Incorporating the Human Hearing Properties in the Signal Subspace Approach for Speech Enhancement

Firas Jabloun and Benoît Champagne

***Abstract*—The major drawback of most noise reduction methods in speech applications is the annoying residual noise known as musical noise. A potential solution to this artifact is the incorporation of a human hearing model in the suppression filter design. However, since the available models are usually developed in the frequency domain, it is not clear how they can be applied in the signal subspace approach for speech enhancement. In this paper, we present a Frequency to Eigendomain Transformation (FET) which permits to calculate a perceptually based eigenfilter. This filter yields an improved result where better shaping of the residual noise, from a perceptual perspective, is achieved. The proposed method can also be used with the general case of colored noise. Spectrogram illustrations and listening test results are given to show the superiority of the proposed method over the conventional signal subspace approach.**

***Index Terms*—Colored noise, Karhunen-Loeve transform (KLT), masking threshold, signal subspace, speech enhancement.**

I. Introduction

THE performance of speech communication systems in applications such as hands-free telephony, degrade considerably in adverse acoustic environments. The presence of noise can cause loss of intelligibility as well as the listener's discomfort and fatigue. Speech enhancement methods seek to improve the performance of these systems and to make the corrupted speech more pleasant to the listener. These methods are also useful in other applications such as automatic speech recognition.

In this paper we focus on the signal subspace approach (SSA) for speech enhancement [1]. This technique is based on the decomposition of the noisy signal vector space into two orthogonal subspaces called the noise subspace and the signal subspace. In this context, the signal subspace decomposition can be achieved either using the Karhunen-Loeve transform (KLT) via eigenvalue decomposition (EVD) of the data covariance matrix [1]–[4], or using the singular value decomposition (SVD) of a data matrix [5]–[7]. The discrete cosine transform (DCT) has also been proposed as an approximation to the KLT [8], [9].

In the SSA, enhancement is obtained by removing the noise subspace as a first step. Then the clean speech is recovered in the remaining signal subspace by optimally weighting the signal coefficients in this subspace. The different SSA methods vary according to the weighting scheme used [6]. The SSA can also be interpreted as a filterbank with the weighting coefficients serving as the subband filters [10].

As in most single channel speech enhancement methods such as spectral subtraction [11], the signal subspace methods suffer from the annoying residual noise known as *musical noise*. Tones at random frequencies, resulting from poor estimation of the signal and noise statistics, are at the origin of this artifact.

In spectral subtraction and its variants, modifications using a human hearing model were proposed to reduce the prominence of the musical noise [12]–[16]. This technique, which was first introduced in audio coding [17], 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 the masking threshold are developed in the frequency domain according to critical band analysis and the excitation pattern of the basilar membrane in the inner ear [18].

Recently, a DCT based SSA imitating the human hearing resolution was proposed [9]. However, no algorithm which employs a sophisticated hearing model with a KLT based SSA is available. The reason is that the SSA do not operate in the frequency domain where the available hearing models are developed. In this paper, we present a frequency to eigendomain transformation (FET) which provides a way to calculate a perceptually based eigenfilter. This is done by estimating an eigenvalue decomposition based power spectral density (PSD) from which a masking threshold is calculated. This threshold is transformed to the speech signal eigendomain using the FET allowing to design the perceptual eigenfilter. This filter yields better residual noise shaping from a psychoacoustic perspective. We provide an analysis of the FET and show how it can be incorporated in the SSA to improve its performance. We also show how the method can be modified to cover the more general case of colored noise.

Informal as well as formal subjective listening test results show that the proposed new method outperforms the conventional SSA. The results also show that our method provides better noise shaping in the sense that for a given speech signal, the residual noise has relatively similar characteristics in different noisy environments.

The paper is organized as follows. In Section II we briefly introduce the signal subspace approach for speech enhancement. The masking model used is described in Section III. The details of the FET are explained in Section IV. Section V deals with the colored noise case and the overall proposed method is given in Section VI. Experimental results are presented in Section VII and finally a conclusion is given in Section VIII.

Manuscript received January 17, 2002; revised June 13, 2003. This work was supported in part by a grant from the Natural Science & Engineering Research Council of Canada. F. Jabloun was supported by "La mission universitaire de Tunisie en Amerique du Nord." The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Li Deng.

The authors are with Department of Electrical & Computer Engineering, McGill University, Montreal, QC, H3A-2A7, Canada.

Digital Object Identifier 10.1109/TSA.2003.818031

II. Signal Subspace Approach

In this section we briefly introduce the signal subspace approach. The reader is referred to [1] for further details.

Let $\mathbf{x} = \mathbf{s} + \mathbf{w}$ be a P-dimensional noisy observation vector where $\mathbf{s}$ is the desired speech vector and $\mathbf{w}$ is the noise vector with covariance matrix $\mathbf{R}_w$. The noise is assumed to be uncorrelated with the speech signal so that the noisy signal covariance matrix $\mathbf{R}_x$ can be written as

$$\mathbf{R}_x = \mathbf{R}_s + \mathbf{R}_w \tag{1}$$

where $\mathbf{R}_s$ is the clean speech covariance matrix. The eigenvalue decomposition (EVD) of $\mathbf{R}_s$ is given by

$$\mathbf{R}_s = \mathbf{U}\mathbf{\Lambda}_s\mathbf{U}_H \tag{2}$$

where $\mathbf{\Lambda}_s = \mathrm{diag}(\lambda_{s,1}, \ldots, \lambda_{s,P})$ with the eigenvalues $\lambda_{s,i}$'s in decreasing order, and $\mathbf{U}$ is the unitary eigenvector matrix expressed as $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_P]$. In this section we assume the noise to be white with variance σ^2, that is $\mathbf{R}_w = \sigma^2\mathbf{I}$. Hence, the EVD of $\mathbf{R}_x$ can be written as

$$\mathbf{R}_x = \mathbf{U}(\Lambda_s + \sigma^2\mathbf{I})\mathbf{U}^H. \tag{3}$$

Note that in this case $\mathbf{R}_x$ and $\mathbf{R}_s$ have the same eigenvectors.

A key assumption in the signal subspace approach is that $\mathbf{R}_s$ is rank deficient with $\mathrm{rank}(\mathbf{R}_s) = Q < P$. Therefore, we have $\lambda_{s,i} = 0$ for $i = Q+1, \ldots, P$. Accordingly, $\mathbf{U}$ can be written as $\mathbf{U} = [\mathbf{U}_1\mathbf{U}_2]$ where $\mathbf{U}_1 = [\mathbf{u}_1, \ldots, \mathbf{u}_Q]$ spans the so-called signal subspace and $\mathbf{U}_2 = [\mathbf{u}_Q + 1, \ldots, \mathbf{u}_P]$ spans the noise subspace.

With these assumptions, a linear filter, $\mathbf{H}$, toestimate the desired speech vector $\mathbf{s}$ from the noisy observation $\mathbf{x}$ is designed as follows: Let $\hat{\mathbf{s}}$ denote the estimate of $\mathbf{s}$ at the filter output

$$\hat{\mathbf{s}} = \mathbf{H}\mathbf{x} = \mathbf{H}\mathbf{s} + \mathbf{H}\mathbf{w}. \tag{4}$$

The residual error signal is defined as

$$\mathbf{r} = \hat{\mathbf{s}} - \mathbf{s} = (\mathbf{H} - \mathbf{I})\mathbf{s} + \mathbf{H}\mathbf{w} \tag{5}$$

with $\mathbf{r}_s \stackrel{\triangle}{=} (\mathbf{H} - \mathbf{I})\mathbf{s}$ being the signal distortion and $\mathbf{r}_w \stackrel{\triangle}{=} \mathbf{H}\mathbf{w}$ being the residual noise.

In the particular form of the SSA called the spectral domain constrained approach (SDC), the enhancement filter $\mathbf{H}$ is the solution to the following optimization problem:

$$\begin{aligned} &\min_{\mathbf{H}} E\{\|\mathbf{r}_s\|^2\} \\ &\text{subject to} \\ &E\left\{|\mathbf{u}_i^H\mathbf{r}_w|^2\right\} \le \alpha_i\sigma^2 \quad \text{for} \quad 1 \le i \le Q \\ &E\left\{|\mathbf{u}_i^H\mathbf{r}_w|^2\right\} = 0 \quad \text{for} \quad Q < i \le P. \end{aligned} \tag{6}$$

The goal here is to minimize the signal distortion subject to keeping every spectral component of the residual noise, in the signal subspace, below some predefined threshold. The solution to this problem is given by [1]

$$\mathbf{H} = \mathbf{U}_1\mathbf{G}\mathbf{U}_1^H \tag{7}$$

where the entries of the gain matrix $\mathbf{G} = \mathrm{diag}(g_1, \ldots, g_Q)$ are chosen to be

$$g_i = e^{-\nu\sigma^2/\lambda_{s,i}} \quad i = 1, \ldots, Q \tag{8}$$

where ν is a parameter that controls the tradeoff between the residual noise level and the signal distortion. Note that other alternatives for the gain function are also possible [1].

The matrix $\mathbf{U}_1^H$ is referred to as the Karhunen-Loeve Transform[1] (KLT) and its effect on the noisy signal vector $\mathbf{x}$ is to calculate the coefficients of its projection onto the signal subspace. These coefficients have the property of being uncorrelated so that they can be processed independently using a diagonal gain matrix. The enhanced signal vector is finally reconstructed in the signal subspace using the matrix $\mathbf{U}_1$, the inverse KLT.

III. Calculating the Masking Threshold

A potentially important development in noise reduction methods is the incorporation of the psychoacoustic properties of human hearing, namely the so called masking [12]. During the past decades, research has been conducted to understand the human auditory system and to develop models which mimic its behavior. Among these we mention that of Johnston [17] and the more sophisticated model 1 and model 2 of the ISO MPEG-1 audio coding standard [19]. Several other models are available for example in [20] and [21].

In this paper we use the MPEG-1 model 1 which was found to be reliable in practice. We provide here a brief description of this model and the interested reader can refer to [19] for further implementation details.

The masking phenomenon can be explained by the so called critical bands. Within one critical band, one sound (the maskee) becomes inaudible in the presence of another sound (the masker) with a higher intensity (here the speech signal is the masker while the undesired noise is the maskee). A perceptual measure, called the Bark scale, relates the acoustic frequency to this nonlinear perceptual frequency resolution, in which one Bark covers one critical bandwidth. The analytical expression which can be used to map the frequency f (in hertz) to the critical-band rate z (in Barks) is [18]

$$z(f) = 13\arctan(0.000\,76f) + 35\arctan\left[\left(\frac{f}{7500}\right)^2\right]. \tag{9}$$

Tonal and nontonal (noise-like) components of the magnitude spectrum of the input signal are identified according to the local spectral maxima. Using the mapping in (9), the masking threshold of each of these individual components is then calculated and the resulting individual thresholds are summed *linearly* to obtain the global masking threshold. A masking component at a particular frequency is discarded if it is below the absolute threshold of hearing at that frequency [18].

The masking threshold of a tonal component is given by

$$T_{\mathrm{tm}}(j,i) = X_{\mathrm{tm}}(j) + O_{\mathrm{tm}}(j) + \mathrm{SF}(j,i) \tag{10}$$

[1]In fact $\mathbf{U}_1^H$ is not exactly the KLT since it does not contain all the components of the "real" KLT, $\mathbf{U}^H$. However, since these missing components have a zero weight in the gain matrix $\mathbf{G}$, $\mathbf{U}_1^H$ can still be considered to be the KLT.

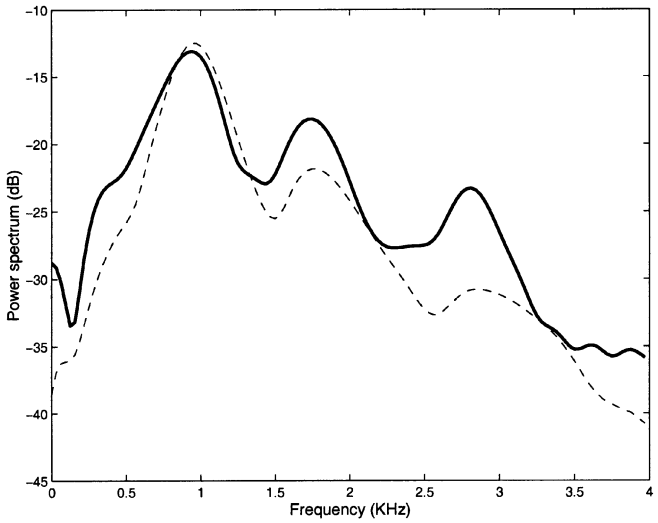

Fig. 1. Power spectrum of a voiced speech frame (the vowel /a/) (continuous) and its corresponding masking threshold (dashed).

where $T_{\rm tm}(j,i)$ is the masking threshold at i barks due to the masking component located at j barks. $X_{\rm tm}(j)$ is the sound pressure level (in dB) of the masking component with critical band index j. The function $O_{\rm tm}(j)$ is the threshold offset given by

$$O_{\rm tm}(j) = -1.525 - 0.275j - 4.5\ \text{dB}. \tag{11}$$

Similarly, the masking threshold of each nontonal component is given by

$$T_{\rm nm}(j,i) = X_{\rm nm}(j) + O_{\rm nm}(j) + \text{SF}(j,i) \tag{12}$$

where

$$O_{\rm nm}(j) = -1.525 - 0.175j - 0.5\ \text{dB}. \tag{13}$$

$\text{SF}(j,i)$ is the spreading function which accounts for the inter-band masking and is given by (see (14) at the bottom of the page) where $dz = i - j$ in barks. The spreading function has no effect on regions of the spectrum that are outside the range of -3 to 8 barks on the critical band rate scale, relative to the location of the masking component. $X(j)$ in (14) stands for either $X_{\rm tm}(j)$ or $X_{\rm nm}(j)$.

Fig. 1 shows the power spectral magnitude of a voiced speech frame (the vowel /a/) and its corresponding masking threshold.

IV. Frequency to Eigendomain Transformation

The filter described in Section II provides some residual noise shaping but this shaping is not based on the masking properties of the human ear. If we can design the gain function (8) based on the masking threshold described in the previous section, then we can achieve better noise shaping from a perceptual viewpoint. Therefore more residual noise can be allowed in the enhanced signal without being perceived by the listener which reduces the signal distortion and hence improves intelligibility. However, as discussed in Section III, the masking threshold is better understood and is calculated in the frequency domain. So to be able to include the hearing properties in the eigenfilter design, a frequency to eigendomain transformation (FET) is required which relates the power spectral density (PSD) of a speech signal to the eigenvalues of its covariance matrix.

A. Derivation

Consider a zero mean stationary signal $x(n)$ with autocorrelation function $\tilde{r}(p) = E\{x(n)x^*(n+p)\}$, where $E\{\cdot\}$ is the expectation operator. The PSD of $x(n)$ is defined as follows:

$$\tilde{\Phi}(\omega) = \sum_{p=-\infty}^{\infty} \tilde{r}(p)e^{-j\omega p}. \tag{15}$$

In practice, however, we need to estimate the PSD from a single realization of $x(n)$ over a finite time interval of length N. To this end, consider the biased autocorrelation estimator given by

$$r(p) = \frac{1}{N}\sum_{n=0}^{N-1-p} x(n)x^*(n+p) \quad p = 0,\ldots,N-1 \tag{16}$$

with $r(-p) = r^*(p)$ and $r(p) = 0$ for $|p| \geq N$. The PSD can then be estimated using the periodogram defined as [22]

$$\Phi(\omega) = \sum_{p=-N+1}^{N-1} r(p)e^{-j\omega p}. \tag{17}$$

Now let $\mathbf{R} = \text{Toeplitz}(r(0),\ldots,r(P-1))$ be the covariance matrix estimate of $x(n)$ with λ_i being its ith eigenvalue and $\mathbf{u}_i = [u_i(0),\ldots,u_i(P-1)]^T$ being the corresponding unit norm eigenvector. $\mathbf{R}$ is assumed in general to have rank $Q \leq P$, so that $\lambda_i = 0$ for $i > Q$.

It is not difficult to show that λ_i can be written in terms of $\Phi(\omega)$ in the following way [22]:

$$\lambda_i = \frac{1}{2\pi}\int_{-\pi}^{\pi} \Phi(\omega)|V^i(\omega)|^2\, d\omega \quad \text{for} \quad i = 1\ldots Q \tag{18}$$

where

$$V_i(\omega) = \sum_{p=0}^{P-1} u_i(p)e^{-j\omega p} \tag{19}$$

$$\text{SF}(j,i) = \begin{cases} 17(dz+1) - 0.4X(j) - 6\ \text{dB} & -3 \leq dz < -1 \\ (0.4X(j)+6)dz\ \text{dB} & -1 \leq dz < 0 \\ -17dz\ \text{dB} & 0 \leq dz < 1 \\ -(dz-1)(17 - 0.15X(j)) - 17\ \text{dB} & 1 \leq dz < 8 \end{cases} \tag{14}$$

is the Discrete-Time Fourier Transform of the entries $u_i(p)$ of the eigenvector $\mathbf{u}_i$. Equation (18) will be called the Frequency to Eigendomain Transformation (FET).

Multiplying $r(p)$ in (17) by a length $P \leq N$ window $w_b(p)$, we obtain the Blackman-Tukey estimator

$$\Phi_B(\omega) = \sum_{p=-P+1}^{P-1} r(p) w_b(p) e^{-j\omega p}. \tag{20}$$

If $w_b(p)$ is a Bartlett (triangular) window, then $\Phi_B(\omega)$ can be written in terms of the eigenvalue decomposition of $\mathbf{R}$ as follows [23]:

$$\Phi_B(\omega) = \frac{1}{P} \sum_{i=1}^{Q} \lambda_i |V_i(\omega)|^2. \tag{21}$$

Equation (21) can be viewed as a sort of "inverse" for (18). Accordingly, we refer to it as the Inverse Frequency to Eigendomain Transformation (IFET). For completeness, a proof of these two relationships is included in the Appendix. The FET is to be used in the new proposed method for speech enhancement described in Section VI.

B. Properties of the Blackman-Tukey Spectrum Estimator

Since the Inverse FET provides a PSD estimate based on the Blackman-Tukey spectrum estimator, we found it necessary to examine the properties of this estimator to verify how adequate it is for the current application.

The periodogram is a very popular spectrum estimator because it can be directly calculated from the samples of $x(n)$. However, it suffers from a high variance [23]

$$\mathrm{Var}\{\Phi(\omega)\} \approx \tilde{\Phi}^2(\omega). \tag{22}$$

This variance is in general considered to be high and can not be tolerated. Precisely, in the current application, the same eigenfilter designed using the FET, will be applied to several overlapping adjacent vectors as will be discussed in Section VI. Therefore, it is preferable that the designed filter have a minimal variance.

In the Blackman-Tukey estimator, the variance is reduced by multiplying the autocorrelation function by the window. The variance in this case is approximately [23]

$$\mathrm{Var}\{\Phi_B(\omega)\} \approx \tilde{\Phi}^2(\omega) \frac{1}{N} \sum_{i=-P}^{P} \omega_b^2(i) \approx \tilde{\Phi}^2(\omega) \frac{2P}{3N} \tag{23}$$

which is less than $\mathrm{Var}\{\Phi(\omega)\}$ since $P \leq N$.

This lower variance is obtained at the expense of a reduced resolution. The Blackman-Tukey estimate is a smoothed version of the periodogram due to the convolution with the Fourier Transform of the window in the frequency domain. So the resolution depends on the bandwidth of the main lobe of the window which in turn depends on its size and type. In our case, for a length $2P-1$ Bartlett window the resolution $\Delta\omega$ is given by [22]

$$\Delta\omega \approx 0.64 \frac{2\pi}{P}. \tag{24}$$

So for $P = 32$ at 8 KHz sampling rate, the resolution will be 160 Hz which will result in a wideband spectrum which smoothes the fine structure of the harmonics while preserving formant structure. For example in the case of vowels, the first three formants, important for speech intelligibility, are on the average 1 KHz apart [24] so they will be well identified with the Blackman-Tukey spectral estimator.

The resolution of the periodogram, on the other hand, is $0.89(2\pi/N)$ that is 28 Hz when $N = 256$ [22]. S o the periodogram will reveal unnecessary details for the present application.

C. Implementation

In this subsection we show how (18) and (21) are implemented as a matrix/vector multiplication and how they are used to calculate perceptually based "eigenvalues".

Define the eigenvalue vector $\boldsymbol{\lambda} = [\lambda_1 \lambda_2 \ldots \lambda_Q]^T$ and the vectors $\mathbf{v}_i = [v_i(0), \ldots, v_i(K-1)]^T$ for $i = 1, \ldots, Q$. Here $v_i(k)$ is the magnitude squared K-point DFT of $u_i(p)$, that is

$$v_i(k) = \left| V_i\left(\frac{2\pi k}{K}\right) \right|^2 \quad \text{for} \quad k = 0, \ldots, K-1. \tag{25}$$

Consider also the vector $\boldsymbol{\Phi}_B = [\Phi_B(0) \cdots \Phi_B(K-1)]^T$ obtained from the K-point DFT implementation of (20). Then (21) is implemented as follows:

$$\boldsymbol{\Phi}_B = \frac{1}{P} \sum_{i=1}^{Q} \lambda_i \mathbf{v}_i. \tag{26}$$

Or in matrix notation

$$\boldsymbol{\Phi}_B = \frac{1}{P} \mathbf{V} \boldsymbol{\lambda} \tag{27}$$

where $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_Q]$.

$\boldsymbol{\Phi}_B$ is used to calculate the masking threshold $\boldsymbol{\Phi}_{\mathrm{thr}}$ as outlined in Section III. Note that the sound pressure level in barks will then be given by

$$X(z(F_s k/K)) = \Phi_B(k) \tag{28}$$

where F_s is the sampling rate and $z(f)$ is given by (9).

This perceptual information is mapped to the eigendomain using (18) which is implemented as follows:

$$\boldsymbol{\theta} = \frac{1}{K} \mathbf{V}^T \boldsymbol{\Phi}_{\mathrm{thr}} \tag{29}$$

where $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_Q]^T$ and the θ_i's hereafter referred to as the "masking energies."

V. Handling the Colored Noise Case

One problem with the signal subspace approach described in Section II is that it is based on the white noise assumption. In [1], prewhitening is proposed as a remedy to this problem. Accordingly, the overall enhancing filter becomes

$$\bar{\mathbf{H}} = \mathbf{R}_\omega^{\frac{1}{2}} \mathbf{H} \mathbf{R}_\omega^{-\frac{1}{2}} \tag{30}$$

where $\mathbf{R}_\omega^{(1/2)}$ is the square root of the colored noise covariance matrix. We shall refer to this modified method as SSA with

Prewhitening (or PWSS). In [3] prewhitening is accomplished using a filter designed from the coefficients of an autoregressive model of the noise whereas in [7], prewhitening is an integral part of a quotient singular value decomposition based algorithm.

The eigenfilter $\mathbf{H}$ in (30) is now the solution of the optimization problem (6) after applying the prewhitening matrix to the input speech vector. Consequently, the noise shaping achieved is obtained according to a modified speech spectrum [2]. So the filter in (30) may not be the best choice to handle the colored noise case.

In this paper, we propose a solution similar to the one suggested in [2] and [4] and which was reported to outperform the PWSS. This solution consists of replacing the noise variance σ^2 in (8) with the noise energies in the directions of the eigenvectors of the speech covariance matrix. This is achieved by calculating ξ_i, the Raleigh quotient associated with $\mathbf{u}_i$ and $\mathbf{R}_w$ for $i = 1 \ldots Q$. Namely,

$$\xi_i = \mathbf{u}_i^H \mathbf{R}_w \mathbf{u}_i. \tag{31}$$

Now using the notation of Section IV, (31) can be written in a similar way to (18) as follows:

$$\xi_i = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Phi_\omega(\omega) |V_i(\omega)|^2 d\omega \tag{32}$$

where $\Phi_\omega(\omega)$ is a PSD estimate of $\mathbf{w}$. On matrix notation we have

$$\boldsymbol{\xi} = \frac{1}{K} \mathbf{V}^T \boldsymbol{\Phi}_\omega. \tag{33}$$

Computing the ξ_i's in this way requires less arithmetic operations than (31) because the matrix $\mathbf{V}$ is already available from the masking threshold computation.

VI. Implementation of the Proposed Method

In this section, we describe in detail the steps required to implement the proposed method.

Although the signal subspace approach outperforms the spectral subtraction methods [1], its major drawback remains the large computational load required to calculate the covariance matrix and especially its eigenvalue decomposition. To reduce this computational burden, we propose to calculate the signal subspace using a modified version of the method used in [1].

We divide the speech signal into overlapping frames of length N with a 50% overlap. The N samples are used to calculate the first P coefficients of the biased autocorrelation function, efficiently implemented using the Fast Fourier Transform. From these coefficients, the Toeplitz covariance matrix $\mathbf{R}_x$ is formed. An eigenfilter is designed using the eigenvalue decomposition of this covariance matrix. Every frame is divided into $(2N/P) - 1$ smaller P-dimensional overlapping vectors[2] with a 50% overlap.

Every such vector is then 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 to obtain one enhanced frame. Finally every frame is multiplied by a second Hanning window and the total enhanced speech signal is recovered using the overlap-add synthesis technique.

[2] N is chosen to be a multiple of P

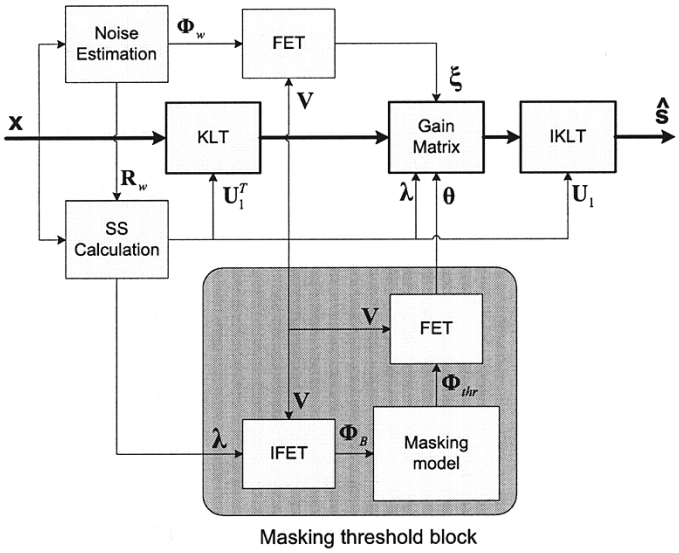

Fig. 2. Block diagram of the proposed method.

With this method, we need to calculate a new eigenfilter less frequently hence reducing the computational load. For example for $N = 256$ and $P = 32$, a new filter is designed every 17 vectors instead of every vector. Obviously, this technique assumes the speech signal to be stationary within one frame of length N. This assumption is reasonable in practice since it is also used in other speech enhancement methods such as the spectral subtraction method [11].

The steps required to calculate the perceptual eigenfilter are shown in the block diagram of Fig. 2 and are explained next.

1) Noise Statistics: During nonspeech activity periods, the biased autocorrelation function estimate of the noise, $r_w(p)$, is obtained. This estimate is both used to calculate the power spectrum $\boldsymbol{\Phi}_w(\omega)$ using the Blackman-Tukey estimator and to form the Toeplitz covariance matrix $\mathbf{R}_w$ of the noise.

Several methods for voice activity detection (VAD) have been proposed in literature [25]. These methods, such as the energy based methods, can be used to complement our proposed approach. However, since VAD is beyond the scope of this paper, voice activity periods had been manually labeled in the experimental results reported here.

2) Calculating the Signal Subspace: Let $\mathbf{R}_x$ denote the noisy covariance matrix estimated as explained above. Since the noise and the speech signal are assumed to be uncorrelated, the clean speech signal covariance matrix is estimated as $\mathbf{R}_s = \mathbf{R}_x - \mathbf{R}^w$ Next perform the eigenvalue decomposition on the matrix $\mathbf{R}_s$ and obtain the eigenvalue vector $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_Q]^T$, the eigenvector matrix $\mathbf{U}_1$ and the corresponding matrix $\mathbf{V}$ as discussed in Section IV-C. $\mathbf{R}_s$ is not guaranteed to be positive definite hence the rank Q of $\mathbf{R}_s$ is chosen to be the number of strictly positive eigenvalues of $\mathbf{R}_s$ [2].

3) Masking Threshold: Use the IFET, (27), and the FET, (29), to get the vector of masking energies $\boldsymbol{\theta}$ from $\boldsymbol{\lambda}$ according to the masking threshold as explained in Section IV-C.

4) Colored Noise Case: To handle the colored noise case, the IFET is used to calculate the noise energies, ξ_i's, in every spectral direction as explained in Section V.

5) KLT: The signal coefficients in the signal subspace are obtained by multiplying the signal vector by the KLT matrix $\mathbf{U}_1^H$.

6) Gain Matrix: The signal coefficients are multiplied by the diagonal gain matrix $\mathbf{G}$. The entries of the matrix are calculated as follows:

$$g_i = e^{-\nu \xi_i / \min(\lambda_i, \theta_i)}. \tag{34}$$

Usually $\theta_i < \lambda_i$, so by replacing λ_i with θ_i in the gain function, g_i becomes smaller and hence more noise suppression is achieved. However since the gain is now obtained via perceptual criteria, the control parameter ν can be reduced in order to obtain less signal distortion without making the residual noise more audible. Nonetheless, during weak energy frames, such as unvoiced fricatives, the spectrum is rarely characterized in terms of formants because low frequencies are not excited and the excited upper resonances have broad bandwidths [24]. In this case, the masking threshold estimate is not accurate and it can happen that λ_i be smaller than θ_i with the result that, if θ_i is used, not enough noise reduction is achieved, due to estimation errors. In particular, at transitions from silence to speech activity periods, the residual noise has a non smooth character which may be uncomfortable to some listeners. Our informal listening tests show that the use of the minimum operation in (34) helps to improve the performance.

7) IKLT: The enhanced signal vector is finally recovered in the signal subspace using the inverse KLT matrix $\mathbf{U}_1$.

The method described above is called the perceptual SS method (PSS). Now, in order to evaluate the merit of using masking, a second method has also been tested. It is exactly similar to PSS but without the masking threshold block. The gain function in this case becomes simply

$$q_i = e^{-\nu \xi_i / \lambda_i} \tag{35}$$

and the ξ_i's are now calculated directly using (31). Since (31) involves the use of the Raleigh Quotient, this method will be referred to as the Raleigh Quotient Signal Subspace (RQSS) method.

VII. Experimental Results

To evaluate the performance of the proposed method, listening experiments were carried out using different speech signals and background noises, where all recordings had a 8 KHz sampling rate. The following parameters were used: $P = 32$ and $N = K = 256$. Several values for the gain function control parameter ν were tested. However, to achieve an acceptable noise reduction level without seriously degrading the desired speech intelligibility, the values $\nu = 2$ (for RQSS and PWSS) and $\nu = 0.8$ (for PSS) were preferred. For comparison purposes, this choice also aimed to maintain the same level of distortion across the three tested methods while keeping some audible residual noise.

During informal listening tests, it was concluded that the proposed method outperformed the PWSS and RQSS in several background noise types and levels and with different sentences. Particularly, the benefit of PSS was more evident at low SNR conditions [26].

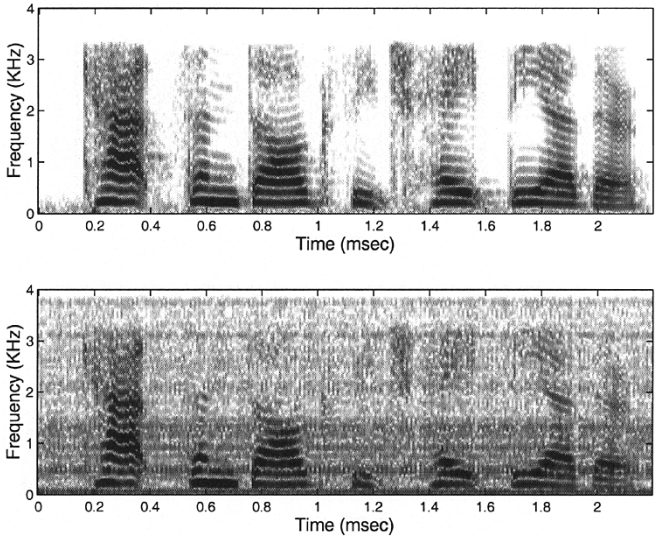

Fig. 3. Spectrograms of the clean (top) and noisy (bottom) female sentence when corrupted with a freezer motor noise.

These results have been confirmed by formal subjective listening tests. In these tests, two sentences and four noise types have been selected from those investigated during informal listening tests. The sentences were a 2.2-s long female sentence (*Cats and dogs each hate the other*) and a 3-s long male sentence (*Post no bills on this office wall*). The noises were those of a Volvo car (VLV), a Leopard military vehicle (LEO), an F16 jet cockpit (JET) and a freezer motor (FRZ). The noises were added to the clean speech signals at a 0 dB segmental SNR except for VLV where the SNR was -5 dB. This was due to the lowpass nature of the Volvo car noise making it relatively difficult for the subjects to discriminate between the different methods' performances at a higher SNR.

In total 18 persons took part of the tests among which three worked in the speech processing area but were unfamiliar with the sentences. The majority of the subjects were in their late twenties.

Fig. 3 shows the spectrogram of the Female sentence used in the tests. Shown are the clean signal and the FRZ corrupted signal. Fig. 4 shows the spectrograms of the same signal enhanced with PWSS and PSS. It can be seen that PSS results in a less noisy signal while maintaining the same level of signal distortion.

A. The A-B Test Results

In this test, the subjects were asked to evaluate the performance of PSS against that of RQSS and PWSS. In total, 8 pairs of recordings per tests were presented to the subjects where each pair consisted of a speech signal enhanced using PSS and a second enhanced with a competing method. A separate test has been conducted for every sentence. For each pair, they were asked to vote for the signal they preferred (A, B, or X if they had no preference) according to three different criteria. Intelligibility: "*which signal is easier to understand ?*" quality "*which*

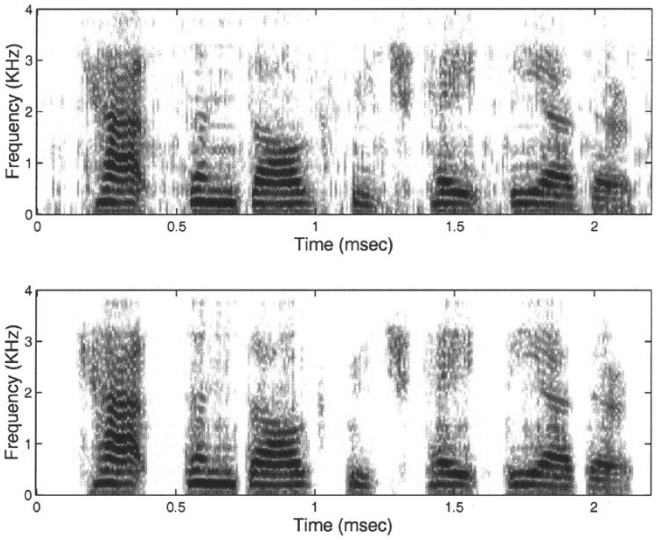

Fig. 4. Spectrograms of the signal in Fig. 3 enhanced with PWSS (top) and PSS (bottom).

TABLE I
A-B TEST PREFERENCE RESULTS FOR THE FEMALE SENTENCE

Noise	Signal Distortion					
	PSS	X	PWSS	PSS	X	RQSS
FRZ	**20%**	73%	7%	**20%**	73%	7%
VLV	**33%**	40%	27%	**27%**	40%	33%
JET	**14%**	50%	36%	**33%**	47%	20%
LEO	**53%**	40%	7%	**33%**	53%	13%
Noise	Residual Noise					
	PSS	X	PWSS	PSS	X	RQSS
FRZ	**67%**	27%	7%	**60%**	40%	0%
VLV	**73%**	20%	7%	**47%**	40%	13%
JET	**57%**	43%	0%	**47%**	47%	7%
LEO	**87%**	0%	13%	**80%**	13%	7%
Noise	Overall					
	PSS	X	PWSS	PSS	X	RQSS
FRZ	**67%**	27%	7%	**60%**	33%	7%
VLV	**80%**	13%	7%	**67%**	20%	13%
JET	**50%**	43%	7%	**47%**	40%	13%
LEO	**87%**	13%	0%	**80%**	20%	0%

signal is less noisy?" and overall: "*putting the previous two criteria together, which signal is preferred?*"

Some initial results showed that including the noisy signal in the test and making the subjects explicitly aware of the signal distortion (due to the test design), lead to biased answers. Therefore, we have decided to remove the noisy signals from the test. The benefit of using PWSS and PSS over the original noisy signal has been reported in [1] and [27] respectively.

Tables I and II show the results of this test for the female and male sentences respectively. It can be seen that the PSS method outperforms the other two methods especially for the female sentence. In general, the subjects found that the three methods provided a relatively similar amount of distortion to the enhanced signals, with the exception on the Female-LEO and Male-VLV cases where the use of PSS resulted also in a less distorted signal than PWSS. Overall, the merit of PSS is in that it succeeds to maintain an acceptable level of distortion

TABLE II
A-B TEST PREFERENCE RESULTS FOR THE MALE SENTENCE

Noise	Signal Distortion					
	PSS	X	PWSS	PSS	X	RQSS
FRZ	**28%**	56%	17%	**28%**	61%	11%
VLV	**72%**	22%	6%	**47%**	29%	24%
JET	**17%**	61%	22%	**11%**	72%	17%
LEO	**39%**	39%	22%	**50%**	28%	22%
Noise	Residual Noise					
	PSS	X	PWSS	PSS	X	RQSS
FRZ	**67%**	22%	11%	**50%**	33%	17%
VLV	**89%**	11%	0%	**59%**	41%	0%
JET	**44%**	39%	17%	**28%**	67%	6%
LEO	**89%**	6%	6%	**83%**	17%	0%
Noise	Overall					
	PSS	X	PWSS	PSS	X	RQSS
FRZ	**61%**	33%	6%	**56%**	39%	6%
VLV	**89%**	6%	6%	**71%**	18%	12%
JET	**39%**	44%	17%	**39%**	56%	6%
LEO	**83%**	11%	6%	**67%**	17%	17%

while offering a better noise reduction (masking) performance. PSS had a considerable success over RQSS and PWSS in the case of LEO, VLV, and to a less extent FRZ. Note that when the subjects *did not* vote for PSS, that was mostly because they were unable to perceive any difference rather than because PSS had a poorer performance.

In the JET case, the improvement achieved over the two competing methods was not as obvious as it is with the other noises. The test results revealed that relatively many subjects found that PSS had the same performance as PWSS and RQSS. The explanation for this is that the spectral characteristics of the JET noise (lowpass with an additional peak at 2.8 KHz) has affected the estimated masking threshold. This estimate was not accurate enough resulting in an increased signal distortion, while still maintaining a high noise reduction performance. PWSS and RQSS, on the other hand, failed to completely suppress that high frequency peak. For this reason, the number of subjects who could not decide whether they prefer a low residual noise or a low signal distortion, was high for this particular noise type.

B. Residual Noise Shaping Score

During informal listening tests, we have observed that signals enhanced with PSS have a residual noise characteristics which are relatively similar, regardless of the original corrupting noise. This result supports our claim that PSS yields improved noise shaping and hence better masking. To confirm this result, we have set up a new subjective test which provides a "*residual noise shaping score*" which serves to compare the performances of the different methods according to the above mentioned criterion.

The subjects were presented with a pair of signals enhanced by the *same* method but corresponding to different noises. Then they were asked to *concentrate* just on the background noise and to compare its characteristics in the two recordings. The comparison is based on how similar or different these characteristics are in the two signals, regardless of the loudness. The sub-

TABLE III
RATING SCHEME FOR THE RESIDUAL NOISE SHAPING SCORE TEST

1	:	Completely different
2	:	Different
3	:	Don't know
4	:	Similar
5	:	Very similar

TABLE IV
RESIDUAL NOISE SHAPING SCORES FOR THE FEMALE (F) AND MALE (M) SENTENCES

Noise pair	PSS		RQSS		PWSS	
	F	M	F	M	F	M
LEO, FRZ	4.3	4.2	2.3	2.2	1.5	1.1
VLV, FRZ	4.1	3.8	2.5	2.1	1.2	1.4
LEO, VLV	4.7	4.5	2.1	1.8	1.5	1.7
LEO, JET	3.3	2.8	1.5	1.2	1.8	1.6
VLV, JET	3.0	2.7	2.1	1.9	2.0	2.1
JET, FRZ	3.8	3.7	3.1	2.9	2.7	3.1
Average	**3.9**	**3.6**	**2.3**	**2.0**	**1.7**	**1.8**

jects had to score their decision according to a five-level rating scheme shown in Table III. Again we have used the same four noises resulting in six pairs for every method. In total, for the three methods, 18 pairs per test were presented to the subjects. Two tests, one for every sentence, had been designed.

The detailed scores for the different noise pairs for the two sentences are given in Table IV. It can be seen that PSS got a higher score on average than RQSS and PWSS which shows that it achieves a relatively better noise shaping than the other two competing methods.

VIII. 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 capable of enhancing 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 method are relatively similar regardless of the original corrupting noise.

APPENDIX

In this Appendix, we prove the FET relationships (18) and (21). To prove (18), we proceed as follows.

Proof: By definition the eigenvalue λ_i can be written

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

Using the relationship between the autocorrelation function estimate and the periodogram

$$r(p) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Phi(\omega) e^{j\omega p} d\omega \tag{36}$$

we have

$$\lambda_i = \sum_{p=0}^{P-1} \sum_{q=0}^{P-1} u_i^*(p) u_i(q) \frac{1}{2\pi} \int_{-\pi}^{\pi} \Phi(\omega) e^{e\omega(p-q)} d\omega. \tag{37}$$

Recalling the definition of $V_i(\omega)$ (19) we get

$$\lambda_i = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Phi(\omega) V_i(\omega) V_i^*(\omega)\, d\omega. \tag{38}$$

■

The proof of (21) is as follows.

Proof: Consider the Blackman-Tukey estimate (20), assuming a triangular window, i.e., $w_b(p) = 1 - (|p|/P)$ for $|p| < P$ we have

$$\Phi_B(\omega) = \frac{1}{P} \sum_{p=-P+1}^{P-1} r(p)(P - |p|) e^{-j\omega p}. \tag{39}$$

The above summation over p is readily expressible as a double summation as

$$\Phi_B(\omega) = \frac{1}{P} \sum_{p=0}^{P-1} \sum_{q=0}^{P-1} r(p-q) e^{-j\omega(p-q)}. \tag{40}$$

From the eigenvalue decomposition formula $\mathbf{R} = \sum_i \lambda_i \mathbf{u}_i \mathbf{u}_i^H$, we note that

$$r(p-q) = \sum_{i=1}^{P} \lambda_i u_i(p) u_i^*(q). \tag{41}$$

Substituting (41) into (40) and recalling the definition of $V_i(\omega)$ (19), we finally obtain

$$\Phi_B(\omega) = \frac{1}{P} \sum_{i=1}^{Q} \lambda_i |V_i(\omega)|^2 \tag{42}$$

where the limit of the summation is changed from P to Q because $\lambda_i = 0$ for $i > Q$. ■

REFERENCES

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

[2] 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, Mar. 2000.

[3] J. Huang and Y. Zhao, "An energy-constrained signal subspace method for speech enhancement and recognition in colored noise," *Speech Commun.*, vol. 1, pp. 165–181, 1998.

[4] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 87–95, Feb. 2001.

[5] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: A regenerative approach," *Speech Commun.*, vol. 10, pp. 45–57, Feb. 1991.

[6] P. S. K. Hansen, P. C. Hansen, S. D. Hansen, and J. A. Sorensen, "Experimental comparison of signal subspace based noise reduction methods," in *Proc. ICASSP'99*, vol. 1, 1999, pp. 101–104.

[7] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sorensen, "Reduction of broad-band noise in speech by truncated QSVD," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 439–448, Nov. 1995.

[8] J. Huang and Y. Zhao, "A DCT-based fast signal subspace technique for robust speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 747–751, Nov. 2000.

[9] R. Vetter, "Single channel speech enhancement using MDL-based subspace approach in bark domain," in *Proc. ICASSP'01*, vol. 1, 2001, pp. 641–644.
[10] M. Jeppesen, C. A. Rodbro, and S. H. Jensen, "Recursively updated eigenfilterbank for speech enhancement," in *Proc. ICASSP'01*, vol. 1, 2001, pp. 653–656.
[11] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, pp. 113–120, Apr. 1979.
[12] 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, Mar. 1999.
[13] S. Gustafsson, P. Jax, and P. Vary, "A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics," in *Proc. ICASSP'98*, 1998, pp. 397–400.
[14] A. Czyzewski and R. Krolikowski, "Noise reduction in audio signals based on the perceptual coding approach," *Proc. IEEE WASPAA*, pp. 147–150, 1999.
[15] A. A. Azirani, R. J. Le Bouquin, and G. Faucon, "Optimizing speech enhancement by exploiting masking properties of the human ear," in *Proc. ICASSP'95*, 1995, pp. 800–803.
[16] 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.
[17] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J Select. Areas Commun.*, vol. 6, pp. 314–323, Feb. 1988.
[18] E. Zwicker and H. Fastl, *Psychoacoustics*, Berlin, Germany: Springer-Verlag, 1990.
[19] K. Brandenburg and G. Stoll, "ISO-MPEG-1 audio: A generic standard for coding of high quality digital audio," *J. Audio Eng. Soc.*, vol. 42, pp. 780–792, Oct. 1994.
[20] C. Colomes, M. Lever, J. B. Rault, Y. F. Dehery, and G. Faucon, "A perceptual model applied to audio bit rate reduction," *J. Audio Eng. Soc.*, vol. 43, pp. 233–240, Apr. 1995.
[21] J. G. Beerends and J. A. Stemerdink, "A perceptual audio quality measure based on psychoacoustic sound representation," *J. Audio Eng. Soc.*, vol. 40, pp. 963–978, Dec. 1992.
[22] M. H. Hayes, *Statistical Digital Signal Processing and Modeling.* New York: Wiley, 1996.
[23] S. M. Kay, *Modern Spectral Estimation Theory and Applications*. Englewood Cliffs, NJ: Prentice-Hall, 1988.
[24] D. O'Shaughnessy, *Speech Communications Human and Machine*, 2nd ed. New York: IEEE Press, 2000.
[25] J. R. Deller and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[26] 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, Germany, 2001, pp. 199–202.
[27] F. Jabloun and B. Champagne, "A perceptual signal subspace approach for speech enhancement in colored noise," in *Proc. ICASSP'02*, vol. 1, 2002, pp. 569–572.

Firas Jabloun was born in Kelibia, Tunisia, in 1974. He received the B.S. and M.S. degrees both in electrical engineering from Bilkent University, Ankara, Turkey in 1996 and 1998, respectively. He is currently pursuing the Ph.D. degree in the Department of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada.

In October 2002, he joined the Toshiba Cambridge Research Laboratory, Cambridge, U.K., as a member of the Speech Technology Group. His research interests include single and multimicrophone speech enhancement, robust speech recognition, and acoustic modeling.

Benoît Champagne received the B.Ing. degree in engineering physics from the Ecole Polytechnique, Montreal, QC, Canada, in 1983, the M.Sc. degree in physics from the University of Montreal in 1985, and the Ph.D. degree in electrical engineering from the University of Toronto, Toronto, ON, Canada, in 1990.

He joined INRS-Télécommunications, Université du Québec as an Assistant Professor in 1990, and was promoted to the rank of Associate Professor in 1995. In September 1999, he joined McGill University, Montreal, as an Associate Professor with the Department of Electrical and Computer Engineering. He remains a Visiting Professor at INRS. His research interests span many areas of digital signal processing, including detection and estimation, sensor array processing, adaptive filtering, and applications thereof in telecommunication engineering. He regularly serves as a reviewer for scientific journals and granting organizations.

Dr. Champagne is a member of the Order des ingénieurs du Québec and a member of the IEEE Signal Processing Society.