A PERCEPTUAL SIGNAL SUBSPACE APPROACH FOR SPEECH ENHANCEMENT IN COLORED NOISE

Firas Jabloun and Benoît Champagne

Department of Electrical & Computer Engineering, McGill University
3480 University Street, Montreal, Canada, H3A 2A7
firas@tsp.ece.mcgill.ca, champagne@ece.mcgill.ca
www.tsp.ece.mcgill.ca/~ firas

ABSTRACT

The major drawback of most noise reduction methods is what is known as musical noise. To cope with this problem, the masking properties of the human ear were used in the spectral subtraction methods. However, no similar approach is available for the signal subspace based methods. In a previous work, we presented a frequency to eigendomain transformation which provides a way to calculate a perceptually based upper bound for the residual noise. This bound, when used in the signal subspace approach, yields an improved result where better shaping of the residual noise is achieved. In this paper, we further improve this method and provide an easy way to generalize it to the colored noise case. Listening tests results are given to show the superiority of the proposed method.

1. INTRODUCTION

Most noise reduction methods for speech enhancement suffer from an annoying residual noise known as *musical noise*. To reduce the effect of this drawback, the use of a human hearing model which was first introduced in audio coding [1], has been proposed (e.g [2], [3]). This model is based on the fact that the human auditory system is able to tolerate additive noise as long as it is below some *masking threshold*. Methods to calculate this threshold are developed in the frequency domain according to critical band analysis and the excitation pattern of the basilar membrane in the inner ear [4]. These masking properties are not used in the signal subspace (SS) approach for noise reduction [5] because it does not operate in the frequency domain as is the case with the spectral subtraction methods.

In [6] we presented a frequency to eigendomain transformation (FET) which provides a way to calculate a perceptually based upper bound for the residual noise. This bound, when used in the signal subspace approach, yields better residual noise shaping from a perceptual perspective. In this paper; we further develop this method using a more sophisticated masking model and a modified gain function. We also provide a generalization of the algorithm to the colored noise case.

Subjective listening tests were carried out and the results show that the proposed new method outperformed other existing methods. The results also show that our method provided better noise shaping which is almost the same no matter what the original background noise is.

The paper is organized as follows. In section 2 we briefly describe the eigenfilter used in the enhancement method. The frequency to eigendomain transformation is described in section 3. The masking model is presented in section 4 and the details of the proposed method are given in section 5. Finally listening tests results are described in section 6 and a conclusion is given in section 7.

2. THE SIGNAL SUBSPACE APPROACH

Let $\mathbf{x} = \mathbf{s} + \mathbf{w}$ be a P-dimensional noisy observation vector where $\mathbf{s}$ is the desired vector and $\mathbf{w}$ is the noise vector with covariance matrix $\mathbf{R}_w$.

The eigenvalue decomposition of the covariance matrix $\mathbf{R}_s$ of the clean vector is given by $\mathbf{R}_s = \mathbf{U}\mathbf{\Lambda}_s\mathbf{U}^H$ where $\mathbf{\Lambda}_s = \text{diag}(\lambda_{s_1}, \ldots, \lambda_{s_P})$ with the eigenvalues λ_{s_i}'s in decreasing order. We first assume the noise to be white with $\mathbf{R}_w = \sigma^2\mathbf{I}$ so that $\mathbf{R}_x$, the covariance matrix of $\mathbf{x}$, will have the same eigenvectors as $\mathbf{R}_s$. We also assume that $\text{rank}(\mathbf{R}_s) = K < P$ so that $\lambda_{s_k} = 0$ for $k = K+1, \ldots, P$. Hence $\mathbf{U}$ can be written as $\mathbf{U} = [\mathbf{U}_1\mathbf{U}_2]$ where $\mathbf{U}_1$ spans the so-called signal subspace and $\mathbf{U}_2$ spans the noise subspace.

We want to find a linear estimate of $\mathbf{s}$ given by $\hat{\mathbf{s}} = \mathbf{H}\mathbf{x} = \mathbf{H}\mathbf{s} + \mathbf{H}\mathbf{w}$. The residual error signal is given by

$$\mathbf{r} = \hat{\mathbf{s}} - \mathbf{s} = (\mathbf{H} - \mathbf{I})\mathbf{s} + \mathbf{H}\mathbf{w} = \mathbf{r}_s + \mathbf{r}_w \qquad (1)$$

In the spectral domain constrained approach (SDC), the enhancement filter $\mathbf{H}$ is obtained by minimizing the signal distortion

$$\min_{\mathbf{H}} \quad \text{tr}(E\{\mathbf{r}_s\mathbf{r}_s^H\}) \qquad (2)$$

0-7803-7402-9/02/$17.00 ©2002 IEEE

subject to

$$E\{|\mathbf{u}_k^H \mathbf{r}_w|^2\} \leq \alpha_k \sigma^2 \quad k = 1 \ldots K \tag{3}$$

which ensures that the k^{th} spectral component of the residual noise is below some threshold. Here $\mathbf{u}_k$ is the k^{th} eigenvector of $\mathbf{R}_s$ with eigenvalue λ_{s_k}. The solution to this problem is given by [5]

$$\mathbf{H} = \mathbf{U}_1 \mathbf{Q} \mathbf{U}_1^H \tag{4}$$

where $\mathbf{Q}$ is a $K \times K$ diagonal gain matrix with entries

$$q_k = \alpha_k^{1/2} = e^{-\nu\sigma^2/\lambda_{s_k}} \quad k = 1, \ldots, K. \tag{5}$$

The major drawback of the above approach is that it requires the noise to be white. In [5], prewhitening is described as a remedy to this problem. Accordingly, the eigenfilter is modified as follows

$$\tilde{\mathbf{H}} = \mathbf{R}_w^{\frac{1}{2}} \mathbf{H} \mathbf{R}_w^{-\frac{1}{2}} \tag{6}$$

where $\mathbf{R}_w^{\frac{1}{2}}$ is the square root of the colored noise covariance matrix. We shall refer to this modified method as the signal subspace (SS) method with Prewhitening (SSwP).

3. THE FREQUENCY TO EIGENDOMAIN TRANSFORMATION

The filter described in the previous section provides some residual noise shaping but this shaping is not based on the masking properties of the human ear. To be able to include these properties in the eigenfilter design, a frequency to eigendomain transformation (FET) is required which relates the power spectrum density (PSD) of a random signal to the eigenvalues of its covariance matrix.

Let $\mathbf{R} = \text{toeplitz}(r(0), \ldots, r(P-1))$ be the covariance matrix of a zero mean random process $x(n)$ with autocorrelation function $r(p) = E\{x(n)x^*(n+p)\}$. Let λ_i and $\mathbf{u}_i = [u_i(0), \ldots, u_i(P-1)]^T$ be the i^{th} eigenvalue and unit norm eigenvector of $\mathbf{R}$ respectively. A well known relationship between λ_i and the PSD $\Phi(\omega) = \sum_{p=-\infty}^{\infty} r(p)e^{-j\omega p}$ of $x(n)$ is as follows [7]:

$$\lambda_i = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Phi(\omega) |V_i(\omega)|^2 d\omega \quad \text{for } i = 1 \ldots P \tag{7}$$

where $V_i(\omega) = \sum_{p=0}^{P-1} u_i(p)e^{-j\omega p}$ is the discrete-time Fourier transform of $u_i(p)$.

In practice only an estimate of the PSD is available. Of interest in the context of this paper is the Blackman-Tukey estimate which is the DTFT of a windowed version of the autocorrelation function $r(p)$, namely

$$\Phi_{BT}(\omega) = \sum_{p=-P+1}^{P-1} r(p) w_b(p) e^{-j\omega p} \tag{8}$$

If $w_b(p)$ is a Bartlett (triangular) window defined as $w_b(p) = 1 - \frac{|p|}{P}$ for $|p| < P$, then the Blackman-Tukey estimate can be written in terms of the eigenvalue decomposition of $\mathbf{R}$ as follows [7]

$$\Phi_{BT}(\omega) = \frac{1}{P} \sum_{i=1}^{P} \lambda_i |V_i(\omega)|^2 \tag{9}$$

Equation (9) can be viewed as a kind of "inverse" of equation (7). A detailed derivation of these two relationships can be found in [6]. The FET is to be used in the new proposed method for speech enhancement described in section 5.

4. CALCULATING THE MASKING THRESHOLD

Unlike in [6] where we used the masking model given in [1], a more sophisticated model is used in this paper. This model is similar to that used in ISO MPEG-1 audio coding standard [8] with some modifications. In this Section we briefly describe the steps required to calculate the masking threshold.

The resolution of the human auditory system is based on critical band analysis which follows a non-linear Bark scale. One Bark is related to the frequency in Hertz as follows [4]

$$z(f) = 13 \arctan(0.00076f) + 35 \arctan[(\frac{f}{7500})^2] \tag{10}$$

Inter-band masking is accounted for by convolution with a spreading function. This function has lower and upper skirts of +25 dB and -10 dB per critical band respectively and is given by [9]

$$SF(z) = 15.81 + 7.5(z + 0.474) - 17.5\sqrt{1 + (z + 0.474)^2} \tag{11}$$

The masking threshold is obtained by subtracting a relative threshold offset depending on the masker type, tone-like or noise-like. The tonality is measured as in [1] using the spectral flatness measure (SFM). Finally the so obtained threshold is compared with the absolute threshold of hearing and the maximum of the two is retained. The global masking threshold is calculated by linear adding these individual thresholds [8].Figure 1 illustrates the power spectrum density of a voiced speech frame together with the masking threshold calculated using the above procedure.

5. THE PROPOSED ALGORITHM

During non-speech activity periods, the noise autocorrelation function $\hat{r}_w(p)$ is estimated. This estimate is both used to calculate the power spectrum $\hat{\mathbf{\Phi}}_w(\omega)$ using the Blackman-Tukey estimator and to form the toeplitz covariance matrix $\hat{\mathbf{R}}_w$ of the noise.

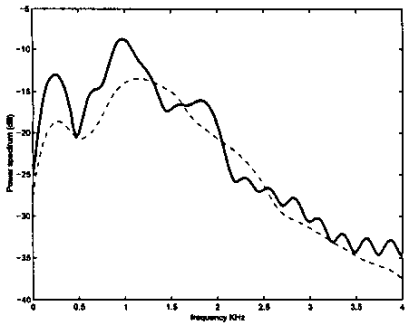

Figure 1: PSD of a voiced speech frame (continuous) and its corresponding masking threshold (dashed)

Let $\hat{\mathbf{R}}_x$ denote the covariance matrix estimate of $\mathbf{x}$. Since the noise and the speech signal are assumed to be uncorrelated, the clean speech signal covariance matrix is estimated as $\hat{\mathbf{R}}_s = \hat{\mathbf{R}}_x - \hat{\mathbf{R}}_w$. This estimate is not guaranteed to be positive definite so the rank K of $\mathbf{R}_s$ is chosen to be the number of strictly positive eigenvalues of $\hat{\mathbf{R}}_s$[10].

Define the vector $\hat{\boldsymbol{\lambda}}_s = [\hat{\lambda}_{s_1}\ \hat{\lambda}_{s_2}\ \ldots\ \hat{\lambda}_{s_K}]^T$ where $\hat{\lambda}_{s_k}$ is the k^{th} eigenvalue of $\hat{\mathbf{R}}_s$. Consider also the $J \times K$ matrix $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_K]$ where $\mathbf{v}_k$ is the magnitude squared of the J-point DFT of $\hat{\mathbf{u}}_k$, the k^{th} eigenvector of $\hat{\mathbf{R}}_s$.

Having defined all the necessary quantities, equation (9) is implemented to calculate the PSD as follows

$$\hat{\boldsymbol{\Phi}} = \frac{1}{P}\mathbf{V}\hat{\boldsymbol{\lambda}}_s \tag{12}$$

$\hat{\boldsymbol{\Phi}}$ is used to calculate a masking threshold $\boldsymbol{\Phi}_\theta$ as described in Section 4. A new set of eigenvalues is then recovered using equation (7) as follows

$$\boldsymbol{\lambda}_\theta = [\lambda_{\theta_1}, \ldots, \lambda_{\theta_K}]^T = \frac{1}{J}\mathbf{V}^H\boldsymbol{\Phi}_\theta \tag{13}$$

The masking properties of the human ear are now embedded in these eigenvalues.

Next, to handle the colored noise case, we calculate the vector $\hat{\boldsymbol{\lambda}}_w = [\hat{\lambda}_{w_1}, \ldots, \hat{\lambda}_{w_K}]^T$ in a similar way

$$\hat{\boldsymbol{\lambda}}_w = \frac{1}{J}\mathbf{V}^H\hat{\boldsymbol{\Phi}}_w \tag{14}$$

The elements of $\hat{\boldsymbol{\lambda}}_w$ will substitute σ^2 in equation (5) to allow the SS method to be generalized to the colored noise case. Actually in terms of $\hat{\mathbf{R}}_s$ and its eigenvectors, we have $\hat{\lambda}_{w_k} = \mathbf{u}_k^H\hat{\mathbf{R}}_s\mathbf{u}_k$, which is the same quantity used in [10] and was reported to have better noise shaping than the SSwP [5].

Now with these tools we can describe two methods for noise reduction which just differ in the gain function of equation (5). In the first method, which we call the modified SS method (MSS), the gain function is given by

$$q_k = e^{-\nu\hat{\lambda}_{w_k}/\hat{\lambda}_{s_k}} \tag{15}$$

This method gives similar results as in [10] and is introduced to evaluate the merit of using masking.

In the second method the gain is defined as follows

$$q_k = e^{-\nu\hat{\lambda}_{w_k}/\min(\hat{\lambda}_{s_k}, \lambda_{\theta_k})} \tag{16}$$

The minimum is used to obtain a smoother transition of the residual noise from silence periods to speech activity periods and consequently achieving better results. This method is called the perceptual SS method (PSS).

6. RESULTS

In the subset of experiments reported here, a 2.2 sec female spoken speech signal sampled at 8 KHz was used. The algorithm was implemented as described above, using the following parameters: $P = 32$, $J = 256$, $\nu = 3$ for SSwP and MSS, and $\nu = 1$ for PSS. A demo of these experiments is available in our web site mentioned in the title.

To evaluate the performance of the proposed algorithm, a subjective test was carried out. The original clean speech signal was corrupted with 4 different types of colored noise: a freezer motor, a military vehicle, a Volvo car and an F-16 cockpit noise. The SNR was different for every noise type in order to have conditions close to those of the real world. The different average and segmental SNR's are shown in Table 1.

A group of 14 people were asked to evaluate the performance of the new PSS method and to compare it with the original noisy signal, the SSwP and the MSS. The subjects ages were in the range of 22 to 35 years and none of them worked in the speech processing field. 12 pairs of recordings were presented to the subjects: for each pair, they were asked to vote for the signal they preferred. A neutral answer was also allowed if they could not perceive any difference.

Table 1 shows the results of this test. It can be seen that the PSS outperformed the other two enhancing methods especially with the military vehicle noise were all the subjects voted for the PSS. We note that in the F-16 cockpit noise case 40% of the subjects voted for the noisy signal because they preferred the existing noise to the obtained signal distortion. However, these subjects said that if the 2.2 sec test signal had been longer they would have changed their preference because the noise would be more disturbing and they would be less able to tolerate it.

To evaluate the noise shaping capabilities of the three speech enhancement methods, a second test was performed. The four signals, corresponding to the four noise types, were enhanced using the PSS method. The resulting enhanced signals were then presented to the subjects. No signal was taken as a reference. The subjects were asked to compare the characteristics of the four residual noises. The comparison is based on how similar or different these characteristics are in the four

Noise Type	SNR (dB) Ave/seg	Compared with noisy signal	Compared with SSwP	Compared with MSS
Freezer motor	5/-4	100%	90%	80%
Military vehicle	5/-4	100%	100%	100%
Volvo car	0/-10	80%	85%	60%
F-16 cockpit	10/-0.4	60%	70%	60%

Table 1: Subjective test results: colored noise case with four different noise types. Shown are the percentage of times where PSS was preferred, compared to noisy signal, SSwP and MSS.

signals. The subjects had to score their decision on a 5 level scale. The following choices were allowed:

0 : Completely different
1 : Different
2 : Don't know
3 : Almost similar
4 : Similar

The same test was then repeated with the other two methods, i.e. SSwP and MSS.

Figure 2 shows the results of this test. For the SSwP and MSS, the average score of 1 and 1.2, respectively, indicates that the subjects found the residual noise different in every case. The average score of 3.5 for PSS shows that the new method provided a residual noise shaping which is almost similar for all background noise types. Hence we conclude that the PSS has a relatively constant performance under different noise conditions.

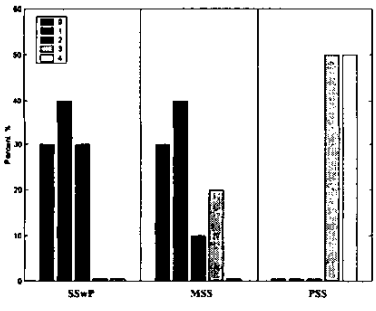

Figure 2: Noise shaping scores for the three different speech enhancement methods.

7. CONCLUSION

In this paper we presented a perceptual spectral domain constrained signal subspace approach for noise reduction. The proposed method uses the masking properties of the human ear within the eigenfilter design. This method is also capable to enhance signals corrupted with colored noise. Listening tests show that our method outperforms other existing signal subspace methods and, unlike these methods, the residual noise characteristics of the proposed PSS are almost similar no matter what the noise type is.

8. REFERENCES

[1] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J Select. Areas Commun*, vol. 6, pp. 314–323, Feb 1988.

[2] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 126–137, March 1999.

[3] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. Speech Audio Proc.*, vol. 5, pp. 479–514, Nov 1997.

[4] E. Zwicker and H. Fastl, *Psychoacoustics.* Berlin, Germany: Springer-Verlag, 1990.

[5] Y. Ephraim and H. L. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 251–266, July 1995.

[6] F. Jabloun and B. Champagne, "On the use of masking properties of the human ear in the signal subspace speech enhancement approach," *in Proc IWAENC, Darmstadt*, pp. 199–202, 2001.

[7] M. H. Hayes, *Statistical Digital Signal Processing and Modeling.* New York: John Wiley & Sons, Inc, 1996.

[8] K. Brandenburg and G. Stoll, "ISO-MPEG-1 Audio: A generic standard for coding of high quality digital audio," *Journal of the Audio Engineering Society*, vol. 42, pp. 780–792, Oct 1994.

[9] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *Journal of the Acoustic Society of America*, vol. 66, pp. 1647–1651, Dec 1979.

[10] U. Mittal and N. Phamdo, "Signal/Noise KLT based approach for enhancing speech degraded by colored noise," *IEEE trans. Speech Audio Processing*, vol. 8, pp. 159–167, march 2000.