PAPER

Speech Enhancement Based on Noise Eigenspace Projection

Dongwen YING[†a)], ***Nonmember***, **Masashi UNOKI**[†b)], ***Member***, **Xugang LU**[†c)], ***Nonmember***, ***and*** **Jianwu DANG**[†d)], ***Member***

SUMMARY How to reduce noise with less speech distortion is a challenging issue for speech enhancement. We propose a novel approach for reducing noise with the cost of less speech distortion. A noise signal can generally be considered to consist of two components, a "white-like" component with a uniform energy distribution and a "color" component with a concentrated energy distribution in some frequency bands. An approach based on noise eigenspace projections is proposed to pack the color component into a subspace, named "noise subspace". This subspace is then removed from the eigenspace to reduce the color component. For the white-like component, a conventional enhancement algorithm is adopted as a complementary processor. We tested our algorithm on a speech enhancement task using speech data from the Texas Instruments and Massachusetts Institute of Technology (TIMIT) dataset and noise data from NOISEX-92. The experimental results show that the proposed algorithm efficiently reduces noise with little speech distortion. Objective and subjective evaluations confirmed that the proposed algorithm outperformed conventional enhancement algorithms.

key words: *speech enhancement, noise eigenspace, dimension reduction (DR), Karhunen-Lóeve transform (KLT)*

1. Introduction

Speech enhancement techniques attempt to improve one or more perceptual aspects of speech communication systems, namely, overall quality and intelligibility of speech sound for human or machine recognizers [1] when a signal is corrupted by noise. The improvement is in a sense of minimizing system degradation due to noise.

Numerous approaches have been proposed for this purpose. Most algorithms were presented in frequency domains such as spectral subtraction [2], [3], Wiener filtering [3], and minimum mean-square error (MMSE) [4] algorithms. They have been used widely because of their simplicity and high computational efficiency. Recently, signal subspace approaches [5]–[9] have been proposed for enhancing speech signals. The core idea of these signal subspace approaches is to decompose a noisy speech into uncorrelated components in a signal space. In such a space, each vector contains a clean signal component and a noise component. The desired signal component is estimated with a linear estimator and synthesized by applying the inverse Karhunen-Lóeve transform (KLT) based on the vectors.

In all these algorithms, the entire noise signal is reduced from noisy speech. The characteristics of noise energy distribution are not taken into account in the noise reduction process. Usually, according to the energy distribution, there are two typical noise signals, a color noise signal with a concentrated energy distribution in certain frequency bands and a white noise signal with a uniform energy distribution. In fact, most environmental noise signals exist between white noise and color noise. In other words, the environmental noise signal generally consists of a "white-like" component and a "color" component. If an appropriate strategy is adopted to deal with each component, better noise reduction is expected.

Based on such considerations, we propose a novel approach for enhancing speech, where different strategies are adopted for different components. A noise eigenspace with a high-energy packing efficiency [10] is proposed to reduce the color component, while a conventional algorithm is used as a complementary processor for processing the white-like component. The aim of this approach is to reduce the noise with as little speech distortion as possible. To track the noise variation, the noise eigenspace should be updated. Since noise signals are assumed to be more stationary than speech, the noise eigenspace does not need to be updated as often as the noisy speech eigenspace [6]. Thus, the computation load can be controlled at an acceptable level.

The rest of the paper is organized as follows. In Sect. 2, we introduce the construction of the noise eigenspace and the noise reduction approach in that space. In Sect. 3, we give details on the implementation of the proposed model for speech enhancement. In Sect. 4, we conduct some experiments for evaluating the algorithm. We focus on discussions in Sect. 5 and summarize and state the conclusions of this study in Sect. 6.

2. Noise Eigenspace and Noise Reduction

Generally speaking, noise can be efficiently reduced if we can clarify the noise energy distribution. To do so, we must find a space for noisy speech representation with two basic characteristics: (1) the distribution of noise energy is easily represented in the space and (2) noise energy is packed into a local area. The noise signal in the local area can be removed from the space to reduce noise.

The noise eigenspace is a desirable space meeting these

Manuscript received August 28, 2008.
Manuscript revised December 3, 2008.

[†]The authors are with the Information School, Japan Advanced Institute of Science and Technology, Nomi-shi, 923–1292 Japan.
a) E-mail: yingdongwen@hccl.ioa.ac.cn
b) E-mail: unoki@jaist.ac.jp
c) E-mail: xugang@jaist.ac.jp
d) E-mail: jdang@jaist.ac.jp

DOI: 10.1587/transinf.E92.D.1137

demands since it has a high-energy packing efficiency to the noise signal, and the noise energy distribution can be described by the eigenvalues. According to the energy distribution, this eigenspace can be separated into a subspace dominated by noise (referred to as "noise subspace" hereafter) and the complementary subspace dominated by speech (referred to as "speech subspace"). The noise subspace is used for noise reduction, and the speech subspace for recovering speech. The details of the process are described in the following subsections.

2.1 Construction of Noise Eigenspace

The noise eigenspace is constructed using the eigenvalue decomposition of a noise covariance matrix according to

$$\mathbf{C}_{ns}\boldsymbol{\varphi}_k = \lambda_k \boldsymbol{\varphi}_k, \quad k = 1, 2, \cdots, K, \tag{1}$$

where $\boldsymbol{\varphi}_k$ is the eigenvector corresponding to the eigenvalue λ_k, K is the number of eigenspace dimensions, and $\mathbf{C}_{ns}$ is a K-by-K matrix of noise covariance. In this study, $\mathbf{C}_{ns}$ is approximated using the noise correlation matrix (Toeplitz matrix), which is derived from the autocorrelation sequence of the noise signal. We detect speech pauses for calculating the noise autocorrelation sequence by using a voice activity detector (VAD, mentioned in Sect. 3).

The projection into the noise eigenspace can be calculated with the KLT. Supposing the noisy speech is a zero-mean signal, projecting a noisy speech in a frame, $\mathbf{y}$, (column vector) into the k-th dimension of the noise eigenspace can be represented as inner product:

$$\left\langle \mathbf{y}, \boldsymbol{\varphi}_k \right\rangle = {\boldsymbol{\varphi}_k}^T \mathbf{y}, \qquad k = 1, 2, \cdots, K \tag{2}$$

where ^{T} denotes the operation of an array transpose. If the noise is additive, a noisy speech frame vector can be described as $\mathbf{y} = \mathbf{x} + \mathbf{n}$, where $\mathbf{x}$ and $\mathbf{n}$ denote the speech and noise in the K-dimension frame vectors, respectively. For a noisy speech utterance, we can decompose it into a frame sequence, $\{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \cdots, \mathbf{y}^{(M)}\}$, where the superscripts represent the time index. When projecting this sequence into the noise eigenspace, the noise and speech energies in each dimension can be represented as

$$d_k = \frac{1}{M} \sum_{m=1}^{M} \left\langle \mathbf{n}^{(m)}, \boldsymbol{\varphi}_k \right\rangle^2 \tag{3}$$

$$s_k = \frac{1}{M} \sum_{m=1}^{M} \left\langle \mathbf{x}^{(m)}, \boldsymbol{\varphi}_k \right\rangle^2 \tag{4}$$

Since we use the noise correlation matrix instead of the covariance matrix here, there is a relationship between the projected noise energies and eigenvalues as $d_k \approx \lambda_k$, where noise eigenvalues can approximately represent the noise energy distribution. For noise mainly consisting of a color component (e.g. car noise), several large-eigenvalue dimensions usually cover most energy.

According to the definition of speech and noise energy, the dimension signal-to-noise ratio (SNR) is defined as

$$Q_k = 10\log_{10}(s_k/d_k), \qquad k = 1, \cdots, K. \tag{5}$$

Note that, in the following sections, the SNR in the noise eigenspace is referred to as dimension SNR if there is no special specification.

We know that the noise eigenspace depends on the noise signal, but it is independent of the speech signal. That is, the noise signal will be packed into a local area, namely certain large-eigenvalue dimensions; while the speech signal is not packed by this projection. Since the speech energy distribution is usually different from that of the noise signal, the large-eigenvalue dimensions are dominated by the noise signal. Therefore, a subspace containing the packed noise signal and little speech signal can be extracted from the eigenspace.

2.2 Noise Reduction in Noise Eigenspace

The criterion of extracting the noise subspace is to make the noise subspace contain as much noise, and as little speech as possible. Based on this consideration, the noise subspace should consist of low-SNR dimensions. Thus, it is necessary to estimate the dimension SNR to determine the noise subspace.

For such purpose, the speech and noise energy distribution should be investigated in the noise eigenspace. We can easily obtain the normalized noise energy, $\tilde{d}_k$, from the noise eigenvalues.

$$\tilde{d}_k = \lambda_k \Big/ \sum_{j=1}^{K} \lambda_j, \quad k = 1, 2, \cdots, K. \tag{6}$$

How to estimate the speech energy distribution in a noise eigenspace is a significant issue for estimating dimension SNR. For the sake of simplicity, we use the long-term average speech distribution to approximate the speech energy distribution. The following procedure is used to estimate the long-term average speech energy distribution in an arbitrary noise eigenspace.

First, a long-term average covariance matrix of a clean speech signal, $\mathbf{C}_{sp}$, is calculated from a large-scale clean speech dataset (Texas Instruments and Massachusetts Institute of Technology (TIMIT) test corpus including 1680 sentences). From the eigenvalue decomposition of $\mathbf{C}_{sp}$, we obtain an average speech eigenspace:

$$\mathbf{C}_{sp}\boldsymbol{\psi}_k = \gamma_k \boldsymbol{\psi}_k, \qquad k = 1, 2, \cdots, K, \tag{7}$$

where the eigenvectors, $\{\boldsymbol{\psi}_k | k = 1, \cdots, K\}$, represent this eigenspace, and the eigenvalues, $\{\gamma_k | k = 1, \cdots, K\}$, represent the average speech energy distribution in this eigenspace. The average eigenspace is a constant space, and it can be taken as prior knowledge of speech signal.

Second, the speech eigenvalues are normalized as

$$\tilde{\gamma}_k = \gamma_k \Big/ \sum_{j=1}^{K} \gamma_j, \quad k = 1, 2, \cdots, K. \tag{8}$$

Then, the normalized speech distribution in the speech eigenspace is projected into the noise eigenspace using the following equation (see details in Appendix).

$$\tilde{s}_k = \sum_{j=1}^{K} \left({\boldsymbol{\psi}_j}^T \boldsymbol{\varphi}_k \right)^2 \tilde{\gamma}_j, \quad k = 1, 2, \cdots, K, \tag{9}$$

where $\tilde{s}_k$ represents the long-term average speech energy distribution in the current noise eigenspace, $\{\boldsymbol{\varphi}_k | k = 1, \cdots, K\}$.

After obtaining the speech and noise distributions, the estimated dimension SNR, $\hat{Q}_k$, is derived as follows:

$$\hat{Q}_k = 10\log_{10}(\tilde{s}_k / \tilde{d}_k) + r, \quad k = 1, 2, \cdots, K, \tag{10}$$

where the average SNR, r, denotes the intensity relationship between speech and noise signals. The common definition of average SNR is the total energy ratio of the speech to noise.

The dimensions of the noise eigenspace are sorted in an ascending order based on the estimated dimension SNR, $\hat{Q}_1 \le \hat{Q}_2 \le \cdots \le \hat{Q}_K$. A threshold, δ_{NR}, for dimension SNR is introduced to determine how many of the low-SNR dimensions belong to the noise subspace. An appropriate threshold should guarantee that the subspace contains as much noise while including as little speech as possible. At the end, the threshold is determined by a tradeoff between the noise and speech energy in the noise subspace. From a preliminary experiment of optimizing the segmental SNR on a dataset, we found out that setting the threshold at −6 dB achieves an optimal tradeoff.

If only the dimension SNR is considered in determining the noise subspace, another problem occurs. Under low average SNR conditions, there are a great number of dimensions with (dimension) SNR less than the threshold, δ_{NR}. If all those dimensions were classified into the noise subspace, the percentage of speech energy inside the subspace (ratio of speech energy in the noise subspace to the total speech energy) would increase so that a larger speech distortion would be introduced in the noise reduction. Therefore, besides the dimension SNR, the speech energy percentage in the noise subspace should also be considered for determining the noise subspace. A threshold, δ_{SD}, for the speech energy percentage is used to prevent large speech loss. The speech energy percentage in the noise subspace should be less than δ_{SD}. It is set as 6% according to our preliminary experiments.

Let q_{NR} denotes the amount of dimensions with SNR less than δ_{NR}, and q_{SD} denotes the amount of first dimensions with an accumulated speech energy percentage less than δ_{SD} (namely $\sum_{j=1}^{q_{\text{SD}}} \tilde{s}_j / \sum_{j=1}^{K} \tilde{s}_j < \delta_{\text{SD}}$). To make the noise subspace satisfy the two conditions simultaneously, the number of the noise subspace dimensions is defined as:

$$q = \min(q_{\text{SD}}, q_{\text{NR}}). \tag{11}$$

Thus, the noise subspace can be represented by a group of eigenvectors, $[\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \cdots, \boldsymbol{\varphi}_q]$, and the remaining subspace, consisting of the last $K - q$ dimensions, $[\boldsymbol{\varphi}_{q+1}, \cdots, \boldsymbol{\varphi}_K]$ is treated as a speech subspace. The speech and noise subspace are orthogonal and implemented to each other. It is worthwhile clarifying that the subspaces defined in this paper are different from those defined in conventional subspace approaches [6]–[8]. The subspaces of conventional approaches contain theoretically either speech or noise signals, while the subspaces in this study include both speech and noise signals. As a result, our proposed approach of noise reduction differs from conventional subspace approaches.

Our approach of noise reduction is based on the two subspaces. Usually, the color component of the noise signal is contained in the noise subspace. If we directly remove the noise subspace from the noise eigenspace, the color component would be reduced with only a little loss of speech. The lost speech is responsible for "speech distortion", which is the cost of noise reduction.

After removing the noise subspace by using the DR algorithm, the projections in the speech subspace are transformed into frames in the time domain by using the following equation:

$$\mathbf{y}_{\text{SS}} = \left[\boldsymbol{\varphi}_{q+1}, \cdots, \boldsymbol{\varphi}_K \right] \times \left[\langle \mathbf{y}, \boldsymbol{\varphi}_{q+1} \rangle, \cdots, \langle \mathbf{y}, \boldsymbol{\varphi}_K \rangle \right] \tag{12}$$

In the above equation, the first item of right side is a $K \times (K - q)$ matrix, and the second item is a $(K - q) \times 1$ vector. Equation (12) can be considered as a noise-dependent filter for noise reduction. The noisy speech frame vector, $\mathbf{y}$, can be regarded as the input signal. Hereafter, the noise reduction algorithm mentioned in this section is referred to as "dimension reduction (DR)" algorithm, which includes Eqs. (6)–(12).

3. Proposed Method for Speech Enhancement

The proposed DR algorithm efficiently reduces the color component of a noise but contributes less to reducing the white-like component. Therefore, a complementary processor is needed to process the remaining white-like component after using the DR algorithm. As the MMSE algorithm [4] is reported to be a simple and efficient approach for noise reduction [11], we adopted it as the complementary processor in our study.

Figure 1 shows a block diagram of the proposed enhancement algorithm, which consists of a subspace determination block, a filter block for reducing the color component, and an MMSE block for reducing the white-like component. A VAD block for tracking noise variation is omitted in this figure.

The noise correlation matrix is calculated from the noise signal in the speech pauses. By eigenvalue decomposition of the matrix, the noise eigenspace is determined. Using Eqs. (6)–(11), the dimension SNR is derived, and then the noise eigenspace is sorted using the dimension SNR. According to the thresholds of δ_{SD} and δ_{NR}, q is obtained for determining a speech subspace. Finally, a filter for noise

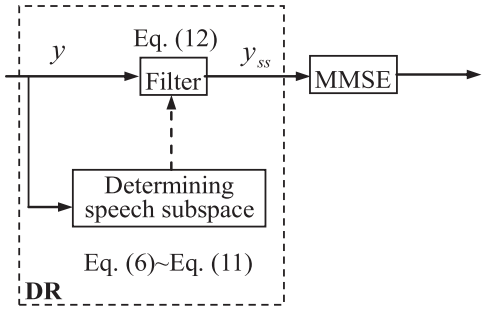

Fig. 1 Diagram of proposed method, which consists of DR block (labeled with a dashed rectangle), and MMSE block.

reduction is constructed.

The input noisy speech signal is segmented into a frame sequence, and is processed with Eq. (12) to suppress the color component of the noise signal. The output signal from the filter, $\mathbf{y}_{SS}$, is transformed into frequency domain and processed using the conventional MMSE algorithm to suppress the white-like component. Finally, the enhanced frames from the MMSE are transformed into the time domain to reconstruct the speech signal by using the overlap-and-add approach with the Hamming window [12].

As noise usually varies with time, we must track noise variation to update the noise eigenspace for efficiently reducing the noise. For this reason, a robust VAD is very important for improving system performance. Since the noise is packed into the noise subspace, the speech subspace is usually associated with a higher SNR. Therefore, we designed a robust VAD using the projections in the speech subspace. The principle of this VAD is closely related to the DR algorithm. Since the space of this paper is limited, the details of the VAD algorithm were introduced in a previous work [13].

When updating the noise eigenspace, the computational load has to be considered because the eigenvalue decomposition is a time-consuming operation. Updating the noise eigenspace more frequently is better for accurately tracking the variation of the noise signal, but it would make the computation load heavier. For a compromise, the updating period in this study was set at 1.5 seconds. As a result, the computational load could be controlled at an acceptable level.

4. Evaluation

Some experiments were conducted for a thorough evaluation of the proposed algorithm. The experiment data was selected from the TIMIT and NOISEX-92 databases. The speech data consisted of twenty utterances from the TIMIT database. Half of the utterances were from male speakers and the other half from female speakers. The noise signals selected from the NOISEX-92 database include leopard (the name of a car), tank, and babble noises. The babble noise was used because its energy distribution is similar to speech. The leopard and tank noises are two typical color noises, the

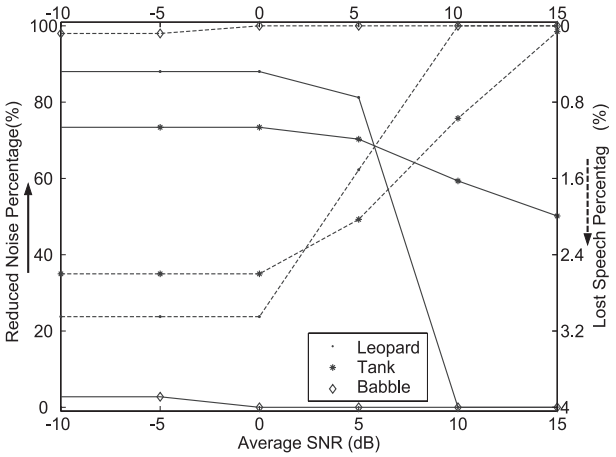

Fig. 2 Noise reduction rate and speech distortion rate under various noisy conditions. Solid lines illustrate reduced noise rate, and dashed lines for speech distortion rate.

energy distribution of which is quite different from speech. All noises were artificially added to the tested clean speech at SNRs ranging from −10 to 15 dB with an interval of 5 dB. The sampling rate was 16 kHz and a 20-ms analysis window with a 10-ms shift was used in the processing. As a result, each frame had 320 sampling-points and the noise eigenspace had the same number of dimensions as the frame length. To give a convincible evaluation, the proposed algorithm is evaluated on different aspects.

4.1 Verification of the DR Algorithm

To evaluate the noise reduction with the DR algorithm, we defined the noise reduction rate, R_{NR} (percentage rate of noise reduced with the DR algorithm) and the speech distortion rate, R_{SD} (percentage rate of lost speech caused by the DR algorithm) as follows:

$$R_{NR} = \sum_{k=1}^{q} d_k \Big/ \sum_{k=1}^{K} d_k \times 100 \quad (\%) \tag{13}$$

$$R_{SD} = \sum_{k=1}^{q} s_k \Big/ \sum_{k=1}^{K} s_k \times 100 \quad (\%) \tag{14}$$

where q is the number of the noise subspace dimensions. We evaluated the DR algorithm by the R_{NR} and R_{SD} averaged over all test samples.

Figure 2 describes the relationship between R_{NR} and R_{SD} under various noise conditions. It shows, under most color noise conditions, that the DR algorithm efficiently reduces most noise with little speech distortion. Generally speaking, the DR algorithm reduces more noise when the color noise has an energy distribution different from the speech signal. For noise with similar energy distribution to speech, such as babble noise, it is found that the DR algorithm hardly contributes to noise reduction. From Fig. 2, it was also found that the DR algorithm contributes more under low-SNR conditions than under high average SNR conditions. If the average SNR is high enough, no dimension

of the noise eigenspace will be reduced. Therefore, both the noise reduction and speech distortion are zero. For the leopard noise, R_{NR} and R_{SD} were zero when the average SNR exceeded 10 dB; and 0 dB for the babble noise.

4.2 Objective Evaluation of the Enhanced Speech

From the above experiment, one can see that the DR algorithm contributes more under low-SNR conditions. Therefore, in the following evaluations, we focused on the performance under low-SNR conditions.

First, we conducted objective evaluations, including two quantitative measures. The first measure was the segmental SNR, in dB, defined by Quackenbush [14]

$$\mathrm{SNR}_{\mathrm{Seg}} = \frac{1}{M}\sum_{m=1}^{M}\left\{10\log_{10}\sum_{k=0}^{N-1}\mathbf{x}^2(k,m) \Big/ \sum_{k=0}^{N-1}\left[\mathbf{x}(k,m)-\hat{\mathbf{x}}(k,m)\right]^2\right\}, \tag{15}$$

where M represents the number of frames containing speech signals, N is the number of samples per frame, $\mathbf{x}(k,m)$ represents the clean speech at the k-th sample in the m-th frame, and $\hat{\mathbf{x}}(k,m)$ for the enhanced speech. The second quantitative measure was the log-spectral distortion (LSD), in dB, defined by Quackenbush [14]

$$\mathrm{LSD} = \frac{1}{M}\sum_{m=1}^{M}\left\{\frac{1}{N/2+1}\sum_{k=0}^{N/2}\left[20\log_{10}\mathbf{X}(k,m) - 20\log_{10}\hat{\mathbf{X}}(k,m)\right]^2\right\}^{1/2}, \tag{16}$$

where $\mathbf{X}(k,m)$ and $\hat{\mathbf{X}}(k,m)$ are the amplitude spectra of clean and enhanced speeches at the k-th frequency component of the m-th frame, respectively. These two objective quality measures take into account both residual noise and speech distortion. The higher the segmental SNR or the lower the log-spectral distortion, the better the perceptual quality.

The performance of the proposed algorithm was evaluated by comparing two well-used algorithms. One is the generalized subspace algorithm, which is reported to be the best of all traditional subspace approaches [8]. The other is MMSE, which is used to show the improvement resulting from the proposed DR algorithm. All these algorithms in the evaluation used the same VAD as in Ying et al. [13].

Tables 1 and 2 show the improvement in segmental SNR and the LSD, respectively, where "KLT" corresponds to the conventional subspace approach [8], the proposed method to "DR+MMSE" method and MMSE to the algorithm in [4]. From the experiments, one can see that the proposed algorithm performed best out of the three algorithms. Under babble noise conditions, the DR algorithm contributed less to noise reduction. The reason is that the DR algorithm efficiently reduces most noises with different energy distributions from speech, but not so efficient for ones with similar energy distributions to speech.

In addition to segmental SNR and LSD, perceptual evaluation of speech quality (PESQ) was also used for evaluating speech quality. The PESQ is a mean opinion score (MOS)-like objective evaluation, which facilitates the objective evaluation of audio signal quality based upon perceptual criteria. This provides a quantifiable voice quality measurement that tightly correlates with voice quality as perceived by humans. The International Telecommunications Union (ITU)-PESQ algorithm converts the disturbance parameters in speech into a PESQ score, which ranges from −0.5 to 4.5. The higher the score, the better is the perceptual quality [15]. The evaluation results are listed in Table 3. The same conclusion as those from Tables 1 and 2 was obtained.

Table 1 Segmental SNR under various noise conditions (dB).

SNR	*Leopard*			*Tank*			*Babble*		
	KLT	MMSE	DR+MMSE	KLT	MMSE	DR+MMSE	KLT	MMSE	DR+MMSE
−5 dB	4.51	3.65	*5.87*	2.60	2.06	*3.50*	−0.47	−1.52	*−1.61*
0 dB	6.99	7.19	*8.52*	5.59	5.23	*6.02*	2.47	1.69	*1.69*
5 dB	9.19	10.36	*10.94*	8.70	8.53	*8.91*	5.55	5.36	*5.36*

Table 2 Log spectral distortion under various noise conditions (dB).

SNR	*Leopard*			*Tank*			*Babble*		
	KLT	MMSE	DR+MMSE	KLT	MMSE	DR+MMSE	KLT	MMSE	DR+MMSE
−5 dB	3.56	3.81	*3.67*	5.61	4.97	*4.88*	5.55	5.46	*5.55*
0 dB	2.86	3.04	*2.91*	4.41	4.24	*4.18*	4.63	4.68	*4.68*
5 dB	2.30	2.37	*2.29*	3.30	3.53	*3.47*	3.74	3.87	*3.87*

Table 3 Perceptual evaluation of speech quality.

SNR	*Leopard*			*Tank*			*Babble*		
	KLT	MMSE	DR+MMSE	KLT	MMSE	DR+MMSE	KLT	MMSE	DR+MMSE
−5 dB	1.86	1.90	*2.11*	1.49	1.73	*1.81*	1.19	1.20	*1.18*
0 dB	2.16	2.18	*2.34*	1.80	2.04	*2.07*	1.56	1.56	*1.56*
5 dB	2.42	2.46	*2.56*	2.12	2.31	*2.33*	1.89	1.88	*1.88*

Table 4 Subjective evaluation (MOS).

SNR	*Leopard*			*Tank*			*Babble*		
	KLT	MMSE	DR+MMSE	KLT	MMSE	DR+MMSE	KLT	MMSE	DR+MMSE
−5 dB	1.90	2.18	*3.27*	1.11	2.27	*2.76*	1.08	1.55	*1.53*
0 dB	2.36	2.74	*3.68*	1.53	3.07	*3.43*	1.42	1.67	*1.67*
5 dB	2.88	3.57	*4.06*	2.27	3.82	*3.78*	2.42	3.00	*3.00*

4.3 Subjective Evaluation of the Enhanced Speech

To validate the objective evaluation, subjective listening tests were performed under various SNR conditions. Twelve TIMIT sentences (subset of the sentences used for the objective evaluation) produced by three male and three female speakers were used in the listening tests. The test dataset consisted of 108 groups (12 sentences × 3 noises × 3 noise levels) and each group included three sentences processed respectively by MMSE, KLT, and the proposed algorithm. This test was taken by four female and four male subjects. They are normal Chinese speaking subjects aged 22 to 28. Each volunteer gave each test sentence a MOS score between one and five, corresponding to the subjective terms 'bad' (1), 'poor' (2), 'fair' (3), 'good' (4), and 'excellent' (5) [12]. This evaluation represented each volunteer's global appreciation of residual noise, background noise and speech distortion. The test sentences were presented to volunteers with headphones. For each speaker, the following procedure was applied: 1) clean and noisy speech signals were played before scoring, and 2) in each group, the test signals were played in random order. The results are presented in Table 4. From the subjective listening tests, we obtained similar conclusions as those from the objective evaluations.

In the above evaluations, the segmental SNR gives a quantitative description of the distance between the waveforms of enhanced and clean speech. LSD describes their distance in spectral amplitude. Both LSD and segmental SNR are the measurement of geometrical distance between clean and enhanced speech signal. The PESQ and MOS are the evaluation of opinion score. The former is produced based upon the objective perceptual criteria while the latter comes from the subjective perception of humans. Summarizing all the above evaluations, although the proposed algorithm is not always the best on all measurements, it is outstanding on most measurements. Especially, the subjective MOS evaluations showed clearly that the proposed algorithm demonstrated the best performance. Accordingly, we can conclude that the proposed algorithm performs better than conventional algorithms.

5. Discussion

As described in the above sections, the proposed algorithm shows advantages under color noise conditions. This section investigates the reasons why the proposed method performs well for color noises. Since the "color" can be easily described in the frequency-domain, it is better to investigate what happens in the frequency domain when the DR algorithm is applied in the noise eigenspace.

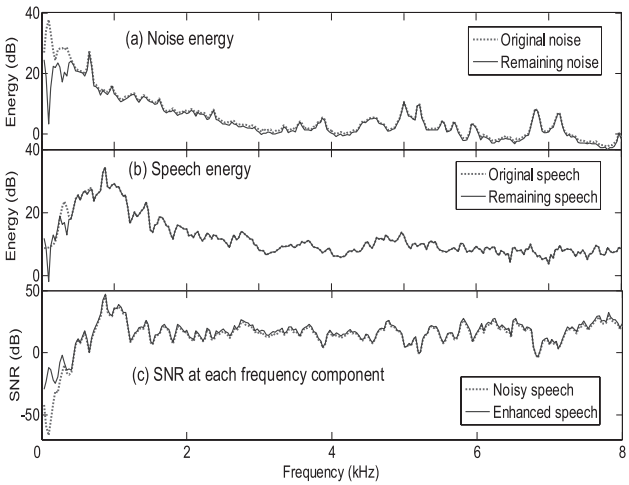

Fig. 3 Energy changes of speech and noise in Fourier space under tank noise conditions.

Here, we use noisy speech utterances to investigate the energy variation of speech and noise in frequency domain. Two typical color noises, tank and leopard, were used. They are respectively added to a clean utterance at average SNR of 0 dB to obtain two noisy speech utterances. By applying the DR algorithm to the leopard noisy utterance, 89.2% of the noise was reduced and 2.1% of speech distortion was introduced. For the utterance contaminated by the tank noise, the noise reduction rate and speech distortion were 74.7% and 2.7%, respectively. This result shows that the DR algorithm produces better results under leopard noise conditions. The reason is that there was more of a color component in the leopard noise than in the tank noise as showned in Fig. 3 (a) and Fig. 4 (a). As a result, the DR algorithm reduces more noise in leopard noise conditions.

After applying the DR algorithm, the signal in the speech subspace was transformed into the frequency domain using fast Fourier transform (FFT). Then, we compare the difference of energy distribution before and after DR algorithm. Figure 3 illustrates the speech and noise energy changes in the frequency domain under tank noise conditions. From Fig. 3 (a), one can see that the noise signal in the low-frequency domain was significantly reduced. At the same time, the speech signal was affected a little, as shown in Fig. 3 (b), which means the speech distortion is controlled at a low level. As a result, the SNR in the low-frequency domain was greatly improved. Figure 4 shows the energy changes under leopard noise conditions. One can see that the noise was generally reduced in almost the whole fre-

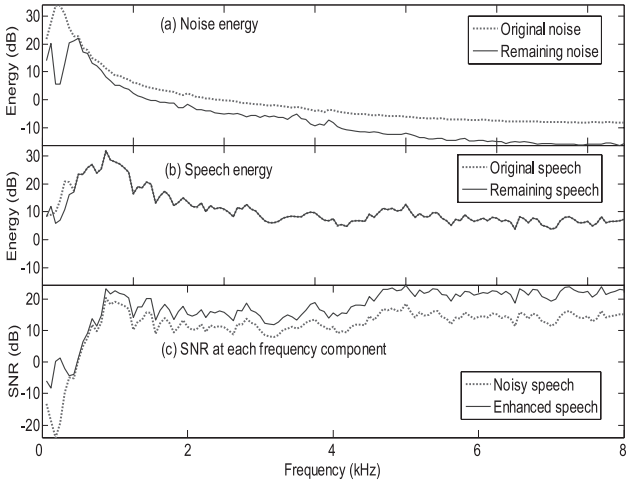

Fig. 4 Energy changes of speech and noise in Fourier space under leopard noise conditions.

quency band, while the speech signal was affected a little, as seen in Fig. 4 (b). Thus, the SNR improved in most frequency components. These figures compare the energy distribution difference before and after DR algorithm. From these comparisons, we can see that the speech distortion can be controlled at a low level in the process of noise reduction. The same phenomena were also observed with other types of noises. In contrast, conventional algorithms do not have such an intrinsic mechanism to control speech distortion.

From a signal decomposition point of view, the DR algorithm decomposes the input noisy speech signal into two sub-signals, as shown in the following equations.

$$\mathbf{y} = \mathbf{y}_{SS} + \mathbf{y}_{NS} \tag{17}$$

$$\mathbf{y}_{NS} = [\boldsymbol{\varphi}_1, \cdots, \boldsymbol{\varphi}_q] \times [<\mathbf{y}, \boldsymbol{\varphi}_1>, \cdots, <\mathbf{y}, \boldsymbol{\varphi}_q>] \tag{18}$$

where $\mathbf{y}_{NS}$ is the sub-signal from the noise subspace and $\mathbf{y}_{SS}$ is the sub-signal from the speech subspace. $\mathbf{y}_{NS}$ includes most of the color component and $\mathbf{y}_{SS}$ contains most of the white-like component. In the proposed algorithm, $\mathbf{y}_{NS}$ is set to zero. In other words, the DR algorithm actually removes a sub-signal with extremely low SNR to reduce the color component of the noise signal.

Conventional algorithms (such as MMSE, spectral subtraction, or KLT) reduce the two sub-signals as a whole. This means, for details of algorithm used by Ephraim [4], not only $\mathbf{y}_{SS}$ but also $\mathbf{y}_{NS}$ are processed with MMSE. We know that $\mathbf{y}_{NS}$ contains little speech, and its actual posterior SNR is extremely low. The theoretical output from $\mathbf{y}_{NS}$ should approach zero. However, the MMSE algorithm is unreliable for processing low-SNR signals. In practical cases, its output from $\mathbf{y}_{NS}$ is not normally zero so that it introduces extra residual noise. The same thing will happen to MMSE and other conventional algorithms. If $\mathbf{y}_{NS}$ is directly cleaned up, the residual noise from $\mathbf{y}_{NS}$ can be avoided and little speech distortion introduced. In this sense, we can say that the DR algorithm efficiently suppresses residual noise as well as prevents significant speech distortion. Therefore, the DR algorithm is more efficient than conventional algorithms for processing the color component.

6. Conclusion and Future Perspectives

Our aim was to develop a speech enhancement algorithm based on noise eigenspace projections. The noise is reduced by using two different strategies: one for removing the white-like components and the other for removing the color components. Experimental evaluations showed that the proposed DR algorithm reduces noise with less speech distortion. It efficiently suppresses residual noise as well as prevents significant speech distortion. By combining DR with MMSE, an efficient speech enhancement can be realized.

There is still room for further improvement in the performance of the proposed algorithm. When removing the noise subspace, a little speech signal is unavoidable to be lost. The speech distortion is expected to be further minimized by recovering the lost speech signal in the noise subspace. This issue will be the aim of a future study.

Acknowledgements

This study is mainly supported by the program for the "Fostering Talent in Emergent Research Fields" in Special Coordination Funds for Promoting Science and Technology by the Ministry of Education, Culture, Sports, Science and Technology. This work was partially supported by a Grant Program of the Yazaki Memorial Foundation for Science. It was also supported by the Strategic Information and Communications R&D Promotion Program (SCOPE) (071705001) of the Ministry of Internal Affairs and Communications (MIC), Japan.

References

[1] Y. Ephraim, H. Lev-Ari, and W.J.J. Roberts, "A brief survey of speech enhancement," The Electronic Handbook, pp.2088–2097, CRC Press, 2005.
[2] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-27, no.2, pp.113–120, April 1979.
[3] P. Scalart and J.V. Filho, "Speech enhancement based on a priori signal to noise estimation," Proc. ICASSP1996, vol.2, pp.629–632, Atlanta, USA, May 1996.
[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short time spectral amplitude estimator," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-32, pp.1109–1121, Dec. 1984.
[5] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: A regenerative approach," Speech Commun., vol.10, no.1, pp.45–57, Feb. 1991.
[6] Y. Ephraim, "A signal subspace approach for speech enhancement," IEEE Trans. Speech Audio Process., vol.3, no.4, pp.251–266, July 1995.
[7] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," IEEE Trans. Speech Audio Process., vol.9, no.2, pp.87–94, Feb. 2001.
[8] Y. Hu and P. Loizou, "A generalized subspace approach for enhancing speech corrupted with colored noise," IEEE Trans. Speech Audio Process., vol.11, no.4, pp.334–341, July 2003.

[9] K. Hermus, P. Wambacq, and H.V. Hamme, "A review of signal subspace speech enhancement and its application to noise robust speech recognition," EURASIP J. Advances in Signal Processing, vol.2007, no.1, pp.195–210, 2007.
[10] P. Yip and K.R. Rao, "Energy packing efficiency for generalized discrete transforms," IEEE Trans. Commun., vol.COM-26, no.6, pp.1257–1262, Aug. 1978.
[11] Y. Hu and P. Loizou, "Subjective evaluation and comparison of speech enhancement algorithms," Speech Commun., vol.49, pp.588–601, 2007.
[12] J.R. Deller, J. Hansen, and J.G. Proakis, Discrete-Time Processing of Speech Signals, IEEE Press, New York, 2000.
[13] D. Ying, Y. Shi, X. Lu, J. Dang, and F. Soong, "Robust voice activity detection based on noise eigenspace," Acoust. Sci. and Tech., vol.28, no.6, pp.413–423, 2007.
[14] S. Quackenbush, T. Barnwell, and M. Clements, Objective Measures of Speech Quality, Prentice-Hall, Englewood Cliffs, NJ, 1988.
[15] ITU-T Rec. P.862, Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs, International Telecommunications Union, Geneva, Switzerland, 2001.
[16] X.D. Zhang, Matrix analysis and applications, Tsinghua & Springer, Beijing, 2004.

Appendix

In this appendix, we prove that the mapping function of speech energy distribution from the speech eigenspace into the arbitrary noise eigenspace can be represented as Eq. (9). On one hand, according to the projection theorem [16], the speech signal in a frame, $\mathbf{x}$, can be represented using the speech eigenvector, $\boldsymbol{\psi}_j$.

$$\mathbf{x} = \sum_{j=1}^{K} \left\langle \mathbf{x}, \boldsymbol{\psi}_j \right\rangle \boldsymbol{\psi}_j. \quad (\mathrm{A}\cdot 1)$$

On the other hand, the projection of the speech frame vector $\mathbf{x}$ can be written as

$$\left\langle \mathbf{x}, \boldsymbol{\varphi}_k \right\rangle = \boldsymbol{\varphi}_k{}^T \mathbf{x}. \quad (\mathrm{A}\cdot 2)$$

Substituting Eq. (A· 1) into Eq. (A· 2), we obtain

$$\left\langle \mathbf{x}, \boldsymbol{\varphi}_k \right\rangle = \sum_{j=1}^{K} \mathbf{x}^T \boldsymbol{\psi}_j \boldsymbol{\varphi}_k{}^T \boldsymbol{\psi}_j. \quad (\mathrm{A}\cdot 3)$$

The expectation of projection energy of the speech signal in the k^{th} dimension of the noise eigenspace is represented as

$$s_k = E\left[\left\langle \mathbf{x}, \boldsymbol{\varphi}_k \right\rangle\right]^2, \quad (\mathrm{A}\cdot 4)$$

where $E(\cdot)$ is the expectation operator. Substituting Eqs. (A· 2) and (A· 3) into Eq. (A· 4), we get

$$\begin{aligned} s_k &= E\left(\boldsymbol{\varphi}_k{}^T \mathbf{x} \sum_{j=1}^{K} \mathbf{x}^T \boldsymbol{\psi}_j \boldsymbol{\varphi}_k{}^T \boldsymbol{\psi}_j \right) \\ &= E\left(\sum_{j=1}^{K} \boldsymbol{\varphi}_k{}^T \mathbf{x}\mathbf{x}^T \boldsymbol{\psi}_j \boldsymbol{\varphi}_k{}^T \boldsymbol{\psi}_j \right) \\ &= \sum_{j=1}^{K} \boldsymbol{\varphi}_k{}^T E(\mathbf{x}\mathbf{x}^T) \boldsymbol{\psi}_j \boldsymbol{\varphi}_k{}^T \boldsymbol{\psi}_j \\ &= \sum_{j=1}^{K} \boldsymbol{\varphi}_k{}^T \mathbf{C}_{sp} \boldsymbol{\psi}_j \boldsymbol{\varphi}_k{}^T \boldsymbol{\psi}_j. \end{aligned} \quad (\mathrm{A}\cdot 5)$$

Substituting Eq. (7) into the Eq. (A· 5), we obtain

$$\begin{aligned} s_k &= \sum_{j=1}^{K} \boldsymbol{\varphi}_k{}^T \boldsymbol{\psi}_j \gamma_j \boldsymbol{\psi}_j^T \boldsymbol{\varphi}_k \\ &= \sum_{j=1}^{K} (\boldsymbol{\varphi}_k{}^T \boldsymbol{\psi}_j)^2 \gamma_j, \quad k = 1, 2, \cdots, K. \end{aligned} \quad (\mathrm{A}\cdot 6)$$

Since the projection is a unit-orthogonal transform, according to the law of energy conservation, we obtain $\sum_{j=1}^{K} s_j = \sum_{j=1}^{K} \gamma_j$. The left and right side of Eq. (A· 6) are respectively divided by $\sum_{j=1}^{K} s_j$ and $\sum_{j=1}^{K} \gamma_j$. Then, Eq. (A· 6) can be written as Eq. (9).

Dongwen Ying received his B.E. degree in 1998 and his M.E. degree in 2000, respectively, from Harbin Institute of Technology. He earned his Ph.D. from the School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), in 2007. He is currently working at the Institute of Acoustics, Chinese Academy of Sciences as an associate researcher. His major interests are speech enhancement and robust speech recognition.

Masashi Unoki received his M.S. and Ph.D. (Information Science) from the Japan Advanced Institute of Science and Technology (JAIST) in 1996 and 1999, respectively. His main research interests are auditory-motivated signal processing and the modeling of auditory systems. He was a JSPS research fellow from 1998 to 2001. He was associated with the Advanced Telecommunications Research Institute (ATR) Human Information Processing Laboratories as a visiting researcher during 1999–2000, and from 2000 to 2001 he was a visiting research associate at the Centre for the Neural Basis of Hearing (CNBH) in the Dept. of Physiology at the University of Cambridge. He has been on the faculty of the School of Information Science at JAIST since 2001, and he is now an associate professor. He is a member of the Research Institute of Signal Processing (RISP), the Acoustical Society of America (ASA), the Acoustical Society of Japan (ASJ), the Institute of Electrical and Electronic Engineering (IEEE), and the International Speech Communication Association (ISCA). Dr. Unoki received the Sato Prize for an Outstanding Paper from the ASJ in 1999 and the Yamashita Taro Prize for Young Researchers from the Yamashita Taro Research Foundation in 2005.

Xugang Lu received the B.S. and M.S. from Harbin Institute of Technology, China, in 1994 and 1996, respectively. He earned his Ph.D. from the National Laboratory of Pattern Recognition, Chinese Academy of Science in 1999. He joined Nanyang Technological University, Singapore, as a research fellow after Oct. 1999. Since Dec. 2001, he was a Postdoctoral fellow at McMaster University, Canada. After April 2003, he has been with the faculty of the School of Information Science of Japan Advanced Institute of Science and Technology, and now is an Assistant professor. Dr. Lu received the President award from the Chinese Academy of Science in 1999. He is a member of the Acoustic Society of Japan. His main research interests include speech signal processing and recognition, statistic learning and pattern recognition, and computational auditory model.

Jianwu Dang received his B.E. and M.S. degrees from Tsinghua University, China in 1982 and 1984, respectively. He worked at Tianjin University as a lecturer from 1984 to 1988. He earned his Ph.D. Eng. from Shizuoka Univ., Japan in 1992. Dr. Dang worked for ATR Human Information Processing Lab., Japan from 1992 to 2001. He joined the University of Waterloo, Canada, as a visiting scholar for one year from 1998. He joined Japan Advanced Institute of Science and Technology (JAIST) in 2001 and is currently a professor there. He joined the Institute of Communication Parlee, Center of National Research Scientific (CNRS), France, as a research scientist the first class for one year beginning in 2002. His research interests are in all of the fields of speech science, especially in speech production. He is a member of the Acoustic Societies of America and Japan, and also a member of the Institute of Image Information and Television Engineers.