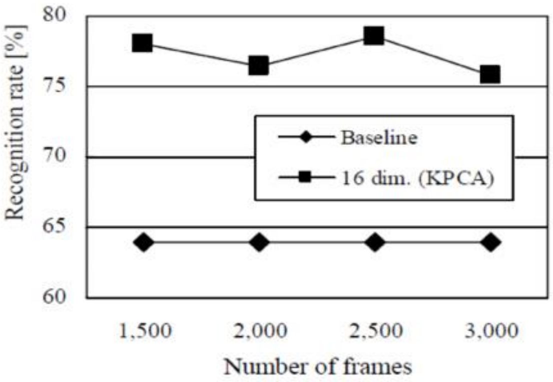

Figure 5. Recognition rates for test speaker3 when kernel PCA is applied using different amounts of training data.

The kernel PCA for the polynomial function for p=1 is almost same as the linear PCA. The recognition rate using the linear PCA described in the Section 2-A is actually 75% on average. Compared to Figure 3, the recognition rate is equal to that of the kernel PCA (p=1). Next, we applied kernel PCA to 16-order MFCCs [13] [14]. The recognition rate improved from 63.9% to 67.8%. as can be seen from figure 4, a further improvement was obtained by the new method, where kernel PCA was applied to mel-scale filter bank output. This is because we can expect that kernel PCA in the spectral domain will project the main speech element onto low-order features, while the reverberant elements will be projected onto high-order features.

Figure 5 shows the performance of test speaker3 when the kernel PCA is applied using different amounts of training data in (6). In this case, increasing the amount of training data does not significantly improve the performance of the kernel PCA. This result shows that the use of 2,500 frames training data is suitable for this experiment.

Figure 6 shows the recognition rates for clean speech by the proposed method. The recognition rate with the new feature extraction was 97.3%. in clean environments, the experiment results indicate that the new method achieves almost the same performance as that of DCT.

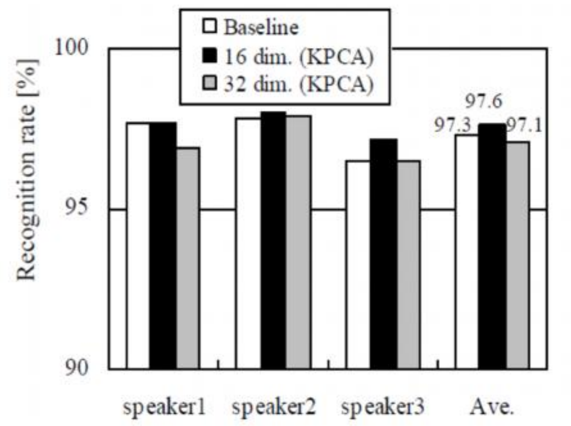

Figure 6. Recognition rates for the clean speech by the proposed method. (p=2 in polynomial function)

Next, Table 1 shows the performance using the sigmoid kernel as shown in (27) instead of the polynomial kernel,

$K(x,y) = \tanh(axy - \sigma)$, (27)

Where $\sigma = 0.01$, and the recognition rates for test speaker3 are shown. The results in table 1 show a decrease in recognition rate, compared to the polynomial kernel. Also, it is difficult to find two appropriate parameters, a and σ, in the sigmoid kernel.

Finally, we examined the performance for the kernel principal component based on the speaker-independent (SI) data instead of the speaker-dependent (SD) data. In this case, 2,500 frames from 25 males were used for calculation of $\bar{K}$ in (15), and the acoustic model was trained using the SD data in order to examine only the accuracy of PCA filter estimated by SI data. (*) shows the recognition rates for the speaker-dependent data. The recognition rate results in a 1.5% decrease on average because of increasing the speaker variability.

Tabele1 Recognition rates [%] with the sigmoid function

	16 dim.	24 dim.	32 dim.
a=0.0001	58.8	60.7	61.7
a=0.00005	71.6	69.7	68.3
a=0.00001	73.0	71.3	72.6
a=0.000005	71.6	72.7	73.4

Table 2 Recognition rates [%] when the kernel principal component is estimated by speaker-independent data

	16 dim.	24 dim.	32 dim.
p=1	70.7 (71.0)	72.9 (74.0)	72.2 (70.1)
p=2	72.0 (73.7)	73.7 (74.8)	74.4 (78.5)
p=3	72.0 (75.6)	73.3 (74.1)	73.3 (76.1)

SUMMARY

This paper has described a PCA-based speech enhancement technique for distorted speech recognition, where kernel PCA is applied to the mel-scale filter bank output. It can be expected that kernel PCA will project the main speech element onto low-order features, while the reverberant (noise) element will be projected onto high-order features, and the PCA-based filter will extract the feature of clean speech only. From our recognition results, it is shown that the use of kernel PCA instead of DCT provides better performance for reverberant speech. (reverberation time:470 msec).

REFERENCES

[1] S. Mika, B. Scholkopf, A.J. Smola, K.-R. Muller, M.Scholz, and G. Ratsch, “Kernel PCA and de-noising in feature spaces,” In M.S. Kearns, S.A. Solla, and D.A. Cohn, editors, Advances in Neural Information Processing Systems 11, pp. 536–542, MIT Press, 1999.

[2] H. Hermansky and N. Morgan, “RASTA Processing of Speech,” IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 4, pp. 578-589, 1994.

[3] C. Avendano, S. Tivrewala, and H. Hermansky, “Multiresolution channel normalization for ASR in reverberant environments,” Eurospeech, pp. 1107-1110, 1997.

[4] U. H. Yapanel and J. H. L. Hansen, “A New Perspective on Feature Extraction for Robust In-Vehicle Speech Recognition,” Eurospeech, pp. 1281-1284, 2003.

[5] B. J. Shannon and K. K. Paliwal, “Influence of Autocorrelation Lag Ranges on Robust Speech Recognition,” ICASSP, pp. 545-548, 2005.

[6] W. Li, K. Itou, K. Takeda and F. Itakura, “Two-Stage Noise Spectra Estimation and Regression Based In-Car Speech Recognition Using Single Distant Microphone,” ICASSP, pp. 533-536, 2005.

[7] M. Fujimoto, S. Nakamura, “Particle Filter Based Non-Stationary Noise Tracking for Robust Speech Recognition,” ICASSP, pp. 257-260, 2005.

[8] K. Kinoshita, T. Nakatani and M. Miyoshi, “Efficient Blind Dereverberation Framework for Automatic Speech Recognition,” Interspeech, pp. 3145-3148, 2005.

[9] M. Tokuhira and Y. Ariki, “Effectiveness of KLTransformation in Spectral Delta Expansion,” Eurospeech99,pp. 359-362, 1999.

[10] R. Vetter, N. Virag, P. Renevey and J.-M. Vesin, “Single Channel Speech Enhancement Using Principal Component Analysis and MDL Subspace Selection,” Eurospeech, 1999.

[11] S-M. Lee, S-H. Fang, J-W. Hung and L-S. Lee, “Improved MFCC Feature Extraction by PCA-Optimized Filter Bank for Speech Recognition,” Automatic Speech Recognition and Understanding, 2001, ASRU, pp. 49-52, 2001.

[12] F. Asano, Y. Motomura, H. Asoh and T. Matsui, “Effect of PCA Filter in Blind Source Separation,” Proc. ICA2000, pp. 57-62, 2000.

[13] A. Lima, H. Zen, Y. Nankaku, C. Miyajima, K. Tokuda, and T. Kitamura, “On the Use of Kernel PCA for Feature Extraction in Speech Recognition,” IEICE Trans. Inf. & Syst., Vol. E87-D, No. 12, pp. 2802-2811, 2004.

[14] A. Lima, H. Zen, Y. Nankaku, K. Tokuda, T. Kitamura and F. G. Resende, “Applying Sparse KPCA for Feature Extraction in Speech Recognition,” IEICE Trans. Inf. & Syst., Vol. E88-D, No. 3, pp. 401-409, 2005.

[15] B. Sch¨olkopf, A. Smola, and K.-R. M¨uller, “Nonlinear component analysis as a kernel eigenvalue problem,” Neural Computation, Vol. 10, pp. 1299-1319, 1998.

[16] S. Nakamura, K. Hiyane, F. Asano, T.Nishiura, T. Yamada, “Acoustical Sound Database in Real Environments for Sound Scene Understanding and Hands-Free Speech Recognition,” Proceedings of International Conference on Language Resources and Evaluation, Vol. 2, pp. 965-968, 2000.