A Space Transformation Approach for Robust Speech Recognition in Noisy Environments

Cun-tai Guan, Shu-hung Leung and Wing-hong Lau
Department of Electronic Engineering
City University of Hong Kong, Kowloon, Hong Kong
Tel: +852 2788 7193, Fax: +852 2784 4262, E-mail: eeguan@cpccux0.cityu.edu.hk

ABSTRACT

To improve the robustness of speech recognition in additive noisy environments, an SVD based space transformation approach is proposed. It is shown that with this approach, not only the signal-to-noise ratio is improved but also a significant recognition error reduction is achieved. A multiple model based on the proposed method is developed and it can provide high recognition rate for a large range of SNRs. Recognition experiments on a speaker-dependent mono-syllabic database with additive noise show that, this new approach outperforms LPC cepstrum, MFCC, and OSALPC cepstrum significantly.

1 Introduction

The performance degradation of a speech recognition system operated in noisy environments is due to the mismatch between training and testing conditions. To solve the problem, a great deal of interest is in the developing of robust front end. Spectral subtraction was first used in speech enhancement to improve the quality of noisy speech. It was then applied to speech recognition[1]. Juang and Rabiner [7] proposed a spectral mapping approach to transform the noisy speech feature vectors into clean feature vectors. MMSE estimation was introduced by Ephraim [6] and Erell and Weitraub [9] obtained a significant increase in recognition performance for noisy speech. A lot of literature focus on the development of robust representations of speech; for instance, auditory based features [2,3], correlation domain based features [4,5]. Jensen *et al.* in [8] proposed a speech enhancement method based on truncated SVD and Quotient SVD (QSVD). They showed that the truncated SVD and QSVD were effective to improve the SNR of broad-band noisy speech. In this paper, we propose a new front-end preprocessing approach which adopts the Least Square(LS) estimation based on SVD to transform speech into a new signal space. We will prove that with this approach, not only the signal-to-noise ratio is improved but also a significant recognition error reduction is achieved. We further apply the processing on the autocorrelation data matrix to enhance the recognition performance in even lower SNR conditions. Finally we propose a multi-model based on this transformation scheme. Recognition experiments on a speaker-dependent mono-syllabic database with additive noise show that, this new approach outperforms LPC cepstrum, MFCC, and OSALPC cepstrum significantly for a large range of SNRs.

2 Space transformation based on SVD

2-1 Singular value decomposition on data matrix

Considering a speech contaminated by an additive noise:

$$x_i = s_i + n_i , \tag{1}$$

where x_i, s_i, and n_i are the noisy speech, clean speech, and noise signal respectively. We can form the following $K \times P$ Hankel data matrices

$\mathbf{S} = [\mathbf{s}_1, \cdots, \mathbf{s}_P]$, $\mathbf{N} = [\mathbf{n}_1, \cdots, \mathbf{n}_P]$, and

$$\mathbf{X} = \mathbf{S} + \mathbf{N} = [\mathbf{x}_1, \cdots, \mathbf{x}_P], \tag{2}$$

where

$\mathbf{s}_i = [s_i, \cdots, s_{i+K-1}]^T$, $\mathbf{n}_i = [n_i, \cdots, n_{i+K-1}]^T$, $\mathbf{x}_i = [x_i, \cdots, x_{i+K-1}]^T$

with $K>P$. For speech signal, it is assumed that rank($\mathbf{S}$)=$L \le P$ and rank($\mathbf{N}$)=rank($\mathbf{X}$)=P.

The singular value decomposition of $\mathbf{X}$ is given by

$$\mathbf{X} = \mathbf{U}_x \Sigma_x \mathbf{V}_x^T , \tag{3}$$

where $\mathbf{U}_x = [\mathbf{u}_{x,1}, \cdots, \mathbf{u}_{x,P}]$ is the orthonormal matrix, $\mathbf{V}_x = [\mathbf{v}_{x,1}, \cdots, \mathbf{v}_{x,P}]$ the unitary matrix, $\mathbf{u}_{x,i} \in R^K$ and $\mathbf{v}_{x,i} \in R^P$ are respectively the left and right singular vectors, $\Sigma_x = diag(\sigma_{x,1}, \cdots, \sigma_{x,P})$ the singular values of the matrix $\mathbf{X}$. The SVD of $\mathbf{X}$ can be rewritten in the following partition form

$$\mathbf{X} = \begin{bmatrix} \mathbf{U}_{x1} & \mathbf{U}_{x2} \end{bmatrix} \begin{bmatrix} \Sigma_{x1} & 0 \\ 0 & \Sigma_{x2} \end{bmatrix} \begin{bmatrix} \mathbf{V}_{x1}^T \\ \mathbf{V}_{x2}^T \end{bmatrix} \tag{4}$$

where $\mathbf{U}_{x1} \in R^{K \times L}, \Sigma_{x1} \in R^{L \times L}, \mathbf{V}_{x1} \in R^{P \times L}$.

The SVD of $\mathbf{S}$ can also be written as

$$\mathbf{S} = \begin{bmatrix} \mathbf{U}_{s1} & \mathbf{U}_{s2} \end{bmatrix} \begin{bmatrix} \Sigma_{s1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{V}_{s1}^T \\ \mathbf{V}_{s2}^T \end{bmatrix} \tag{5}$$

For ease of discussion, it is assumed that the noise is white and uncorrelated with the clean speech, i.e.

$$\mathbf{n}_i^T \mathbf{n}_j = \lambda^2 \delta(j-i) , \text{ and } \mathbf{s}_i^T \mathbf{n}_j = 0 \tag{6}$$

where $\delta(j)$ is a Kronecker delta function and λ^2 the energy of noise vector. The relationship of the SVD of $\mathbf{X}$ with regard to $\mathbf{N}$ and the SVD of $\mathbf{S}$ is given as[8]

$$\mathbf{U}_x = \left[(\mathbf{U}_{s1}\Sigma_{s1} + \mathbf{N}\mathbf{V}_{s1})(\Sigma_{s1}^2 + \lambda^2 \mathbf{I})^{-1/2} \quad \lambda^{-1}\mathbf{N}\mathbf{V}_{s2} \right] \tag{7}$$

$$\Sigma_x = \begin{bmatrix} (\Sigma_{s1}^2 + \lambda^2 \mathbf{I})^{1/2} & 0 \\ 0 & \lambda \mathbf{I} \end{bmatrix} \tag{8}$$

$$\mathbf{V}_x = [\mathbf{V}_{s1} \quad \mathbf{V}_{s2}] \tag{9}$$

2-2 Space transformation as a preprocessing

In [8], Jensen *et al* constructed a speech enhancement method from the SVD estimation. However, for speech recognition, it is not necessarily needed to transform a noisy speech to approximate the original clean speech. We should try to alleviate the mismatch between clean and noisy speech. With this motivation, we formulate the following preprocessing procedures for speech recognition, in which both the training clean speech and the testing noisy speech are transformed into a new signal space before the feature estimation.

A) Preprocessing of clean training speech

From the SVD of clean speech data matrix in (5), the least square (LS) estimation of $\mathbf{S}$ at rank l is

$$\mathbf{S}^{(l)} = \mathbf{U}_{s1}^{(l)} \Sigma_{s1}^{(l)} \mathbf{V}_{s1}^{(l)T} \tag{10}$$

where $\mathbf{U}_{s1}^{(l)} = [\mathbf{u}_{s1,1}, \mathbf{u}_{s1,2}, \cdots, \mathbf{u}_{s1,l}]$,

$\mathbf{V}_{s1}^{(l)} = [\mathbf{v}_{s1,1}, \mathbf{v}_{s1,2}, \cdots, \mathbf{v}_{s1,l}]$,

$\Sigma_{s1}^{(l)} = diag(\sigma_{s1,1}, \sigma_{s1,2}, \cdots, \sigma_{s1,l})$, $l \le L$.

It is reasonable to restore the "speech" vector from any column of $\mathbf{S}^{(l)}$, for instance, the first column. Then we have the preprocessed vector of clean speech

$$\mathbf{s}^{(l)} = \mathbf{U}_{s1}^{(l)} \Sigma_{s1}^{(l)} \mathbf{v}_{s1}^{(l,1)T} \tag{11}$$

where $\mathbf{v}_{s1}^{(l,1)}$ is the first row vector of $\mathbf{V}_{s1}^{(l)}$. It is easy to see that $\mathbf{s}^{(l)}$ contains only partial information of the original speech due to the reduced rank approximation in (11). We refer this signal as quasi-speech.

B) Preprocessing of noisy testing speech

The noisy testing speech can be processed in the same way as the clean speech. By utilizing the equations (3) through (9), we have an LS estimate from the noisy speech as

$$\overline{\mathbf{S}}^{(l)} = \mathbf{U}_{x1}^{(l)} \Sigma_{x1}^{(l)} \mathbf{V}_{x1}^{(l)T} = (\mathbf{U}_{s1}^{(l)} \Sigma_{s1}^{(l)} + \mathbf{N}\mathbf{V}_{s1}^{(l)}) \mathbf{V}_{s1}^{(l)T} \tag{12}$$

The estimated speech vector is derived as

$$\overline{\mathbf{s}}^{(l)} = (\mathbf{U}_{s1}^{(l)} \Sigma_{s1}^{(l)} + \mathbf{N}\mathbf{V}_{s1}^{(l)}) \mathbf{v}_{s1}^{(l,1)T} = \mathbf{s}^{(l)} + \mathbf{n}^{(l)} \tag{13}$$

C) SNR improvement due to the preprocessing

The SNR of the original noisy speech is defined as

$$\overline{SNR} = 10\log_{10} \frac{\|\mathbf{s}\|^2}{\|\mathbf{x}-\mathbf{s}\|^2} = 10\log_{10} \frac{\|\mathbf{s}\|^2}{\|\mathbf{n}\|^2} = 10\log_{10} \frac{E_s}{\lambda^2} \tag{14}$$

where the energy of clean speech is related to singular values as $E_s = \|\mathbf{s}\|^2 \approx \frac{1}{P}\|\mathbf{S}\|_F^2 = \frac{1}{P}\sum_{j=1}^{L} \sigma_{s1,j}^2$ (15)

Therefore, $\overline{SNR} \approx 10\log_{10}\{(1/P \cdot \sum_{j=1}^{L} \sigma_{s1,j}^2)/\lambda^2\}$. (16)

In the similar way, the SNR of preprocessed signal, or noisy quasi-speech, is defined and derived as

$$\hat{SNR}^{(l)} = 10\log_{10} \frac{\|\mathbf{s}^{(l)}\|^2}{\|\overline{\mathbf{s}}^{(l)} - \mathbf{s}^{(l)}\|^2} = 10\log_{10} \frac{\sum_{j=1}^{l} v_{s1,j,1}^2 \sigma_{s1,j}^2}{\lambda^2 \sum_{j=1}^{l} v_{s1,j,1}^2} \tag{17}$$

It is easy to see that $\hat{SNR}^{(l-1)} > \hat{SNR}^{(l)}$, for $1 < l \le L$. Considering that the original noisy speech is the special case of noisy quasi-speech with $l=L$, we come to the conclusion that the quasi-speech possesses higher SNR than the original noisy speech, i.e.

$$\hat{SNR}^{(1)} > \hat{SNR}^{(2)} > \cdots > \hat{SNR}^{(L)} = \overline{SNR} \tag{18}$$

Fig. 1 depicts the average $\hat{SNR}^{(l)}$ versus l for the noisy speech data formed by artificially adding white noise to a sequence of about 150 second clean Cantonese speech. The signal-to-noise ratios are 10dB and 5dB in the sense of SNRMAX measurement, which are 9.714dB and 4.715dB respectively in the measurement of segmental SNR, or average $\overline{SNR}$. For both cases, when l=1, approximate 10dB improvement of SNR is obtained. Furthermore, if the quasi-speech is reconstructed by averaging the anti-diagonal elements of (10) and (12), it can be shown that SNR is further improved (the proof is omitted for the sake of brevity) which is evidently indicated in Fig 1 with the lines labeled by "Hankel".

3 Recognition experiments

3-1 Database

A Cantonese speech database for telephony application was collected, sampled at 16KHz, and end-points were manually marked. The vocabulary of the database contains 10 digits and 4 control words. The speech data are recorded from 4 female and 4 male speakers. Each speaker uttered the words 22 times. 12 clean utterances form the training database and the remaining 10 utterances of the words constitute the testing database which is added with various levels of white noise to generate different SNRs ranging from 40dB to 0dB, where the SNR is measured in the sense of SNRMAX.

3-2 Preprocessing in waveform domain

In this experiment, the preprocessing is accomplished in the waveform domain. After the training and testing speech are preprocessed, any feature analysis method, e.g. LPC, MFCC, can be used to create the feature parameters. We here use LPC analysis and transform LPC coefficients into cepstral coefficients and delta cepstral coefficients. The recognition results are listed in Table 1. The recognition rates of quasi-speech are denoted with QLPC. LPC and MFCC analysis without the preprocessing are also tested on the same database. The feature vector contains 12D cepstral and 12D delta cepstral coefficients. All analysis uses 20ms frame window with 10ms shift. The orders of LPC and QLPC are both 16. MFCC is analyzed with 40 filters. For each word a continuous HMM model is trained, which is composed of 6 states, 3 mixtures, and diagonal Gaussian probability density. In these experiments, the number of

columns of data matrix for SVD is P=16. The recognition results show that the performance of QLPC is significantly better than those of LPC and MFCC

3-3 Preprocessing in autocorrelation domain

Hernando and Nadeu[5] proposed a robust speech feature OSALPC, which was demonstrated to be more robust than LPC and MFCC in noisy environments. The preprocessing method proposed in 2-2 can straightforward be extended to the autocorrelation domain. We can form a data matrix in autocorrelation domain as

$$\mathbf{Y} = [\mathbf{r}_1^+, \mathbf{r}_2^+, \cdots, \mathbf{r}_P^+]. \quad (19)$$

where $\mathbf{r}_i^+ = \left[r_i^+, r_{i+1}^+, \cdots, r_{K+i-1}^+\right]^T$, and r_i^+ is the one-sided autocorrelation sequence. We can carry out the similar preprocessing described in section 2-2 on this autocorrelation matrix. The resulting sequence is called quasi-autocorrelation sequence. After the preprocessing has been performed, the same feature analysis procedure as in QLPC can be implemented to result in QOSALPC parameters. Speech recognition with OSALPC and QOSALPC are carried out on the same database as above. The analysis order of OSALPC and QOSALPC are both 16. The number of column for the SVD preprocessing is still P=16. The recognition results are listed in Table 2. The results show that the performance of QOSALPC is significantly better than that of OSALPC for SNR below 10dB. It can been seen, from Table 1 and Table 2, that in the cases of SNR> 20dB, QLPC gives the best performance, and at the cases of SNR<20dB, QOSALPC is the best.

4 Multiple model approach

In section 2, we proved that the lower the reconstruction rank, the higher the SNR of the reconstructed quasi-speech can be obtained; however the preprocessing may lose some information when the reconstruction rank is relatively low. Therefore, to eliminate noise and to retain as much speech information as possible are conflicted. There must be a compromise between them. The experiments in section 3 demonstrate the statements, where at each noise condition, the best performances of QLPC and QOSALPC are obtained at a specific rank of reconstruction and the best rank is $l \le 3$. When SNR is relatively high, the loss of speech information dominates, so the optimum rank is higher, while SNR is low, the impact of noise is more important, so the optimum rank must be lower. Generally speaking, the optimum rank is SNR dependent. Furthermore, from the experiments in section 3, it can be seen that there exists a distinct SNR division, say 20dB, above and below which QLPC and QOSALPC respectively give the best performances. To combine the two analysis methods in a unified framework will yield the best performance over all SNR conditions. Based on this motivation, we propose a multiple model approach in this section. The approach is based on two considerations:

A) For each word, in the training phase, we train several HMM models using different ranks of quasi-speech (rank $l \le 3$), and in the recognizing phase, each testing speech is preprocessed to yield several observation sequences with different ranks of reconstruction. The optimum model is selected by scoring each observation sequence with corresponding HMM models. This scheme avoids the determination of optimum rank by exactly estimating SNR which is not an easy job;

B) In order to obtain the best performance over the whole SNR range (0-40dB), an SNR threshold should be determined in order to automatically select a proper analysis method from QLPC and QOSALPC in different SNR conditions. This SNR should be easily obtained based on the SVD and do not use any a priori knowledge about the noise.

4-1 SNR estimation based on singular values

The 2-norm condition number $\kappa_2(\mathbf{Y})$, which is defined as $\kappa_2(\mathbf{Y}) = \|\mathbf{Y}\|_2 \|\mathbf{Y}^{-1}\|_2 = \sigma_1/\sigma_P$ (20)

where σ_1 and σ_P are the maximum and minimum singular values of matrix $\mathbf{Y}$. In the case of noisy speech $\mathbf{Y} = \mathbf{S} + \mathbf{N}$, $\kappa_2(\mathbf{Y})$ will decrease with the increase of noise level. Therefor, the 2-norm condition number of a noisy speech matrix is a good measurement of SNR. Based on this property, we can explicitly defined an estimated SNR based on SVD as

$$SNR_{SVD} = \frac{10}{T}\sum_{t=1}^{T} \log_{10} \kappa_2(\mathbf{Y}_t), \quad (21)$$

where T denotes the length of a word, i.e., the average is over the whole word. Experiments, on the training database with various levels of noises, show that the distribution of segmental SNR and the distribution of SNR_{SVD} are very similar, especially under the conditions of SNR<25dB. Recalling that there is a SNR division at 20dB between QLPC and QOSALPC, then a threshold can be easily determined with the SNR_{SVD}

$TH_{SNR} = \max\{SNR_{SVD},\ SNR_{SVD} \in x(SNR = 20dB)\}$.

In our experiments, $TH_{SNR} = 19.4dB$.

4-2 Multiple model recognition experiment

The recognition experiment with the multiple model is carried out as follows. In the training phase, eight HMM models are trained using clean speech for each word. Four models are trained by deploying full rank and rank 1, 2 and 3 QLPC parameters and the other four are trained by deploying full rank and rank 1,2, and 3 QOSALPC parameters. During the testing phase, SNR_{SVD} is firstly estimated. Then, with the guidance of the SNR_{SVD}, the speech is transformed and analyzed with either QLPC or QOSALPC method and is scored with

corresponding HMM models. The final recognition results are listed in Table 3.

5 Conclusion and discussion

A new preprocessing approach for robust speech recognition is proposed. Multiple model approach is shown to provide very good results. It is worthwhile to mention that even with rank one preprocessing, the recognition performance is acceptable in some extent. The computational complexity can be significantly reduced by calculating only the largest singular value and its corresponding vectors and make the robust method well suitable for low cost implementation.

ACKNOWLEDGMENT

This research is supported by UGC under the Grant No. 9040176 and CityU STRATEGIC under the Grant No. 7000330.

REFERENCES

[1] A. Acero, Acoustical and environmental robustness in automatic speech recognition, Kluwer Academic Publishers, Boston/Dordrecht/London, 1993.

[2] H. Hermansky, "Perceptual Linear Prediction (PLP) analysis of speech", *J. Acoust. Soc. Am.*, vol. 87, pp. 1738-1752, Apr. 1990.

[3] O. Ghitza, "Robustness against noise: The role of time-synchrony measurement", in *Proc. Int. Conf. Acoust. Speech, Signal Processing*, Apr. 1987, pp.2372-2375.

[4] D. Mansour, and B. Juang, "The short-time modified coherence representation and its application for noisy speech recognition", in *Proc. Int. Conf. Acoust. Speech, Signal Processing*, Apr. 1988, pp.525-528.

[5] J. Hernando and C. Nadeu, "Speech recognition in noisy car environment based on OSALPC representation and robust similarity measuring techniques", in *Proc. Int. Conf. Acoust. Speech, Signal Processing*, Apr. 1994, pp. II-69-72.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator", *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 32, pp. 1109-1121, Dec. 1984.

[7] B. Juang and L. Rabiner, "Signal restoration by spectral mapping", in *Proc. Int. Conf. Acoust. Speech, Signal Processing*, Apr. 1987, pp. 2368-2371.

[8] S.H. Jensen, P.C. Hansen, S.D. Hansen and J.A. Sorensen, "Reduction of broad-band noise in speech by truncated QSVD", *IEEE Trans. On Speech and Audio Processing*, Vol. 3, pp. 439-448, June 1997.

[9] A. Erell, M. Weintraub, "Energy conditioned spectral estimation for recognition of noisy speech",*IEEE Trans. On Speech and Audio Processing*, Vol. 1, pp. 84-89, Jan. 1993.

Table 1 Recognition rates of LPC, MFCC, and QLPC under various SNRs (P=16).

rank	clean	SNR(dB) 40	35	30	25	20	15
	LPC						
	100	83.57	62.86	37.14	27.86	20.71	15.00
	MFCC						
	100	97.14	78.57	57.14	43.57	34.29	23.57
	QLPC						
1	97.86	97.86	97.86	97.86	**97.86**	**94.29**	**50.00**
2	98.57	98.57	**100**	**99.29**	95.00	77.86	38.31
3	97.86	**100**	99.29	90.71	67.86	32.86	29.22
4	97.86	98.57	97.86	88.57	57.86	35.00	20.97
5	97.14	97.14	94.29	77.14	38.57	26.43	19.73
6	97.14	96.43	92.86	70.00	37.86	27.14	16.48
8	99.28	95.71	87.86	51.43	31.43	20.00	15.00
10	100	93.57	80.71	39.29	27.86	20.00	15.00
16	100	83.57	62.86	37.14	27.86	20.71	15.00

Table 2 Recognition rates of OSALPC and QOSALPC under various SNRs (P=16).

rank	clean	SNR(dB) 40-25	20	15	10	5	0
	OSALPC						
	92.14	92.14	92.14	91.42	85.00	40.71	14.29
	QOSALPC						
1	87.14	87.86	87.86	87.86	87.86	**87.14**	**69.29**
2	89.29	90.00	90.71	90.71	87.86	82.86	60
3	**94.29**	**94.29**	**94.29**	**92.86**	**90.71**	77.86	48.57
4	92.14	92.14	92.86	92.14	87.14	60.00	17.14
5	92.86	92.86	92.14	92.86	85.00	43.57	14.29
6	91.43	91.43	91.43	92.14	87.14	45.00	14.29
8	90.00	90.00	90.71	92.14	87.14	46.43	14.29
10	90.71	90.71	90.71	92.14	87.14	42.14	14.29
16	92.14	92.14	92.14	91.42	85.00	40.71	14.29

Table 3 Recognition accuracy rates of the multiple model approach.

clean	40dB	35dB	30dB	25Db
100	100	100	99.29	97.86

20dB	15dB	10dB	5dB	0dB
94.29	92.86	90.71	87.14	69.29

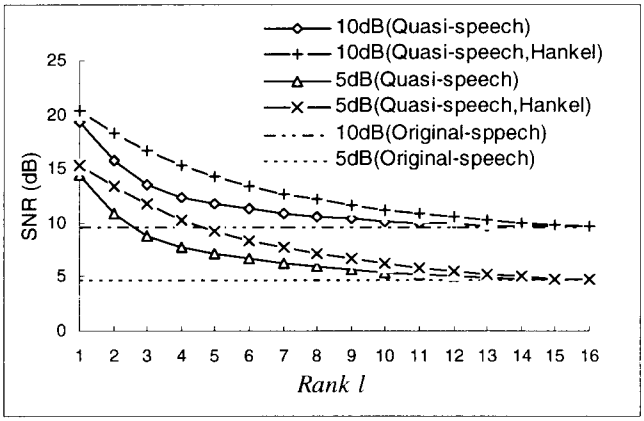

Figure 1. Average $\hat{SNR}$ versus l at the conditions of global SNR of 5dB and 10dB, which correspond to average $\overline{SNR}$ of 4.715dB and 9.714dB respectively.