SIGNAL SUBSPACE SPEECH ENHANCEMENT FOR AUDIBLE NOISE REDUCTION

Chang Huai You[+], *Soo Ngee Koh*[*], *Susanto Rahardja*[+]

[+]Institute for Infocomm Research, Singapore 119613
[*],[+] Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798

ABSTRACT

A novel subspace-based speech enhancement scheme based on a criterion of audible noise reduction is considered. Masking properties of the human auditory system is used to define the audible noise quantity in the eigen-domain. Subsequently, an audible noise reduction scheme is developed based on a signal subspace technique. We derive the eigen-decomposition of the estimated speech autocorrelation matrix with the assumption of white noise and outline the implementation of our proposed scheme. We further extend the scheme to the colored noise case. Simulation results show that our proposed scheme outperforms many existing subspace methods in terms of segmental signal-to-noise ratio (SNR), perceptual evaluation of speech quality (PESQ) and informal listening tests.

1. INTRODUCTION

Spectral domain speech enhancement always suffers from tonal noise problem to a varying degree. While autoregressive and hidden Markov models have been shown to be very effective in speech coding and recognition of clean speech signals, they have not been found to be sufficiently effective for speech enhancement. Subspace-based approach, however, has been shown to be largely useful for reducing wideband additive noise. The underlying principle of the subspace method is to decompose the vector space of a noisy signal into a signal-plus-noise subspace and a noise subspace. Enhancement is then performed by removing the noise subspace and estimating the clean speech from the remaining signal-plus-noise subspace [1].

Though subspace methods have been shown to have a number of advantages in speech enhancement, psychoacoustic properties of the human auditory system have not been sufficiently exploited in these methods to date. Masking properties that researchers used for speech coding and enhancement are usually exploited through the critical frequency domain. In [2], a spectral domain perceptual post-filter is applied to the output of the signal subspace filter to improve the subjective quality of the enhanced speech. However, it cannot help to preserve the weak speech components which have already been erased by the conventional subspace method. In [3], a frequency to eigen-domain transformation is introduced so as to apply masking properties in subspace speech enhancement. It computes the power spectral density with respect to the estimated eigenvalues and eigenvectors according to Blackman-Tuckey cross-spectral analysis with a Bartlett window. However, in the conventional perceptual properties-based subspace speech enhancement methods, the audible and inaudible components are neither separated nor individually analyzed, which is in contrast to the frequency domain enhancement approach proposed in [5].

Based on the knowledge of the human auditory system, we postulate that the audible noise quantity in the eigen-domain can be suppressed without significantly distorting the underlying speech signal. The organization of the paper is as follows. Section 2 introduces the subspace enhancement method. Our proposed audible noise reduction scheme based on perceptual masking is presented in section 3. Section 4 provides the simulation results. Section 5 concludes this paper.

2. SUBSPACE SPEECH ENHANCEMENT

A noisy speech sequence $\mathbf{x}(l)$=[$x(hl-K+1), x(hl-K+2), \cdots, x(hl-1), x(hl)$]$^{\mathrm{T}}$ is assumed as the sum of a clean speech sequence $\mathbf{s}(l)$=[$s(hl-K+1), s(hl-K+2), \cdots, s(hl-1), s(hl)$]$^{\mathrm{T}}$ and a white noise sequence $\mathbf{w}(l)$=[$w(hl-K+1), w(hl-K+2), \cdots, w(hl-1), w(hl)$]$^{\mathrm{T}}$ with variance σ_w^2, i.e.,

$$\mathbf{x}(l) = \mathbf{s}(l) + \mathbf{w}(l) \tag{1}$$

where l, h and K denote the frame number, frame advance size in sample and the vector dimension respectively. Let R_s denote the autocorrelation matrix of clean speech. The eigen-decomposition of R_s is given by

$$R_s = U\Lambda_s U^{\mathrm{T}} \tag{2}$$

where $\Lambda_s = \mathrm{diag}[\lambda_{s,0}, \cdots, \lambda_{s,K-1}]$ and $U = [\mathbf{u}_0, \cdots, \mathbf{u}_{K-1}]$ are the eigenvalue matrix and the orthonormal eigenvector matrix of R_s respectively.

The spectral domain constrained (SDC) estimator proposed by the Ephraim $et\ al$ [1] can be written as

$$H = UGU^{\mathrm{T}} \tag{3}$$

where a possible solution of the SDC estimator is that G is a diagonal matrix given by

$$G = \Lambda_s(\Lambda_s + \sigma_w^2\Lambda_\mu)^{-1}. \quad (4)$$

Λ_μ = diag[$\mu_0, \cdots, \mu_{K-1}$] is a diagonal matrix of Lagrange multipliers when deriving the estimator H based on the criterion of minimizing the speech signal distortion with the permissible residual noise level [1]. Subsequently the k-th diagonal element of G is given by

$$g_k = \frac{\lambda_{s,k}}{\lambda_{s,k} + \mu_k\sigma_w^2}, \quad k = 0, \cdots, K-1. \quad (5)$$

3. PROPOSED AUDIBLE NOISE REDUCTION

3.1. Algorithm

Since the estimated speech signal is given by $\hat{\mathbf{s}}(l) = H\mathbf{x}(l)$, we can derive the autocorrelation matrix of the estimated speech signal, $R_{\hat{s}}$, as follows:

$$\begin{aligned} R_{\hat{s}} &= \mathrm{E}\{\hat{\mathbf{s}}(l)\hat{\mathbf{s}}(l)^{\mathrm{T}}\} = \mathrm{E}\{H\mathbf{x}(l)\mathbf{x}(l)^{\mathrm{T}}H^{\mathrm{T}}\} \\ &= \mathrm{E}\{H[\mathbf{s}(l) + \mathbf{w}(l)][\mathbf{s}(l) + \mathbf{w}(l)]^{\mathrm{T}}H^{\mathrm{T}}\} \\ &= H(R_s + \sigma_w^2\mathrm{I})H^{\mathrm{T}} = UGU^{\mathrm{T}}(R_s + \sigma_w^2\mathrm{I})UG^{\mathrm{T}}U^{\mathrm{T}} \\ &= UGU^{\mathrm{T}}(U\Lambda_s U^{\mathrm{T}} + U\sigma_w^2U^{\mathrm{T}})UG^{\mathrm{T}}U^{\mathrm{T}} \\ &= UG(\Lambda_s + \sigma_w^2\mathrm{I})G^{\mathrm{T}}U^{\mathrm{T}} = U\Lambda_{\hat{s}}U^{\mathrm{T}}. \end{aligned} \quad (6)$$

Substituting (4) into (6), we have $\Lambda_{\hat{s}}$ given by

$$\Lambda_{\hat{s}} = \Lambda_s(\Lambda_s + \sigma_w^2\Lambda_\mu)^{-1}(\Lambda_s + \sigma_w^2\mathrm{I})[\Lambda_s(\Lambda_s + \sigma_w^2\Lambda_\mu)^{-1}]^{\mathrm{T}}. \quad (7)$$

Obviously, $\Lambda_{\hat{s}}$ is a diagonal matrix with its k-th diagonal element $\lambda_{\hat{s},k}$ given by

$$\lambda_{\hat{s},k} = \frac{\lambda_{s,k}^2}{(\lambda_{s,k} + \mu_k\sigma_w^2)^2}(\lambda_{s,k} + \sigma_w^2). \quad (8)$$

Hence, by the definition of eigen-decomposition, we note that $\lambda_{\hat{s},k}$ is actually the eigenvalue of the autocorrelation matrix of the estimated speech, and U is also the eigenvector matrix of the autocorrelation matrix of the estimated speech.

The main objective of psychoacoustic signal enhancement is to reduce the audible noise. Provided that the masking threshold of the clean speech signal is known, those noisy components below this threshold is inaudible due to the effect of masking.

We define the masking-eigenvalues, λ_t, as a function of masking threshold according to the frequency eigen-domain transformation. Consequently, we define the audible eigenvalues of the autocorrelation matrices of the clean speech and the estimated speech as the maximum values of their individual eigenvalues and the corresponding masking-eigenvalue respectively, i.e.,

$$\vartheta_{s,k} = \max(\lambda_{s,k}, \quad \lambda_{t,k}) \quad (9)$$

$$\vartheta_{\hat{s},k} = \max(\lambda_{\hat{s},k}, \quad \lambda_{t,k}). \quad (10)$$

We define the error between the two audible eigenvalues in (9) and (10) as the residual audible eigen-noise, i.e.,

$$\varepsilon(k) = \vartheta_{\hat{s},k} - \vartheta_{s,k}, \quad k = 0, \cdots, K-1. \quad (11)$$

Substituting (9) and (10) into (11), the residual audible eigen-noise can be further expressed as

$$\varepsilon(k) = \begin{cases} \lambda_{\hat{s},k} - \lambda_{s,k}, & \text{if } \lambda_{\hat{s},k} > \lambda_{t,k} \;\&\; \lambda_{s,k} > \lambda_{t,k} \text{ (I)} \\ \lambda_{\hat{s},k} - \lambda_{t,k}, & \text{if } \lambda_{\hat{s},k} > \lambda_{t,k} \;\&\; \lambda_{s,k} \le \lambda_{t,k} \text{ (II)} \\ \lambda_{t,k} - \lambda_{s,k}, & \text{if } \lambda_{\hat{s},k} \le \lambda_{t,k} \;\&\; \lambda_{s,k} > \lambda_{t,k} \text{ (III)} \\ 0, & \text{if } \lambda_{\hat{s},k} \le \lambda_{t,k} \;\&\; \lambda_{s,k} \le \lambda_{t,k} \text{ (IV).} \end{cases} \quad (12)$$

Since we only have the noisy speech available, and the estimated speech is obtained through modifying the noisy signal, the corresponding condition of (12) can be extended to the comparisons between the eigenvalues of the autocorrelation matrices of the noisy and clean speech and the masking-eigenvalue. Therefore, we define the modified residual audible eigen-noise as

$$\tilde{\varepsilon}(k) = \begin{cases} \lambda_{\hat{s},k} - \lambda_{s,k}, & \text{if } \lambda_{x,k} > \lambda_{t,k} \;\&\; \lambda_{s,k} > \lambda_{t,k} \text{ (I)} \\ \lambda_{\hat{s},k} - \lambda_{t,k}, & \text{if } \lambda_{x,k} > \lambda_{t,k} \;\&\; \lambda_{s,k} \le \lambda_{t,k} \text{ (II)} \\ \lambda_{t,k} - \lambda_{s,k}, & \text{if } \lambda_{x,k} \le \lambda_{t,k} \;\&\; \lambda_{s,k} > \lambda_{t,k} \text{ (III)} \\ 0, & \text{if } \lambda_{x,k} \le \lambda_{t,k} \;\&\; \lambda_{s,k} \le \lambda_{t,k} \text{ (IV).} \end{cases} \quad (13)$$

It should be noted that the main difference between the modified residual audible eigen-noise and the discrepancy between the eigenvalues of the noisy and clean speech is that the former has a smaller dynamic range than the latter as a result of involving λ_t. Consequently, it results in more emphasis of audible noise in the processing algorithm. As an eigenvalue represents the power component along its corresponding eigenvector and, in the additive white noise case, the eigenvalue for noisy signal is the sum of the corresponding eigenvalues of clean speech and noise, it is therefore always greater than that of clean speech. The goal of enhancement algorithm is to modify the noisy eigenvalue by attenuating it to obtain the estimated eigenvalue for clean speech. Thus, the objective of our proposed enhancement method is to make the modified residual audible eigen-noise less than or equal to zero, i.e.,

$$\tilde{\varepsilon}(k) \le 0, \quad k = 0, \cdots, K-1. \quad (14)$$

When $\tilde{\varepsilon}(k)$ is negative, it means that either the speech eigenvalue is underestimated (condition of (13, I)) or the speech eigenvalue is correctly estimated to be below the threshold (condition of (13, II)) and is hence inaudible. Observing (13), we may give the following conclusions for each condition:

- (13, I and II): $\tilde{\varepsilon}(k)$ may be positive, negative or zero, depending on the values of $\lambda_{\hat{s},k}$, $\lambda_{s,k}$ and $\lambda_{t,k}$
- (13, III): $\tilde{\varepsilon}(k)$ is always negative or zero

- (13, IV): $\tilde{\varepsilon}(k)$ is zero.

Consequently, (13, III) and (13, IV) will not be affected by the introduction of $\lambda_{\hat{s},k}$, wherein $\lambda_{x,k} \leq \lambda_{t,k}$. Only (13, I) and (13, II) need to be modified, wherein $\lambda_{x,k} > \lambda_{t,k}$. Considering (5), it is noted that the Lagrange multipliers play an important role on the estimation of speech signal. The value of μ_k ($k = 0, \cdots, K-1$) regulates the trade-off between the residual noise and speech distortion. As discussed in [1], when μ_k increases from zero to infinity, the level of the residual noise decreases whilst the level of signal distortion increases. An approach for choosing μ that compromises between signal distortion and residual noise is to make μ dependent on the SNR in each frame [6]. A simpler approach is to set μ_k to a fixed value in the time domain constrain (TDC) method proposed in [1], but it leads to a suboptimal selection of μ_k. To overcome this drawback, we propose to adapt the μ_k value based on the criterion of audible noise reduction defined in (14).

As indicated by the above analysis, the condition of (14) is always satisfied in the case of $\lambda_{x,k} \leq \lambda_{t,k}$ and we should therefore keep the eigenvalue unchanged, i.e., $\lambda_{\hat{s},k} = \lambda_{x,k}$. According to (8), we then have

$$\frac{\lambda_{s,k}^2}{(\lambda_{s,k} + \mu_k \sigma_w^2)^2}(\lambda_{s,k} + \sigma_w^2) = \lambda_{x,k}. \tag{15}$$

Since $\lambda_{x,k} = \lambda_{s,k} + \sigma_w^2$ for white additive noise, it can be seen that $\mu_k = 0$.

In the case of $\lambda_{x,k} > \lambda_{t,k}$, according to the constraint of (14), substituting (8) into (13, I) and (13, II) respectively yields

$$\begin{aligned} \mu_{I,k} &\geq \Big(\sqrt{\frac{\lambda_{s,k} + \sigma_w^2}{\lambda_{s,k}}} - 1\Big)\frac{\lambda_{s,k}}{\sigma_w^2}, \quad \lambda_{s,k} > \lambda_{t,k} \\ \mu_{II,k} &\geq \Big(\sqrt{\frac{\lambda_{s,k} + \sigma_w^2}{\lambda_{t,k}}} - 1\Big)\frac{\lambda_{s,k}}{\sigma_w^2}, \quad \lambda_{s,k} \leq \lambda_{t,k} \end{aligned} \tag{16}$$

where $\mu_{I,k}$ and $\mu_{II,k}$ represent the μ_k value in (13, I) and (13, II), respectively. We use the masking-eigenvalue in place of *a priori* eigenvalue ($\lambda_{s,k}$) required in the square root part of (16) and $\mu_{t,k} = (\sqrt{\frac{\lambda_{t,k}+\sigma_w^2}{\lambda_{t,k}}} - 1)\frac{\lambda_{s,k}}{\sigma_w^2}$ to replace both $\mu_{I,k}$ and $\mu_{II,k}$. It can be proven that $\mu_{t,k}$ satisfies both the conditions of (16) simultaneously. In other words, it implies that $\mu_{t,k}$ satisfies the criterion of (14) under both the conditions of (13, I) and (13, II) simultaneously. Thus the estimate of μ_k may be determined by

$$\hat{\mu}_k = \Big(\sqrt{\frac{\lambda_{t,k} + \sigma_w^2}{\lambda_{t,k}}} - 1\Big)\frac{\lambda_{s,k}}{\sigma_w^2} + \delta \tag{17}$$

where δ is a nonnegative value, which is empirically set to 3.8 to give the best subjective quality for enhanced speech.

3.2. Implementation

The relationship between eigenvalues and power spectral density (PSD) can be found through an autocorrelation function. In practice, a speech signal is analyzed in the short term. A short-term PSD definition through Bartlett windowed autocorrelation is given as follows [3]

$$\psi_s(p) = \sum_{k=-K+1}^{K-1} (1 - \frac{|k|}{K}) r_s(k) e^{-j2\pi pk/P}, \quad p = 0, \cdots, P-1 \tag{18}$$

where P denotes the number of frequency bins, and $r_s(k)$ is the autocorrelation function with lag k. Based on the above definition, the following relationship is satisfied [3]

$$\psi_s(p) = \frac{1}{K}\sum_{k=1}^{K} \lambda_{s,k} |V_k(p)|^2. \tag{19}$$

where $V_k(p)$ is the p-th frequency bin of the Fourier transform of the k-th orthonormal eigenvector, $\mathbf{u}_k$=[$u_k(0)$ $u_k(1)$, $\cdots$, $u_k(K-1)$]$^{\mathrm{T}}$, of the autocorrelation matrix, and it is defined as

$$V_k(p) = \sum_{i=0}^{K-1} u_k(i) e^{-j2\pi pi/P}. \tag{20}$$

With ψ_s, we can obtain the masking threshold ψ_t according to [[7], pp. 317-319]. Using the method proposed in [3], the masking-eigenvalues, λ_t, can be approximately obtained based on the masking-threshold, ψ_t, as follows:

$$\lambda_{t,k} \approx \frac{1}{P}\sum_{p=0}^{P-1} \psi_t(p) |V_k(p)|^2, \quad k = 0, \cdots, K-1. \tag{21}$$

Next, we give the procedure for the implementation of the proposed algorithm

1. Estimation of σ_w^2, followed by pre-estimation of R_s by $\hat{R}_s = R_x - \hat{\sigma}_w^2 \mathrm{I}$
2. Perform eigen-decomposition of $\hat{R}_s$ according to (2)
3. Assuming that the eigenvalues of R_s are ordered as $\lambda_{s,0} \geq \lambda_{s,1} \geq \cdots \geq \lambda_{s,K-1}$, determine the dimensionality of the speech signal subspace as follows:
$$M = \arg \max_{k \in [0,\cdots,K-1]} \big\{k \mid \lambda_{s,k} > 0\big\} \tag{22}$$
4. Compute $V_i(p)$ according to (20) and the pre-estimated PSD of speech signal $\hat{\psi}_s$ according to (19) with K being changed into M, i.e.,
$$\hat{\psi}_s(p) = \frac{1}{K}\sum_{i=1}^{M} \lambda_i |V_i(p)|^2 \tag{23}$$
5. Use $\hat{\psi}_s$ to obtain the masking threshold $\hat{\psi}_t$

6. Use $\hat{\psi}_t$ to obtain $\lambda_{t,k}$ according to (21)

7. Compute μ_k according to (17) when $\lambda_{x,k} > \lambda_{t,k}$, and $\mu_k = 0$ when $\lambda_{x,k} \leq \lambda_{t,k}$

8. Compute g_k according to (5), with the following modification

$$g_k = \begin{cases} \frac{\lambda_{s,k}}{\lambda_{s,k}+\mu_k \sigma_w^2}, & k = 0, \cdots, M-1 \\ 0, & k = M, \cdots, K-1 \end{cases} \quad (24)$$

9. Compute H according to (3) and then estimate the enhanced speech signal by $\hat{\mathbf{s}}(l) = H\mathbf{x}(l)$.

For colored noise with its autocorrelation matrix R_n, we need to pre-whiten the noise by multiplying $R_n^{-\frac{1}{2}}$ on the observed noisy signal $\mathbf{x}(l)$ so that the input noisy signal becomes $R_n^{-\frac{1}{2}}\mathbf{x}(l)$ in which the noise signal is white. We then input the whitened noisy signal to the speech enhancement system. After processing, we recover the estimated speech by multiplying the estimated speech signal $\hat{\mathbf{s}}(l)$ by $R_n^{\frac{1}{2}}$.

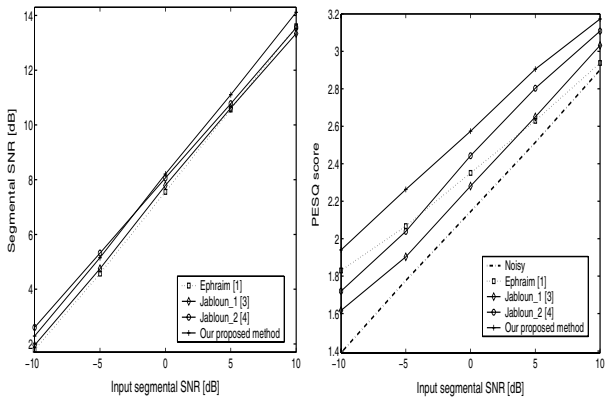
4. PERFORMANCE EVALUATION

Different noise types, taken from the NOISEX-92 database and 10 different utterances from the TIMIT database are used in our performance evaluations. In order to maintain the whiteness of the input noise, i.e., $R_n = \sigma_w^2 \mathrm{I}$, a rectangular analysis window is used. The proposed enhancement parameters are set to: sampling rate=8kHz, number of samples used to calculate autocorrelation=256, h=16, K=32 and P=256. The enhanced vectors are Hanning windowed and combined using the overlap-add-synthesis method. For colored noise, pre-whitening is used before subspace-enhancement processing. The estimation of noise may be obtained during speech pause intervals. In our simulations, since we only focus on speech enhancement algorithm study, and the noise signal is assumed to be stationary, the statistical characteristics of the noise signal is directly obtained from an initial segment of the noisy signal where only pure noise is present.

Fig. 1 shows the comparison results in terms of segmental SNR and PESQ for different subspace methods including conventional subspace method proposed by [1], a masking-based subspace by [3] and its improved version [4], as well as our proposed method in white noise case. Informal listening tests show that it is advantageous to exploit masking properties in subspace speech enhancement.

5. CONCLUSION

We have introduced an audible noise reduction scheme based on the signal subspace technique for enhancing noisy speech. The masking properties of human auditory system is incorporated into the signal subspace technique through the definition of the residual audible eigen-noise. Gain of eigen-filter is proposed to be regulated by an audible noise reduction criterion for the white noise model. Our proposed scheme is further extended to the colored noise model. Both objective and subjective tests indicate that our proposed scheme outperforms a number of conventional subspace methods.



(a) (b)

Fig. 1. Objective comparisons in the case of white noise (a) Segmental SNR; (b) PESQ score.

6. REFERENCES

[1] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech and Audio Processing*, vol. 3, July 1995.

[2] M. Klein and P. Kabal, "Signal Subspace Speech Enhancement with Perceptual Post-filtering," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-02, Vol. 1, pp. 537-540, May 2002.

[3] F. Jabloun and B. Champagne, "On the Use of Masking Properties of the Human Ear in the Signal Subspace Speech Enhancement Approach," in proc. of *Int. Workshop on Acoustic Echo and Noise Control*, IWAENC, Darmstadt, Germany, Sept. 2001.

[4] F. Jabloun and B. Champagne, "A Perceptual Signal Subspace Approach for Speech Enhancement in Colored Noise," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-02, Vol. 1, pp. 569-572, 2002.

[5] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. Speech and Audio Processing*, vol. 5, pp. 497-514, Nov. 1997.

[6] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-93, Vol. 2, pp. 355-358, Apr. 1993.

[7] J. D. Johnston, "Transform Coding of Audio Signal Using Perceptual Noise Criteria," *IEEE J. Select Areas Commun.*, Vol. 6, pp. 314-323, Feb. 1988.