·ROBUST SPEECH RECOGNITION USING SINGULAR VALUE DECOMPOSITION BASED SPEECH ENHANCEMENT

B.T. Lilly and K.K. Paliwal

Signal Processing Laboratory
School of Microelectronic Engineering
Griffith University
Brisbane, QLD 4111, Australia
B.Lilly, K.Paliwal@me.gu.edu.au

ABSTRACT

Speech recognition systems work reasonably well in laboratory conditions, but their performance deteriorates drastically when they are deployed in practical situations where the speech is corrupted by additive noise. One way to improve the performance of a speech recognition system in the presence of noise, is to enhance the speech prior to its recognition. Two singular value decomposition based techniques have been recently proposed for speech enhancement [5] [6]. In these techniques, singular value decomposition has been applied to an over-determined, over-extended data matrix formed from the noisy speech signal. A noise-free, low rank approximation was obtained by retaining a specific number of singular values. This technique was applied here as a preprocessor for recognising speech in the presence of noise. It was found to improve the recognition performance significantly for signal-to-noise ratios less than 15dB.

1. INTRODUCTION

Speech recognition has been shown to perform well under clean, ideal conditions. However, when deployed in practical environments, the performance of the system degrades rapidly, due to the presence of noise and other distortions. For example, an isolated word recognizer that can recognize 10 English digits perfectly when spoken in laboratory environment, recognizes only 30% of the spoken digits when white noise is added to the signal with 10 dB signal-to-noise ratio (SNR) [1].

Much research has been devoted to improving the robustness of speech recognition systems in noisy environments. As speech in noisy or distorted environments is not always available, a practical solution is to train recognition systems on clean (noise free) speech utterances and use speech enhancement algorithms to clean the noisy speech utterances prior to their recognition.

Spectral subtraction has been shown to work to some extent for speech recognition [2] [4]. The spectral subtraction algorithm works on the assumption that the noise contained in the noisy speech signal is additive and uncorrelated with the clean speech signal. In this method, the short-time magnitude and phase spectra of the noisy signal is computed. The magnitude spectrum is modified by subtracting the short-time spectral magnitude of the noise estimated during the non-speech periods, while the phase spectrum is left unchanged. The signal reconstructed from the modified magnitude and phase spectra is an enhanced version of the original signal [3].

A singular value decomposition (SVD) based speech enhancement technique has recently been proposed in the literature [5]. Like the spectral subtraction method, this technique also assumes that the noise contained in the noisy speech is additive and uncorrelated with the clean speech signal. The SVD technique enhances a noisy signal by retaining a few of the singular values from the decomposition of an over-determined, over-extended data matrix. The singular values that are ignored are associated with the noisy part of the signal. The signal reconstructed from the reduced rank matrix is the enhanced speech signal.

In the experiments presented in this paper, we use singular value decomposition to enhance noisy speech corrupted with white noise as a preprocessor for recognition. Two SVD based techniques have been proposed in the literature - a least squares method [5] and a minimum variance technique [6]. These two methods are used in speech recognition experiments presented in this paper. The results obtained using these SVD algorithms are then compared to those obtained using spectral subtraction as a preprocessor for recognition.

Section 2 of this paper describes the enhancement algorithms used in [5] [6]. Section 3 describes the experimental setup for the enhancement algorithms, databases used and the recognition system. Section 4 shows results for both speech enhancement in terms of SNR and speech recognition performance using the TIMIT continuous speech database. Section 5 presents our conclusions.

2. SINGULAR VALUE DECOMPOSITION BASED SPEECH ENHANCEMENT

SVD-based speech enhancement relies on the assumption that the noise $\mathbf{n}$ contained in noisy speech signal $\mathbf{x}$ is additive and uncorrelated with the clean speech signal $\bar{\mathbf{x}}$, ie.

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{n} \tag{1}$$

If we take $\mathbf{x}$ as a vector having length N, we can construct a LxM Hankel matrix $\mathbf{H}$, where M is the over-determined, over-extended order (full rank) of L data vectors.

$$\mathbf{H} = \begin{bmatrix} x_0 & x_1 & \dots & x_{M-1} \\ x_1 & x_2 & \dots & x_M \\ . & . & \dots & . \\ . & . & \dots & . \\ x_{L-1} & x_L & \dots & x_{N-1} \end{bmatrix} \tag{2}$$

The choice of L is determined by the length of N due to the conditions L+M=N+1 and L≥M. Using singular value decomposition, we can decompose the matrix **H** into matrices **U** and **V** (with orthonormal columns) and $\mathbf{\Sigma}$ [9], which represents the singular values of the noisy matrix **H**, ie.

$$\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{3}$$

where $\mathbf{\Sigma} = \text{diag}(\ \sigma_1\ ,\ \sigma_2\ ,\ ...\ ,\ \sigma_M\ )\ ,\ \sigma_1 \gg ... \gg \sigma_M$.

We then predict a rank K such that the first K singular values of the diagonal matrix $\mathbf{\Sigma}$ represent the clean speech signal and the last M-K singular values represent the noise. By setting the singular values representing the noise to zero, we can reconstruct a new reduced rank matrix $\bar{\mathbf{H}}$ representing the clean speech signal $\bar{\mathbf{x}}$ using (4). To obtain the vector $\bar{\mathbf{x}}$ from $\bar{\mathbf{H}}$, the anti-diagonal components of $\bar{\mathbf{H}}$ are then averaged.

$$\bar{\mathbf{H}} = \mathbf{U}\mathbf{\Sigma}_\mathbf{K}\mathbf{V}^T \tag{4}$$

where $\mathbf{\Sigma}_\mathbf{K}$ represents the first diagonal K values of $\mathbf{\Sigma}$.

The final problem is how to predict the value for K. Two methods exist in the literature which calculate the value of K in a least squares sense [5] and using a minimum variance estimate [6].

The least squares method assumes the noise variance σ^2 contained in the noisy speech signal is known. This can be calculated by averaging all the singular values over a number of non-speech frames. The rank K of the matrix $\bar{\mathbf{H}}$ can then be calculated by first decomposing the noisy data matrix **H** into the three matrices **U**, **V** and $\mathbf{\Sigma}$ as in (3). The value K is chosen when the difference between the calculated noise variance for an estimated K value Σ_K^2 and the known noise variance σ^2 is a minimum,

$$\min_{1\leq K\leq M} \left|\Sigma_K^2 - \sigma^2\right| \tag{5}$$

where

$$\Sigma_K^2 = \left(\sum_{i=K+1}^{M} \Sigma_i\right) \tag{6}$$

This is known as the NEE criterion [5] used to implement the SVD algorithm for speech enhancement in a least squares sense and is the first of the two algorithms used in the recognition experiments presented in Section 4.

A problem with the least squares method is that it is sensitive to the selected value of K. If the value is not selected correctly, the speech enhancement performance is degraded. The SVD speech enhancement algorithm based on the minimum variance technique in [6] attempts to solve this problem by transforming the singular values before reconstruction of the reduced rank data matrix $\bar{\mathbf{H}}$ [6] [7]. Thus, the reconstruction of the data matrix becomes,

$$\bar{\mathbf{H}} = \mathbf{U}\mathbf{F}_{\mathbf{corr}}\mathbf{\Sigma}_\mathbf{K}\mathbf{V}^T \tag{7}$$

where F_{corr} is the identity matrix for the least squares algorithm,

$$\mathbf{F}_{\mathbf{corr}} = \mathbf{I}_\mathbf{k} \tag{8}$$

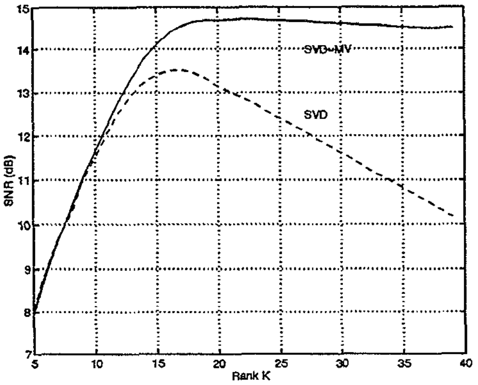

Figure 1. Plot of SNR versus Rank for the SVD-MV method (solid) and the SVD algorithm alone (dotted) for M=40 and N=256.

For the minimum variance estimation algorithm, F_{corr} is a diagonal matrix,

$$\mathbf{F}_{\mathbf{corr}} = diag\left(\left(1-\frac{\sigma_{noise}^2}{\sigma_1}\right),...,\left(1-\frac{\sigma_{noise}^2}{\sigma_K}\right)\right) \tag{9}$$

for which the proof is found in [6] [7].

3. EXPERIMENTAL SETUP

3.1. Enhancement Setup

3.1.1. Spectral Subtraction Algorithm

The spectral subtraction based experiments in this paper use an algorithm similar to the one in [3]. Firstly, the noise magnitude spectrum was estimated by averaging several noise frames. In the experiments presented in this paper, we regenerated noise based on the known variance of the noisy signal and estimated the short-term, average magnitude spectrum from it.

A Bartlett window of length 256 samples, is then applied to the noisy speech every 128 samples and a 256-point FFT is computed. The magnitude and phase spectrums are calculated from the short term spectrum of the noisy signal. The average magnitude spectrum of the noise is then subtracted from the magnitude spectrum of the original noisy signal. The resulting magnitude spectrum, with its negative spectrum values set to zero, is then combined with the phase spectrum of the noisy signal. An inverse FFT (IFFT) on this new complex spectrum then produces a noise reduced signal.

3.1.2. SVD Algorithms

The frame length N and the over-estimation M are two parameters common to both the SVD algorithms. The frame length was set to be the same as for the spectral subtraction algorithm, which was 256 samples with an overlap of 128 samples. A trade-off had to be made for the value of M with higher values of M giving better resolution and SNR improvements at the cost of computation time. From experimental results, a value of M=40 produced good results.

We also examined the effect of the chosen rank K on the SNR performance for the SVD algorithm. To evaluate the SNR performance in these experiments, white noise was added to clean speech for 2 male and 2 female utterances from the TIMIT database. The noise variance was calculated to give an overall SNR of 10dB. Figure 1 shows the SNR performance of the SVD algorithm on its own (dotted line) versus the chosen rank K. In this figure we notice that the SNR performance shows a peak for K=17. The SNR performance decreases relatively quickly either side of this value. Therefore, the rank K must be chosen carefully in order to obtain significant SNR improvements using the standard SVD algorithm.

The least squares algorithm solves this problem by choosing a rank based on prior knowledge of the noise variance. That is, it attempts to calculate the value of the rank K which corresponds to the peak in Figure 1.

In the minimum variance technique however, a high enough rank is selected so that the performance is independent of the final K value. This is shown in Figure 1 (solid line). Once the rank K is above 19, the SNR performance stays reasonably steady. Based on this figure and other experiments performed using these algorithms, a rank value of 24 was initially chosen for use with the minimum variance technique and was used for all SNR experiments presented in this paper. In the recognition experiments however, rank values of K=24 and K=35 were used for the minimum variance technique to examine the effect different values of K have on the recognition performance.

3.2. Recognition and Database

Speech recognition experiments presented in this paper use the TIMIT speaker-independent, continuous speech recognition database. Only the first dialect of this database was used to demonstrate the recognition performance of the SVD algorithms. This dialect consists of 304 training utterances from 24 male and 14 female speakers and 88 utterances for testing from 7 male and 4 female speakers (different from training). As the database has a sampling rate of 16kHz, all utterances in the training and testing sets were decimated to 8kHz using a low pass FIR filter with a cutoff frequency of 3.5kHz.

We used the 304 training utterances to train 48 context-independent models, as proposed in [8]. These models are trained on "clean" PCM-coded speech. Recognition was then performed using spectral subtraction and the two SVD enhancement algorithms as pre-processors to the estimation of the cepstrum coefficients. We also performed recognition on the noisy speech to analyze any improvements.

The HTK (HMM Tool Kit) package was used to observe the effect of using SVD speech enhancement on the testing data containing white noise with various global SNR values. We used simple left-to-right, 5-mixture HMM models with each model containing 3 states, except for the closures and silence models (epi, sil, vcl, cl) which were modeled with only a single state. All Gaussian densities used diagonal covariances and we used a bigram model for added robustness.

Speech recognition performance was evaluated using a robust speech recognition feature extraction method. Mel-spaced frequency cepstral coefficients (MFCC) were used as the primary extraction method with delta, log energy and delta log energy coefficients appended to them. Twelve MFCC's were computed from 19 filter bank energies every 10ms with a 20ms Hamming window.

Table 1. SNR Improvements for the SVD algorithms and spectral subtraction for different SNR values.

SNR (dB)	SVD-LS	SVD-MV	Spectral Subtraction
25	0.41	-3.43	1.79
20	1.90	0.26	2.88
15	3.36	3.12	4.05
10	5.05	5.03	5.23
5	6.65	6.34	6.32
0	8.23	7.39	7.27

4. EXPERIMENTAL RESULTS

4.1. SNR Performance

The SVD techniques and the spectral subtraction method described in previous sections were used in these experiments to evaluate their performance in terms of SNR improvement. White noise was added to clean speech for 2 male and 2 female utterances from the TIMIT database to produce SNR values ranging from 0 to 25dB. Table 1 shows the SNR improvement of the three enhancement schemes.

These results showed that both SVD algorithms improve the SNR of the speech for SNR values less than 25dB. The least squares technique outperformed the minimum variance technique at all SNR values since the least squares algorithm utilizes prior knowledge of the noise variance. At SNR values above 25dB, both SVD algorithms do not improve the speech in terms of SNR. However, only a small amount of distortion was audible at this level. This is tolerable to the human listener and does not degrade the audible quality. For SNR values less than 20dB, a small amount of "musical noise" was evident and became quite distracting at very low SNR values.

The spectral subtraction algorithm showed improvements at all SNR levels. For higher SNR values the spectral subtraction technique outperformed both SVD algorithms. However, for SNR values less than 15dB, the SVD techniques were comparable to or better than the spectral subtraction method. The amount of distortion introduced by the spectral subtraction scheme at lower SNR values was also comparable to the SVD techniques, with some musical noise present.

4.2. Recognition Performance

The first column of Table 2 shows the recognition performance of white noise corrupted speech for SNR values ranging 0 to 25dB. As expected, as the noise level increased the recognition performance degraded quickly. In comparison, the recognition performance using the SVD algorithms (Table 2) showed a significant increase in recognition performance for SNR's equaled to and less than 15dB.

This table shows that the recognition performance of the least squares method exhibited a steady decline with SNR. For a rank of K=24, the minimum variance SVD technique did not show the same linear decline in recognition performance. For higher SNR values, the recognition performance for the minimum variance technique did not change. This can attributed to the chosen value of the rank K being

Table 2. Recognition Performance at different SNR values.

SNR	Recognition Accuracy (%)			
dB	No Enhance	SVD-LS	SVD-MV (K=24)	SVD-MV (K=35)
∞	56.87			
25	42.97	39.40	41.08	46.02
20	38.01	35.52	41.95	45.30
15	31.28	32.19	41.64	44.43
10	24.68	30.31	40.21	42.44
5	20.38	27.29	37.00	36.88
0	18.23	25.89	32.77	32.00

Table 3. Recognition performance using the SVD and spectral subtraction speech enhancement algorithms.

SNR	Recognition Accuracy (%)		
(dB)	Spectral Subtraction	SVD-LS	SVD-MV (K=35)
25	51.48	39.40	46.02
20	48.15	35.52	45.30
15	44.49	32.19	44.43
10	39.70	30.31	42.44
5	35.58	27.29	36.88
0	31.22	25.89	32.00

too small for the higher SNR values but adequate for low SNR values. Increasing the value of K to 35 for the minimum variance technique (Column 4 of Table 2), showed that for lower SNR values the recognition performance did not degrade significantly. However at higher SNR values, the recognition performance increased as expected.

As found in section 4.1 for the SNR performance, the recognition performance using spectral subtraction outperformed the SVD enhancement algorithms at higher SNR values (Table 3). At lower values of SNR however, the minimum variance method performed comparable to or better than spectral subtraction. In Section 4.1 we demonstrated that the least squares SVD method performed better than the spectral subtraction method at low SNR levels. Surprisingly, in terms of recognition performance, the least squares SVD algorithm showed no improvement over spectral subtraction at any SNR level.

5. CONCLUSIONS

Using the least squares singular value decomposition method as a front-end pre-processor before recognition, improves the recognition performance for signal-to-noise ratios less than 15dB. The least squares method however, performs poorly compared to the spectral subtraction method. This is a surprising result as both these methods utilise a known noise variance of the signal.

The minimum variance technique shows good improvements in recognition performance for all SNR values as long as the rank K is large enough. For SNR values less than 20dB, the minimum variance technique obtains similar or better results than spectral subtraction. These results show that transforming the singular values in a minimum variance sense, provides a significant improvement in recognition performance.

REFERENCES

[1] Juang B.H., "Speech recognition in adverse environments" Computer Speech and Language, Vol. 5, pp. 275-294, 1991.

[2] Flores J.A.N., Young S.J., "Adapting a HMM-based recognizer for noisy speech enhanced by spectral subtraction", Proc. Eurospeech, pp829-832, Sept. 1993.

[3] Boll S.F., "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. Speech and Audio Processing, Vol. ASSP-27, No. 2, pp 113-120, April 1979.

[4] Mokbel C.E., Chollet G.F.A., "Automatic Word Recognition in Cars", IEEE Trans. Speech and Audio Processing, Vol. 3, No. 5, pp346-356, Sept 1995.

[5] Dendrinos M., Bakamidis S., Carayannis G., "Speech enhancement from noise: A regenerative approach" Speech Communication, Vol. 10, No. 2, pp 45-57, Feb. 1991.

[6] Jensen S., Hansen P., Hansen S., Sorensen J., "Reduction of Broad-Band Noise in Speech by Truncated QSVD" IEEE Trans. Speech and Audio Processing, Vol. 3, No. 6, pp. 439-448, Nov. 1995.

[7] Huffel S.V., "Enhanced resolution based on minimum variance estimation and exponential data modeling", Signal Processing, Vol. 33, No. 3, pp 333-355, Sept. 1993.

[8] Lee K., Hon H., "Speaker-Independent phone recognition using Hidden Markov Models", IEEE Trans. Speech and Audio Processing, Vol. 37, No. 11, pp 1641-1648, Nov. 1989.

[9] Strang G., "Introduction to Linear Algebra", USA: Wellesley-Cambridge Press, 1993.