SPEECH ENHANCEMENT WITH MASKING PROPERTIES IN EIGEN-DOMAIN FOR COLORED NOISE

Chang Huai You, Kong Aik Lee, and Cheung Chi Leung

Institute for Infocomm Research (I2R), A*STAR, Singapore 138632

ABSTRACT

In this paper, we study speech enhancement in eigen-domain. In our previous work on audible noise reduction, we use masking properties of the human auditory system to define the audible noise quantity in the eigen-domain. We then derived the speech enhancement algorithm using white noise model. Without loss of generality, in this paper, we introduce the noise reduction method for colored noise. Through many simulations, we show that colored noise modeling is superior over other existing eigen-decomposition methods in terms of objective and subjective evaluations.

Index Terms— Speech enhancement, eigen decomposition, human auditory system

1. INTRODUCTION

For speech enhancement, subspace approach is generally used for reducing wide-band additive noise. The underlying principle of 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 use for speech signal processing such as coding and enhancement are usually exploited through the critical frequency domain. In [2], a perceptual weighting filter is incorporated in a subspace approach to shape the residual noise. The perceptual weighting filter only reflects very limited properties of the human auditory system as compared to the masking threshold which is more encompassing. In [3], 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 signal. However, it cannot help in preserving the weak speech components which have already been erased by the conventional subspace method. Since the masking properties are related to the critical frequency band, which is derived from the characteristics of human cochlea, they could not be directly used in subspace methods as they do not function in the frequency domain.

In the conventional perceptual properties-based subspace speech-enhancement methods, the audible and inaudible components are neither explicitly separated nor individually analyzed, which is in contrast to the spectral suppression approach proposed in [4]. Based on the characteristics of the human auditory system, it was introduced in our previous work in [5] the audible noise quantity in the eigen-domain. It frees the reduction degree for inaudible noise so that some weak components of the underlying speech signal can be effectively preserved and the distortion of the speech signal can therefore be mitigated. In this paper, we extend the white noise model study to colored noise model for the reduction of audible noise.

The organization of the paper is as follows. In Section 2, we give a brief introduction of the invertible frequency eigen-domain transformation and the audible noise reduction scheme based on perceptual masking. In Section 3, we introduce the scheme for colored noise solution in eigen-domain. Section 4 shows the simulation results, and Section 5 concludes this paper.

2. EIGEN-DOMAIN SPEECH ENHANCEMENT AND MASKING PROPERTIES

The typical way of subspace speech enhancement is given by the following procedure: Firstly, the noisy signal is decomposed into a signal-plus-noise subspace and a noise subspace by applying the Karhunen-Loève transform (KLT); secondly, the KLT components are modified by a gain function; and thirdly, the enhanced signal is obtain by inverse KLT of the modified components [1].

A noisy speech sequence \( x(l) = [x(\omega l - P + 1), x(\omega l - P + 2), \cdots, x(\omega l - 1), x(\omega l)]^T \) is assumed to be the sum of a clean speech sequence \( s(l) = [s(\omega l - P + 1), s(\omega l - P + 2), \cdots, s(\omega l - 1), s(\omega l)]^T \) and a white noise sequence \( w(l) = [w(\omega l - P + 1), w(\omega l - P + 2), \cdots, w(\omega l - 1), w(\omega l)]^T \) with variance \( \sigma_w^2 \), i.e.,

\[
x(l) = s(l) + w(l)
\]  

(1)

where \( l, \omega \) and \( P \) denote the frame number, frame advance size and the vector dimension respectively. Let \( R_s \) denote the autocorrelation matrix of clean speech. The eigen-decomposition of \( R_s \) is given by

\[ R_s = U \Lambda_s U^T \]  

(2)

where \( \Lambda_s = \text{diag}[\lambda_{s0}, \cdots, \lambda_{sp-1}] \) and \( U = [u_0, \cdots, u_{p-1}] \) are the eigenvalue matrix and the orthonormal eigenvector matrix of \( R_s \) respectively.
Let $H$ be a linear filter derived according to the spectral domain constrained (SDC) estimator proposed by Ephraim and Van Trees in [1], we have

$$H = U \begin{bmatrix} G_1 & 0 \\ 0 & 0 \end{bmatrix} U^T$$

(3)

where $G_1$ is a full rank matrix. Substituting $\begin{bmatrix} G_1 & 0 \\ 0 & 0 \end{bmatrix}$ with $G$, we have $H = UGU^T$. A possible solution of the SDC estimator is that $G$ is a diagonal matrix given by

$$G = \Lambda_s (\Lambda_s + \sigma_w^2 \Lambda_m)^{-1},$$

(4)

where $\Lambda_s = \text{diag}[\mu_0, \cdots, \mu_{P-1}]$ is a diagonal matrix of Lagrange multipliers when deriving the estimator $H$ based on the criterion of minimizing the speech signal distortion with a permissible residual noise level [1].

We can use the auditory masking effects for noise suppression to lessen the classic tradeoff between noise reduction and speech distortion, wherein the audible noise is masked rather than being suppressed, and thereby reducing the chances of distorting the speech signal.

### 2.1. Frequency eigen-domain transformation

In order to introduce masking properties into the signal subspace field, one way is to map the masking threshold from the original frequency domain into the eigen-domain.

The relation between eigenvalue and the power spectral density (PSD) can be derived through the autocorrelation function. In practice, speech signals are analyzed on the short-term basis. A short-term PSD definition through Bartlett windowed autocorrelation is given as follows:

$$\psi(k) = \sum_{p=-P}^{P-1} (1 - |p|/P)r(p)e^{-j2\pi pk/K}$$

(5)

where $K$ denotes the number of frequency bins so that $k = 0, \cdots, K - 1$, and $r(p)$ is the autocorrelation function with lag $p$. Based on the above definition, the following relation is satisfied

$$\psi(k) = \frac{1}{P} \sum_{p=0}^{P-1} \lambda_p |V_p(k)|^2$$

(6)

where $V_p(k)$ is the Fourier transform of the unit-norm eigenvector, $u_p = [u_p(0), u_p(1), \cdots, u_p(P - 1)]^T$, of the autocorrelation matrix in the $k$th bin, which is defined as

$$V_p(k) = \sum_{i=0}^{P-1} u_p(i)e^{-j2\pi ki/K}, \quad p = 0, \cdots, P - 1.$$  

(7)

When $K \geq 2P + 1$, the inverse transformation of (5) is given by

$$r(p) = \frac{1}{K} \left( \frac{P}{P - p} \right) \sum_{k=0}^{K-1} \psi(k)e^{j2\pi kp/K}, \quad p = 0, \cdots, P - 1.$$  

(8)

In [6], we show that the inverse transformation of eigenvalue of (6) can be given by

$$\lambda_p = \frac{1}{K} \sum_{k=0}^{K-1} \psi(k)\Omega_p(k), \quad p = 0, \cdots, P - 1$$

(9)

where $\Omega_p(k)$ is given by

$$\Omega_p(k) = \sum_{m=0}^{P-1} u_p(m)u_p(m) +$$

$$\sum_{m=0}^{P-2} \sum_{b=m+1}^{P-1} u_p(m)u_p(b)(\frac{2P}{P - m - b})\cos[2\pi(m - b)k/K].$$

(10)

This means that the transformation pair of eigenvalue and PSD are given by (6) and (9) with $V_p(k)$ and $\Omega_p(k)$ being the functions of eigenvectors given by (7) and (10) respectively. In contrast to the conventional frequency eigen-domain transformation (FET) [7], the above transformation is called the invertible FET [6].

#### 2.2. Audible noise reduction

The masking threshold, $\Psi(k)$, in the frequency domain can be estimated by using critical band analysis, a spreading function, the relative threshold offset and the absolute auditory threshold. We define the masking-eigenvalues, $\lambda_t$, as a function of masking threshold according to the frequency eigen-domain transformation. Using (9), we have the masking-eigenvalue

$$\lambda_{t,p} = \frac{1}{K} \sum_{k=0}^{K-1} \Psi(k)\Omega_p(k), \quad p = 0, \cdots, P - 1.$$  

(11)

In (11), $\Omega_p(k)$ comes from the eigenvectors of the speech signal. As a result, the masking-eigenvalues, $\lambda_{t,p}$, can be interpreted as the transform of the masking threshold on the eigenvectors of the speech signal, i.e., they are the masking components distributed along the eigenvectors of the speech signal. As such, $\lambda_{t,p}$ can be used for the purpose of the arithmetic manipulation with the eigenvalues of speech signals in the same eigenvector space. Consequently, the masking-based subspace speech-enhancement algorithm can be realized.

According to the psychoacoustics theory, a speech signal can mask the coexisting noise. The noise degradation perceived by listeners will vary in time according to the time-varying properties of short-time speech components. For audible noise reduction, only the audible noise components must be removed by the enhancement algorithm. Consequently, the enhancement approach adopted here is based on the definition of an audible noise component, which is extended and used for the derivation of an optimal modifier that achieves audible noise suppression.

In our previous work [5], the audible noise reduction can be expressed by modifying the Lagrange multipliers as follows

$$\hat{\mu}_p = \begin{cases} \left(\frac{\lambda_{t,p} + \sigma_w^2}{\lambda_{t,p}} - 1\right)\frac{\lambda_{s,p} + \sigma_w^2}{\lambda_{s,p} + \sigma_w^2} + \delta, & \lambda_{s,p} + \sigma_w^2 > \lambda_{t,p} \\ \delta, & \lambda_{s,p} + \sigma_w^2 < \lambda_{t,p}. \end{cases}$$

(12)
3. COLORED NOISE SCHEME

Actually, (12) is set up based on white noise assumption. In order to extend to colored noise situation, we describe some approaches for colored noise solution based on the algorithm of the audible noise reduction [6].

3.1. Pre-whitening for colored noise

The first approach is to adopt pre-whitening for colored noise. For colored noise with its autocorrelation matrix \( R_n \), we may pre-whiten the noise by multiplying \( R_n^{-1/2} \) with the observed noisy signal \( x(l) \) so that the input noisy signal becomes \( R_n^{-1/2}x(l) \) in which the noise signal is white [8]. We then input the whitened noisy signal to the speech enhancement system. After processing, we obtain the final estimated speech by multiplying the estimated speech signal \( \hat{s}(l) \) by \( R_n^{1/2} \).

The autocorrelation function of the pre-whitened noisy signal is given by:

\[
R_{\text{prew}} = E\{ R_n^{-1/2}x(l)x^T(l) R_n^{-1/2}\} = R_n^{-1/2}R_n R_n^{-1/2} = R_n^{-1/2}(R_n + \mu R_n)(R_n^{-1/2})^T = R_n^{-1/2}R_n R_n^{1/2} + I = R_{\text{prew}} + I.
\]

where \( R_{\text{prew}} \) is the speech autocorrelation matrix and \( I \) the identity matrix. The estimated speech is given by the following equation:

\[
\hat{s}(l) = R_n^{1/2}UG_{\text{prew}}U^T R_n^{-1/2}x(l).
\]

where \( G_{\text{prew}} = \Lambda_{\text{prew}}(\Lambda_{\text{prew}} + \Lambda_\mu)^{-1} \), and \( \Lambda_{\text{prew}} \) is the eigenvalue matrix of \( R_{\text{prew}} \). Here, the pre-whitening method has changed the focus from the original speech signal into a pre-whitened clean speech signal. In other word, the original optimization constraint used for minimizing the variance of speech distortion is shifted to minimizing the variance of pre-whitened speech distortion. It therefore may not result in the optimal estimation of the clean speech signal.

In [9], a generalized subspace approach is proposed for enhancing speech corrupted by colored noise. The orthogonal KLT used for white noise is replaced by a nonsingular and nonorthogonal transformation, and the eigenvalues of the clean speech are essentially replaced by those of the whitened signal. It results in no explicit solution of the estimation of clean speech, but the whitened speech instead.

3.2. Gain modification for colored noise

The second approach is to modify the gain in the eigen-domain for colored noise. The total distortion of the estimation is given as follows

\[
d = \hat{s} - s = Hx - s = (H - I)s + Hw = d_s + d_n
\]

where \( d_s = (H - I)s \) is the signal distortion and \( d_n = Hw \) the residual noise. In order to provide proper spectral shaping of the residual noise and minimization of the variance of the signal distortion, \( H \) is obtained via

\[
\min_H \text{tr}\left\{ E\{d_n d_n^T\} \right\}
\]

subject to the constraints

\[
E\left\{ \left| u_n^T d_n \right|^2 \right\} \leq \alpha_p \sigma_n^2(p), \quad p = 0, \ldots, P - 1
\]

where \( \sigma_n^2(p) \) is the pth diagonal element of \( U^T R_n U \), and \( 0 \leq \alpha_p \leq 1 \) is the constraint constant. According to the time domain constraint estimator defined in [1],

\[
H_{\text{opt}} = \arg \left\{ \min_H \text{tr}\left\{ E\{d_n d_n^T\} \right\} \right\}
\]

\[
\text{subject to } : \frac{1}{P} \text{tr}\left\{ E\{d_n d_n^T\} \right\} \leq \sigma^2
\]

where \( \sigma^2 \) is a positive constant, the solution is given by

\[
H_{\text{opt}} = R_n(\mu R_n)^{-1} = U\Lambda_s(\mu U^T R_n U)^{-1}U^T.
\]

\( U^T R_n U \) can be approximated by a diagonal matrix with its pth diagonal element equal to the variance of noise along the pth eigenvector of the clean speech autocorrelation matrix, i.e.,

\[
U^T R_n U \simeq \begin{bmatrix}
\sigma_n^2(0) & 0 & \cdots & 0 \\
0 & \sigma_n^2(1) & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \sigma_n^2(P - 1)
\end{bmatrix}
\]

By modifying (4) with the connection of the above analysis, we arrive at

\[
\hat{G} = \Lambda_s(\Lambda_s + \Sigma_n \Lambda_n)^{-1}
\]

\[
\Sigma_n = \begin{bmatrix}
\sigma_n^2(0) & 0 & \cdots & 0 \\
0 & \sigma_n^2(1) & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \sigma_n^2(P - 1)
\end{bmatrix}
\]

and \( \tilde{\theta} \) is a diagonal matrix with its pth element \( \tilde{\theta}_p \) modified according to (12) as

\[
\tilde{\theta}_p = \left( \frac{\sqrt{\lambda_p + \sigma_n^2(p)}}{\lambda_p} - 1 \right) \frac{\lambda_p - \lambda_{p-1}}{\lambda_p - \lambda_{p+1}} + \tilde{\delta}_p,
\]

where \( \lambda_p \) is the pth diagonal element of \( \tilde{\theta} \) and \( \delta_p \) is set to 1.8 ~ 3.8. Consequently the estimate of the speech signal degraded by colored noise is obtained by

\[
\hat{s}(l) = U\tilde{G}U^T x(l).
\]

4. PERFORMANCE EVALUATION

Different noise types from the NOISEX-92 database and twenty different speech samples from the TIMT database are used in our performance evaluations. The parameters in the proposed algorithm are set as follows: sampling rate = 8.
kHz, number of samples used to calculate autocorrelation = 256, \( \varpi = 16, P = 32 \) and \( K = 256 \). The objective measurements used for performance evaluation are segmental signal-to-noise ratio (SNR) improvement, perceptual evaluation of speech quality (ITU-P.862) (PESQ) and modified Bark spectral distortion (MBSD).

The eigen-domain methods for colored noise modeling that are used for comparison in the evaluation include the pre-whitening subspace method by Lev-Ari and Ephraim (PWH-LE) [8], Jabloun and Champagne’s work (JC) [10], the proposed audible noise reduction method with (i) pre-whitening (14) with invertible FET (PWH-ANR), (ii) the colored noise model (24) with conventional FET (CFE-ANR) and (iii) the colored noise model (24) with invertible FET (IFE-ANR).

Fig. 1 compares the results by using F-16 cockpit noise. It can be observed that the performances of CFE-ANR and IFE-ANR are generally better than the others as expected. It can also be noted that IFE-ANR is always better than CFE-ANR. It shows, at least based on simulations with the range of input SNRs under consideration, that the invertible FET leads to better results than the conventional FET. Another observation is that for the case of colored noise shown in Fig. 1, there is an apparent drop in performance arising from the use of pre-whitening. It is largely because the original characteristics of the speech signal have been changed and the characteristics of the resulting signal can be very much unlike those of a speech signal. When using eigen-decomposition, pre-whitening may greatly weaken the advantage arising from the characteristics of principal component assumption of speech signal. Since the deviations of the eigenvectors of the pre-whitened signal from the eigenvectors of the original speech are usually large, when we discard the noise components, we might also discard more speech components due to the use of the criterion of separation of the signal and noise subspaces as well as eigen-filtering. Statistically, the pre-whitened speech is more like noise and therefore the difference between the pre-whitened speech and noise is less apparent than that between the original speech and noise. We therefore may not be able to separate the speech signal and noise signal effectively through the use of pre-whitening. This is why JC, CFE-ANR and IFE-ANR have better performances than PWH-LE and PWH-ANR.

The aim of speech enhancement is to improve the quality of a degraded speech signal while keeping the intelligibility at a certain level. Speech quality is a subjective measure that reflects the perception of listeners. The informal listening test results show that most listeners prefer the enhanced speech obtained by using IFE-ANR as compared to PWH-LE and JC.

5. CONCLUSION

This paper extended our previous work [5] on audible noise reduction, with a new solution to colored noise modeling. We have introduced two audible noise reduction approaches for the colored noise: pre-whitening (14) and gain modification (24). From the simulation results obtained, it is shown that the performance of the proposed gain modification (24) outperforms the two pre-whitening subspace methods that are cited in [8] and given by (14) respectively, as well as JC subspace [10] method for colored noise in terms of various measurements. It is also observed that the proposed eigen-domain method with the invertible FET is better than that based on the conventional FET.

6. REFERENCES