# ON THE USE OF MASKING PROPERTIES OF THE HUMAN EAR IN THE SIGNAL SUBSPACE SPEECH ENHANCEMENT APPROACH

*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 this paper we present a relationship between the signal subspace domain and the frequency domain 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 the residual noise is much less annoying than the usual musical noise.

## 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, a human hearing model has been used (e.g [1][2][3]). This model was first introduced and is widely used in audio coding [4]. It 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 approach for noise reduction [5] because it does not operate in the frequency domain as is the case with the spectral subtraction methods. In this paper we present a relationship between the signal subspace domain and the frequency domain 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 the residual noise is much less annoying than the usual musical noise.

This paper is organized as follows. In section 2 we briefly describe the eigenfilter used in the enhancement method. A relationship between the frequency domain and the eigen domain which allows the use of the masking properties into the signal subspace approach is described in section 3. The masking threshold is explained in section 4 and the overall algorithm is summarized in section 5. Finally results are given in section 6 and a conclusion in section 7.

This work has been partially financed by *"La Mission Universitaire de la Tunisie en Amerique du Nord"*.

## 2. THE SIGNAL SUBSPACE APPROACH

In this section we introduce the signal subspace approach for speech enhancement presented by Ephraim and Van Trees in [5]. We just focus on the spectral domain constraint (SDC) estimator since the time domain estimator (TDC) can be viewed as a special case of the SDC.

Let $\mathbf{x} = \mathbf{s} + \mathbf{w}$ be the noisy observed vector where $\mathbf{s}$ is the desired vector and $\mathbf{w}$ is a white noise vector with variance $\sigma^2$. The eigendecomposition of the covariance matrix $\mathbf{R}_s$ of the clean vector is given by $\mathbf{R}_s = \mathbf{U}\mathbf{\Lambda}_s\mathbf{U}^H$. We want to find a linear estimator 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

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

The enhancement filter $\mathbf{H}$ is found by minimizing the signal distortion

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

subject to

$$E\{|\mathbf{u}_k^H\mathbf{r}_w|^2\} \leq \alpha_k\sigma^2 \qquad (3)$$

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

$$\mathbf{H} = \mathbf{U}\mathbf{Q}\mathbf{U}^H \qquad (4)$$

where $\mathbf{Q} = \mathrm{diag}(q_k) = \mathrm{diag}(\alpha_k^{1/2})$.

Several choices are available for the diagonal components of the matrix $Q$. To put these choices under a unique interpretation, we define the quantity

$$\gamma_k = \sigma^2/\lambda_{sk} \tag{5}$$

which is the inverse of the SNR on the $k$th spectral component. Then we let $q_k = f(\gamma_k)$ where $f(.)$ is a decreasing function satisfying $f(0) = 1$ and $f(\infty) \to 0$. A possible choice for this function is $f_1(\gamma) = \frac{1/\mu}{\gamma+1/\mu}$ leading to

$$q_k = \frac{\lambda_{s_k}}{\lambda_{s_k} + \mu\sigma^2}$$

which is the solution of the time domain constraint linear estimator. A second choice which has more noise suppression capabilities is $f_2(\gamma) = \exp(-\nu\gamma)$ which gives

$$q_k = e^{-\nu\sigma^2/\lambda_{s_k}}.$$

In practice $\mathbf{\Lambda}_s$ is not available so it is approximated as $\hat{\mathbf{\Lambda}}_s = \mathbf{\Lambda}_x - \beta\sigma^2\mathbf{I}$ where $\mathbf{\Lambda}_x$ is the eigenvalue matrix of the noisy vector $\mathbf{x}$ and $\beta$ is a scalar usually chosen to be one. This approximation tends to be one of the reasons behind the annoying residual musical noise. In the method we are proposing in this paper we try to replace $\mathbf{\Lambda}_s$ with another quantity which takes into account the masking properties of the human ear in order to shape the noise spectrum like the desired speech signal and eventually mask it.

## 3. RELATIONSHIP BETWEEN THE EIGENVALUES AND THE PSD

The properties of the human auditory system are especially understood in the frequency domain[1]. Therefore these properties have to be mapped to the eigen domain so that they can be used to design the eigenfilter presented in section 2. Namely we need a two-way relationship which relates the power spectrum density (PSD) of a random signal to the eigenvalues of its covariance matrix.

Let $\mathbf{R} = \text{toeplitz}\big(r(0), \ldots, r(P-1)\big)$ be the covariance matrix of a zero mean random process $x(n)$ with autocorrelation function $r(p) = E\{x(n)x^*(n+p)\}$. Let $\lambda_i$ 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, then $\lambda_i$ is related to the PSD $\Phi(\omega) = \sum_{p=-\infty}^{\infty} r(p)e^{-j\omega p}$ of $x(n)$ as follows:

$$\lambda_i = \frac{1}{2\pi}\int_{-\pi}^{\pi} \Phi(\omega)|V_i(\omega)|^2 d\omega \quad \text{for } i = 1 \ldots P \tag{6}$$

where

$$V_i(\omega) = \sum_{p=0}^{P-1} u_i(p)e^{-j\omega p} \tag{7}$$

[1]In this paper we are interested in the simultaneous masking which is a frequency domain phenomenon.

is the discrete-time Fourier transform of $u_i(p)$.

This relationship can be found in many statistical signal processing books and is basically used to prove the *Eigenvalue Extremal Property* [6].

In practice just an estimate of the PSD is available. Of interest in the context of this paper is the Blackman-Tuckey estimate which is the discrete-time Fourier transform of a windowed version of the autocorrelation function $r(p)$,

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

If $w_b(p)$ is a Bartlett (triangular) window defined as $w_b(p) = 1 - \frac{|p|}{P}$ for $|p| < P$ then the Blackman-Tuckey estimate can be written in terms of the eigendecomposition of $\mathbf{R}$ as follows [6]

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

Equation (9) can be considered as the "inverse" of equation (6) although mathematically speaking this is not correct. A detailed derivation of these two relationships is given in the appendix.

The power spectrum estimate $\Phi_{BT}(\omega)$ is a smeared version of the PSD of $x(n)$ obtained by convolving $\Phi(\omega)$ with the Fourier transform of $w_b(p)$. However, in our current application this is not a problem since we will eventually be applying some transformations to the spectrum which will cause more severe smearing [4]. These relationships are to be used in the new proposed method for speech enhancement described in section 5.

## 4. CALCULATING THE MASKING THRESHOLD

In this section we briefly describe the steps required to calculate the masking threshold.

The human ear can not distinguish between two frequencies belonging to the same critical band and its resolution is usually linear up to 1KHz and logarithmic thereafter. So the first step in calculating the masking threshold is critical band analysis which converts the linear frequency scale to the logarithmic Bark scale [4]. After that, masking between different critical bands is taken into account by convolution with a spreading function. This imitates the excitation pattern of the basilar membrane in the inner ear where the cells of the basilar membrane corresponding to a critical band are also excited by other frequencies in neighboring bands. The spreading function we used has lower & upper skirts with slopes of +25 dB and -10 dB per critical band respectively [7].

The final step is the subtraction of a relative threshold offset depending on the masker type (tonal or non-tonal). As in [1] we use the method suggested in [8]

which estimates the tonality instead of calculating it exactly. It is based on the fact that the speech signal has a tone like nature in lower critical bands and a noise like nature in higher critical bands. The step which accounts for absolute threshold of hearing is not included since it can be done using the control parameters of the eigenfilter. The detailed steps to calculate the masking threshold can be found in [4] and [1].

## 5. THE PROPOSED ALGORITHM

In this section we show how to put everything together and describe the steps required to implement the algorithm. We try to use matrix notation whenever possible to make the straight forward implementation of the algorithm as easy as possible.

### 5.1. Implementation

Although the signal subspace approach outperforms the spectral subtraction methods [5] especially at very low SNR conditions, its major drawback remains the large computational load required to calculate the covariance matrix and especially its eigendecomposition. To solve this problem Gazor and Rezayee [9], for example, propose an adaptive approach based on the PASTd subspace tracking algorithm. However this method is based on estimating the covariance matrix using a sliding exponential window which introduces some undesired reverberation to the enhanced signal. Therefore we prefer a slightly modified version of the method proposed in [5] to calculate the signal subspace.

We divide the speech signal into overlapped frames of length $N$. The $N$ samples are used to calculate the first $P$ coefficients of the biased autocorrelation function, efficiently implemented using the FFT. From these coefficients, a toeplitz covariance matrix is formed. An eigenfilter is designed using the eigendecomposition of this covariance matrix as explained in the next subsection. Every frame is divided into smaller P-dimensional overlapping vectors, and every vector is enhanced using the same eigenfilter of the current frame. The vectors are then multiplied by a Hanning window and synthesized using the overlap-add method. Finally every frame is multiplied by a second Hanning window and the total speech signal is recovered using the overlap-add synthesis technique.

With this method, we need to calculate a new eigenfilter less frequently hence reducing the computational load. Besides, this frame by frame processing makes it easy to merge the signal subspace approach with other speech enhancement techniques like spectral subtraction to further reduce the computational cost without degrading the enhanced speech quality.

### 5.2. Calculating the eigenfilter

Given the P-dimensional noisy speech vector $\mathbf{x}$ and the eigendecomposition of its covariance matrix $\mathbf{R}$. Consider the vector $\boldsymbol{\lambda}_s = [\lambda_{s_1}\ \lambda_{s_2}\ \ldots\ \lambda_{s_P}]^T$ where $\lambda_{s_i} = \lambda_i - \beta\sigma^2$. $\lambda_i$ is the $i^{th}$ eigenvalue of $\mathbf{R}$, $\sigma^2$ is the noise variance and $\beta$ is a scalar. Define the matrix $\mathbf{V}$ which has the amplitude squared of the discrete Fourier transform of the eigenvectors of $\mathbf{R}$ on its columns. Having defined all the necessary quantities, equation (9) is implemented to calculate the PSD as follows

$$\boldsymbol{\Phi} = \frac{1}{P}\mathbf{V}\boldsymbol{\lambda}_s \tag{10}$$

With this PSD a masking threshold $\boldsymbol{\Phi}_\theta$ can be calculated as described in section 4. After that a new set of eigenvalues is recovered using equation (6) as follows

$$\boldsymbol{\lambda_\theta} = \frac{1}{K}\mathbf{V}^H\boldsymbol{\Phi}_\theta \tag{11}$$

where $K$ is the DFT size. The masking properties of the human ear are now embedded in this new set of eigenvalues. So the eigenfilter of equation (4) designed using these eigenvalues, will shape the residual noise spectrum like that of the desired speech signal and eventually reduces the musical noise.

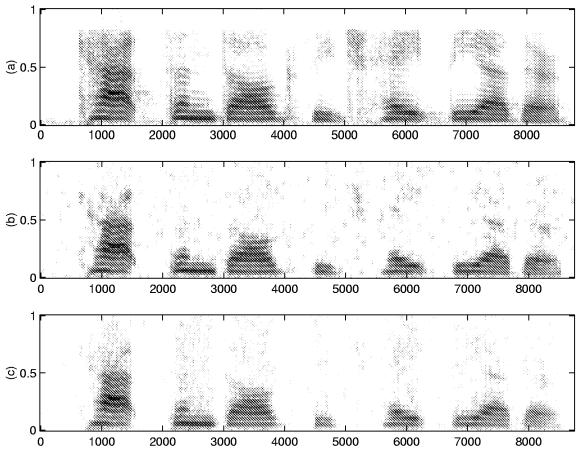


Figure 1: Spectrograms of (a)clean speech (b)Ephraim's method (c) proposed method.

## 6. RESULTS

In our simulations a speech signal sampled at 8Kz was used. The algorithm was implemented using the following parameters. $P = 32$, $N = 256$, $\beta = 1$, $f_2(\gamma)$ was used for the gain matrix $Q$ with $\nu = 3$. The size of the DFT was $K = 256$. Figure 1 shows the spectrograms of a clean speech signal, a signal enhanced using Ephraim's signal subspace method and finally using our proposed method. It can be seen that our method reduces the musical noise.

Besides, ten people were asked to take a listening test to evaluate the proposed method and compare it with the non processed noisy signal and with Ephraim's signal subspace approach. Every recording consisted of two sentences spoken consecutively by two male and two female speaker. Computer generated white noise was added to the recordings at an average SNR of 20 dB, 10 dB and 5 dB. The recordings were presented to the listeners in pairs each representing two different processing methods. The listeners were asked to compare the two recordings and choose the one they prefer. Table 1 shows the results of this test. On the average the listeners preferred the enhanced signal over the noisy signal 88% of the times and preferred the use of masking threshold in 73% of the times. The proposed method becomes more useful at very low SNR conditions where the subjects voted for the use of masking threshold to enhance the speech signals in 95% of the times. The recordings used in this test can be found in the web site mentioned in the title as a demo.

| Input SNR | Compared with noisy signal | Compared with Ephraim's method |
|---|---|---|
| 20 dB | 95% | 60% |
| 10 dB | 85% | 65% |
| 5 dB | 85% | 95% |

Table 1: Listening test results

## 7. CONCLUSION

In this paper we presented a signal subspace noise reduction method which uses the masking properties of the human ear. Listening tests show that our method largely reduces the effect of musical noise and hence improves the quality of the enhanced speech.

## 8. APPENDIX

In this appendix we prove the relations (6) and (9)given above.

*Proof.* By definition we have

$$\lambda_i = \mathbf{u}_i^H R u_i = \sum_{p=0}^{P-1}\sum_{q=0}^{P-1} u_i^*(p) r(p-q) u_i(q)$$

So using the relation

$$r(p) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \Phi(\omega) e^{j\omega p} d\omega$$

we get

$$\begin{aligned}\lambda_i &= \sum_{p=0}^{P-1}\sum_{q=0}^{P-1} u_i^*(p) u_i(q) \frac{1}{2\pi}\int_{-\pi}^{\pi} \Phi(\omega) e^{j\omega(p-q)} d\omega \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \Phi(\omega) V_i(\omega) V_i^*(\omega) d\omega\end{aligned}$$

And this completes the proof for (6). In a similar way we have

$$\begin{aligned}&\frac{1}{P}\sum_{i=1}^{P} \lambda_i |V_i(\omega)|^2 \\ &= \frac{1}{P}\sum_{p=0}^{P-1}\sum_{q=0}^{P-1} e^{-j\omega(p-q)} \sum_{i=1}^{P} \lambda_i u_i(p) u_i^*(q) \\ &= \frac{1}{P}\sum_{p=0}^{P-1}\sum_{q=0}^{P-1} r(p-q) e^{-j\omega(p-q)} \\ &= \frac{1}{P}\sum_{p=-P+1}^{P-1} r(p)(P-|p|) e^{-j\omega p}\end{aligned}$$

Which proves (9) □

## 9. REFERENCES


[1] 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.

[2] S. Gustafsson, P. Jax, and P. Vary, "A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics," *in Proc ICASSP 98*, pp. 397–400, 1998.

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

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

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

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

[8] D. Sinha and A. H. Tewfik, "Low bit rate transparent audio compression using adapted wavelets," *IEEE Trans. Signal Processing*, vol. 41, pp. 3463–3479, Dec 1993.

[9] S. Gazor and A. Rezayee, "An adaptive subspace approach for speech enhancement," *in Proc ICASSP 2000*, pp. 1839–1842, 2000.