ABSTRACT
The singular value decomposition (SVD)-based signal-subspace approach for noise reduction has received high interests in recent years. With this approach, we can diagonalize the matrices constructed from noisy speech frames and divide the whole feature-space into signal-subspace and noise-subspace by the singular values obtained from the matrices. We then reconstruct speech from the signal-subspace only. In this way, speech signals can be successfully enhanced. This approach is very effective when the additive noise is white. If the noise is not white, we have to first whiten the noise spectrum prior to SVD-based approach and perform the inverse whitening procedure afterwards. Not only the process is complicate, but extra distortion may be introduced in such a process. In this paper, a generalised SVD (GSVD)-based approach for speech enhancement is proposed, which is useful regardless of whether the added noise is white or not. Experimental results show that this new approach can provide very good performance, specially better than the conventional spectral subtraction algorithm and SVD-based approach, in particular when the additive noise is non-white.

The rest of the paper is organized as follows. In section 2, the principles of the GSVD-based approach are introduced. Experimental environment setup is then described in section 3, and experimental results are discussed in section 4. Some conclusions are finally given in the last section.

1. INTRODUCTION
In wired or wireless communication systems, voice quality and intelligibility is important for either human-to-human connections or human-to-machine interactions. In order to obtain good quality for communications, for example via mobile phones, speech enhancement techniques have been used to improve the quality and intelligibility of the noise-corrupted speech, to reduce the listener’s fatigue, and to improve the speech recognition accuracy. From the viewpoint of digital signal processing, we are very often classified the noise sources into two types: convolutional noise and additive noise. The former one is usually properly handled by some channel equalization and bias removal approaches [1,2]. For the additive noise, on the other hand, the spectral subtraction approach (SS) has been very popular [3]. However, for most of the additive noise sources the spectra are usually non-white. It is well known that the SS algorithm can offer significant performance improvements in the case of additive white noise, but becomes less helpful when the noise spectrum is non-white. This is the main issue we wish to address in this paper.

The singular value decomposition (SVD)-based signal-subspace approach for noise reduction has received high interests in recent years [4,5,6]. With this approach, we can diagonalize the matrices constructed from noisy speech frames and divide the whole feature-space into signal-subspace and noise-subspace by the singular values obtained from the matrices. We then reconstruct speech from the signal-subspace only. In this way, speech signals can be successfully enhanced. This approach is very effective when the additive noise is white. If the noise is not white, we have to first whiten the noise spectrum prior to SVD-based approach and perform the inverse whitening procedure afterwards. Not only the process is complicate, but extra distortion may be introduced in such a process. In this paper, a generalised SVD (GSVD)-based approach for speech enhancement is proposed, which is useful regardless of whether the added noise is white or not. Experimental results show that this new approach can provide very good performance, specially better than the conventional spectral subtraction algorithm and SVD-based approach, in particular when the additive noise is non-white.
\[ H_y = \begin{bmatrix} y_0 & y_1 & \cdots & y_{L-1} \\ y_1 & y_2 & \cdots & y_K \\ \vdots & \vdots & \ddots & \vdots \\ y_{L-1} & y_{L-2} & \cdots & y_{K-1} \end{bmatrix}, \] \hspace{1cm} (2)\]

and so does \( \hat{H}_y \). Here, \( H_y, \hat{H}_y \in \mathbb{R}^{L \times K} \), where \( L+K=M+1 \) and in generally \( K \) is smaller than \( L \). Under noise free situation, the value of \( K \) is chosen such that the matrix \( H_y \) is rank deficient, i.e. the rank of \( H_y \) is smaller than \( K \). This rank deficiency condition makes it easy for us to choose a boundary to separate the signal-subspace and the noise-subspace later on when noise is presented.

Next, by GSVD algorithm [4,7], a nonsingular matrix \( X \in \mathbb{R}^{L \times K} \) and two orthogonal matrices \( U, V \in \mathbb{R}^{L \times L} \) can be found which simultaneously transforms both \( H_y \) and \( \hat{H}_y \) into nonnegative diagonal form matrices \( C \) and \( S \),

\[ U^T H_y X = C = \text{diag}(c_1, \ldots, c_K), \quad c_i \geq 0, 2 \leq i \leq K \]
\[ V^T \hat{H}_y X = S = \text{diag}(s_1, \ldots, s_K), \quad s_i \geq 0, 2 \leq i \leq K, \] \hspace{1cm} (3)

where superscript \( T \) means the transpose of a matrix, the diagonal elements of the matrix \( C \) are arranged in descending order, and those of \( S \) are similar except in ascending order. There is a constraint for the matrices \( C \) and \( S \), \( C^T C + S^T S = I_K \), where \( I_K \) is a \( K \times K \) identity matrix. The values \( c_i/s_i, \ldots, c_K/s_K \) and the columns of the matrix \( X \) are respectively referred to as the generated singular values and generalized singular vectors of the matrices \( H_y \) and \( \hat{H}_y \).

From equations (1) and (2), it is evident that the matrix \( H_Y \) can be represented as the summation of two \( L \times K \) Hankel-form matrices \( H_D \) and \( H_S \), \( H_Y = H_D + H_S \), each of which is respectively constructed from the clean speech frame and the noise frame. Note here the matrix \( H_S \) is unknown but can be approximated by \( \hat{H}_s \) mentioned above. Therefore, we can estimate the matrix \( H_D \) from the matrices \( H_Y \) and \( \hat{H}_s \) by the GSVD algorithm as summarized below. The enhanced speech frame can then be obtained by regenerating the estimated matrix \( \hat{H}_s \) from the signal-subspace of \( H_D \) and concatenated frame-by-frame with the overlap-and-add method to form the enhanced speech.

### 2.2 Signal-subspace Construction by Minimum Variance Estimation

When the additive noise is white, the diagonal elements of the matrix \( S, s_i, 1 \leq i \leq K \), in equation (3), are almost identical. These values become quite different when the noise is non-white. For such cases of non-white noise, we then need to perform the whitening process for the additive noise for equation (3) by assigning each value \( s_i, 1 \leq i \leq K \), to unity and normalizing the matrix \( C \) by replacing the values of \( c_i \) by \( c_i/s_i \), \( 1 \leq i \leq K \) [4]. In here, the diagonal elements \( \tau_i \) of the normalized matrix \( C \) are just the generalized singular values of the matrices \( H_Y \) and \( \hat{H}_s \). Since the noise has been whitened, we can then employ the well-developed signal-subspace speech enhancement algorithms [4,5]. After that, we have to perform the inverse whitening procedure by de-normalizing the processed diagonal elements of the matrix \( C \).

In this paper, a modified approach to handle the non-white noise case is proposed. This is based on the minimum variance (MV) estimation [8] previously developed to split the diagonal elements of the matrix \( C \) into two sets, referred to as the signal-subspace and the noise-subspace. The concept of MV estimation is to find a transformation matrix \( \hat{P} \) which minimizes the distance of two matrices \( H_y \cdot \hat{P} \) and \( H_D \), here,

\[ \hat{P} = \arg \min_{P \in \mathbb{R}^{L \times K}} \| H_y \cdot P - H_D \|_F^2, \] \hspace{1cm} (4)

where \( \| \cdot \|_F \) is the Frobenius norm of a matrix. We can then obtain the estimated matrix \( \hat{H}_D \) for the matrix \( H_D \) as follows.

\[ \hat{H}_D = H_y \cdot \hat{P} = X^T \cdot \hat{C} \cdot U, \] \hspace{1cm} (5)

where the matrices \( X \) and \( U \) are those in equation (3). The diagonal elements \( \hat{\tau}_i, 1 \leq i \leq K \), of the matrix \( \hat{C} \) is,

\[ \hat{\tau}_i = \begin{cases} \tau_i, & \frac{1 - \tau_i}{\tau_i} = \frac{c_i}{c_1}, \frac{c_i}{c_1} > 1 \\ 0, & \tau_i \leq 1 \end{cases}, \] \hspace{1cm} (6)

where \( c_i \) and \( s_i \) are those in equation (3). In this way, we can see that the signal-subspace of the matrix \( H_Y \) can be determined from equations (5) and (6).

The estimated matrix \( \hat{H}_D \) obtained here may not have the Hankel-structure. We can simply average the anti-diagonal elements of \( \hat{H}_S \) to recover the Hankel-structure and thus the enhanced speech frame,

\[ \Pi_D = \begin{bmatrix} d_0 & d_{0,1} & \cdots & d_{0,K-1} \\ d_{L,0} & d_{L,1} & \cdots & d_{L,K-1} \\ \vdots & \vdots & \ddots & \vdots \\ d_{L,M-2} & d_{L,M-1} & \cdots & d_{L,M-1} \end{bmatrix}, \] \hspace{1cm} (7)

\[ \tilde{a}(0) = [\tilde{a}_0, \tilde{a}_1, \ldots, \tilde{a}_M-2, \tilde{a}_M-1]^T. \] \hspace{1cm} (8)

### 3. EXPERIMENTAL ENVIRONMENT

The experimental environment is as follows. Sampling rate of the speech signals is 8000 Hz. The test speech includes 50 sentences, each with an average length of about 5 seconds. We choose two types of additive noise from the NOISEX-92...
database, white and Volvo car noise respectively. The latter is non-white. For white noise case, we evaluated the subjective performance of noisy speech with SNR ranging from 0dB to 30dB. For Volvo car noise case, we evaluated it from -10dB to 20dB. This is because Volvo car noise is low-passed, and therefore it could be masked by the voiced speech due to human ear perception effect. This is why we evaluated the performance for Volvo car noise case in lower SNR range only.

In our experiments for the GSVD-based approach, the frame length, M, is 256 samples with a 50% overlap, and the dimension of the Hankel-form matrices, referred to L and K, is 217 and 40 respectively. A simple silence detection algorithm was applied to extract the silence frames to be employed to construct the matrix $H_w$. The performance of the propose approach was compared with the conventional spectral subtraction algorithm and SVD-based approach.

Two objective performance measures were used here for analysis. The first is the averaged time domain signal-to-noise ratio (TDSNR), which can be formulated as follows.

$$\text{TDSNR} = 10 \cdot \log_{10} \left\{ \frac{1}{N} \sum_{n=1}^{N} d_n^2 - \frac{1}{N} \sum_{n=1}^{N} |d_n - 2\bar{d}_n|^2 \right\}, \quad (9)$$

with $d_n$ being the clean speech signal, $\bar{d}_n$ being the enhanced speech, and N being the total number of samples. The second measure is the averaged spectral distance (SD),

$$\text{SD} = 20 \cdot \log_{10} \left\{ \frac{1}{N} \sum_{k=1}^{128} \sum_{c=1}^{256} |d_{kc} - 2\bar{d}_{kc}| \right\}, \quad (10)$$

where $d_{kc}$ and $\bar{d}_{kc}$ are respectively the 256-point DFT with 128-point shift for $d_n$ and $\bar{d}_n$. Furthermore, the spectrogram and several other parameters for the processed speech were also observed and discussed.

4. EXPERIMENTAL RESULTS

Figure 1 shows the time domain signal-to-noise ratio (TDSNR) measures in equation (9) for the enhanced speech for the two types of additive noise, where ‘BSL’ is for baseline noisy speech without any processing, ‘SS’ is the conventional spectral subtraction algorithm [3], ‘SVD’ stands for SVD-MV based approach [5], and ‘GSVD’ is the proposed GSVD-based approach. In here, we do not perform the pre-whitening process for ‘SVD’ when the corrupted noise is non-white. From Figure 1a for the white noise case, we can see that GSVD is always better than SS and BSL in time-domain evaluation, regardless of whether SNR is 0dB or 30dB. This holds true for the non-white Volvo car noise in Figure 1b. In fact, for non-white noise case as in Figure 1b, GSVD significantly outperforms SS and SVD especially when the SNR is low. Therefore the proposed GSVD-based approach is shown to offer performance improvements in TDSNR measure, especially when the noise is not white and SNR is low.

Figure 2 is the spectral distance (SD) measures. From Figure 2a and 2b, it is evident that GSVD gives much smaller spectral distance than SS and SVD approach, either the noise is white or not, and the spectral distance is specially small for non-white noise. This indicates that the proposed new algorithm generates less distortion than SS and SVD in frequency domain, and this is naturally extended to human perception.

From Figures 1 and 2, we can see that GSVD also offers improvements when the SNR is 20dB or more, but SS and SVD very often does not. This implies GSVD will not degrade the intelligibility and quality of the speech signal at the presence of very light additive noise.

The spectrogram plots for clean, noisy, SS-processed, SVD-processed, and GSVD-processed speech for one typical test sentence are shown in Figure 3. The noisy speech is corrupted by Volvo car noise at -10dB of SNR. It can be observed that GSVD also offers improvements when the SNR is 20dB or more, but SS and SVD very often does not. This implies GSVD will not degrade the intelligibility and quality of the speech signal at the presence of very light additive noise.
Moreover, we found that the ratio (in dB) of the maximum generalized singular value $c_1$ to the minimum generalized singular value $c_4$, say $\sigma$, where $\sigma = c_1/c_4$, obtained in the GSVD process is much less sensitive to the SNR for different types of noise and offers a much better parameter for speech/silence detection than the relatively simple energy-based parameters. Figure 4a displays the waveform of one typical test speech utterance. We depicted its contour of $\sigma$ for various SNR for white and Volvo car noise in Figures 4b and 4c respectively. A threshold around 13dB was obtained empirically to separate the speech and silence frames very well regardless of what the SNR value is, with which the detected silence frames can be employed to track the noise statistics. Apparently this insensitivity property of $\sigma$ can be further explored in end-point detection problems for speech recognition tasks.

In addition, it is well known that the enhanced speech very often suffers some undesired interference from artificially produced noise, usually called musical noise, during the speech enhancement process. Some informal subjective listening tests were conducted, and it was found that the musical noise induced by the proposed GSVD approach is obviously less than that by SS. This is because the over-estimation of noise spectrum may cause the processed spectrum of noisy speech to be negative in the SS algorithm and thus distorts the enhanced speech. On the other hand, the minimum variance estimation criterion for determining the signal-subspace of noisy speech is closer to a linear operation, thus reducing the non-linear distortion. This is again in good consistency with the results in Figures 1 and 2.

5. CONCLUSIONS

In this paper, we proposed a generalized singular value decomposition (GSVD)-based approach for speech enhancement, which is useful regardless of whether the noise is white or not. Experimental results show that this new approach can provide better performance than the conventional spectral subtraction algorithm and SVD-based approach, especially when the additive noise is non-white. Informal subjective listening test and a few other tests also indicated attractive features of this approach.

The main drawback of GSVD-based approach is its computation load. Unlike DFT, GSVD has no fast algorithm invented presently. However, we can employ adaptive algorithms, such as the recursive least square method, to try to fast the GSVD operations, which degrade the performance as little as possible. We are now put our attentions in this direction.

6. REFERENCES