A FAMILY OF BAYESIAN STSA ESTIMATORS FOR THE ENHANCEMENT OF SPEECH WITH CORRELATED FREQUENCY COMPONENTS

Eric Plourde and Benoît Champagne

McGill University
Department of Electrical and Computer Engineering
Montreal, Quebec, Canada, H3A 2A7
e-mail: eric.plourde@mail.mcgill.ca, benoit.champagne@mcgill.ca

ABSTRACT

In Bayesian short-time spectral amplitude (STSA) estimation for single-channel speech enhancement, the spectral components are traditionally assumed uncorrelated. However, this assumption is inexact since some correlation is present in practice. We thus investigate a multi-dimensional STSA estimator that assumes correlated frequency components. Since the closed-form solution of this optimum estimator is not readily available, we previously derived closed-form expressions for an upper and a lower bound on the desired estimator. In this paper, we study the proximity between the upper and the lower bounds. Moreover, we propose a new family of speech enhancement estimators that are derived from these bounds and characterized by a scalar parameter $0 \leq \gamma \leq 1$, with $\gamma = 0$ corresponding to the lower bound and $\gamma = 1$ to the upper bound. Results using the wideband Perceptual Evaluation of Speech Quality (PESQ) and Log-likelihood Ratio (LLR) measures as well as informal listening experiments show that the newly proposed estimators achieve a better performance than existing estimators, especially at high SNR.

Index Terms— Speech enhancement, Bayesian estimation, short-time spectral amplitude

1. INTRODUCTION

Speech enhancement algorithms are used to remove background noise in a speech signal. They are present in many common devices such as cell phones and hearing aids. In the Baysian short-time spectral amplitude (STSA) estimation approach [1], an estimator of the clean speech is derived by minimizing the statistical expectation of a cost function that penalizes errors in the clean speech estimate. In this approach, it is always assumed that the different spectral components of speech are uncorrelated so that the different frequency components of the noisy speech can be processed independently.

The latter assumption is however inexact as there are some sources of correlation between the spectral components [2]. Firstly, the use of a finite window function in the short-time processing introduces some correlation between adjacent frequency components. This is due to the spectral smearing phenomenon which is a known effect of the windowing process [3]. Secondly, voiced speech is characterized by the vibration of the vocal cords at a fundamental frequency F0 and has several harmonics at multiples of F0 [4]. The frequencies corresponding to these different harmonics will therefore be inherently correlated. The estimators following the traditional uncorrelated approach are thus sub-optimal. Correlated frequency components in Bayesian STSA estimation has apparently not been considered by other authors in the recent speech and audio litterature.

We investigate a multi-dimensional Bayesian STSA estimator that considers the spectral components to be correlated. Since a closed-form solution for such an estimator is not readily available, we previously developed closed-form expressions for a lower and an upper bound on the desired estimator [5]. Here, we study the proximity between the upper and lower bounds. Moreover, we propose a new family of speech enhancement estimators that are derived from these bounds and characterized by a scalar parameter $0 \leq \gamma \leq 1$, with $\gamma = 0$ corresponding to the lower bound and $\gamma = 1$ to the upper bound. Knowledge of the clean speech and noise correlation matrices is needed to implement the proposed estimators. Since speech is mostly correlated in voiced parts, we also modify the clean speech correlation matrix to give it a full structure in voiced sections and a diagonal structure in unvoiced sections. Informal listening experiments as well as results using the wideband Perceptual Evaluation of Speech Quality (PESQ) and Log-likelihood Ratio (LLR) measures show that the proposed estimators achieve better performance than the benchmark estimators for several noise types and signal-to-noise ratio (SNR) conditions.

The paper is organized as follows. In Section 2, we briefly describe the bounds derived in [5] and present the proposed family of estimators. Section 3 studies the proximity between the upper and lower bounds and addresses the estimation of the associated correlation matrices. Section 4 presents the experimental results and Section 5 concludes the work.

The following notation is used in this article: for any vector $\mathbf{A} = [a_k] \in \mathbb{R}^{N\times 1}$ and any positive real α, we define $\mathbf{A}^{[\alpha]} = [a_k^\alpha]$; for any vector $\mathbf{A} \in \mathbb{C}^{N\times 1}$, we define $|\mathbf{A}| = [|a_k|]$; for any matrix $\mathbf{A} \in \mathbb{C}^{N\times N}$ we define $\mathrm{diag}\{\mathbf{A}\}$ as the column vector containing the diagonal elements of matrix $\mathbf{A}$; $\mathbf{I}_N$ is the $N \times N$ identity matrix.

2. MULTI-DIMENSIONAL FAMILY OF STSA ESTIMATORS ALLOWING CORRELATED FREQUENCY COMPONENTS

In this section, we proceed to obtain the family of multi-dimensional clean speech STSA estimators that assume correlated frequency components.

Let $\mathbf{Y}_i = \mathbf{X}_i + \mathbf{W}_i$ be an N-dimensional column vector representing the short-time Fourier transform (STFT) coefficients of noisy speech observations for time frame i. $\mathbf{X}_i$ and $\mathbf{W}_i$ are respec-

This work was supported by the *Fonds québécois de la recherche sur la nature et les technologies* and a grant from the *Natural Sciences and Engineering Research Council of Canada*.

tively the clean speech STFT vector and the noise STFT vector. To simplify the notation, we will usually omit the subscript i and consider the processing of one particular frame. The elements of $\mathbf{X}$ are $X_k = \mathcal{X}_k e^{j\alpha_k}, 1 \leq k \leq N$, where $\mathcal{X}_k$ is the positive and real STSA and $\alpha \in [-\pi, \pi)$. We also define $\boldsymbol{\mathcal{X}} = [\mathcal{X}_1\ \mathcal{X}_2\ \cdots\ \mathcal{X}_N]^T$ and $\boldsymbol{\alpha} = [\alpha_1\ \alpha_2\ \cdots\ \alpha_N]^T$. We assume that $\mathbf{X}$ and $\mathbf{W}$ are independent, zero-mean and circular Gaussians with probability density functions:

$$f_{\mathbf{X}}(\mathbf{X}) = \frac{1}{\pi^N \det(\mathbf{R}_{\mathbf{X}})} e^{-\mathbf{X}^H \mathbf{R}_{\mathbf{X}}^{-1} \mathbf{X}}, \quad (1)$$

$$f_{\mathbf{W}}(\mathbf{W}) = \frac{1}{\pi^N \det(\mathbf{R}_{\mathbf{W}})} e^{-\mathbf{W}^H \mathbf{R}_{\mathbf{W}}^{-1} \mathbf{W}}. \quad (2)$$

In these expressions $\mathbf{R}_{\mathbf{X}} = E\{\mathbf{X}\mathbf{X}^H\}$ and $\mathbf{R}_{\mathbf{W}} = E\{\mathbf{W}\mathbf{W}^H\}$ are the correlation matrices of the clean speech and of the noise respectively, superscript H indicates the conjugate transpose and $\mathbf{R}_{\mathbf{W}} > 0$ (positive definite) is assumed. Traditional Bayesian STSA estimation approaches (e.g. [1]) assume that $\mathbf{R}_{\mathbf{X}}$ and $\mathbf{R}_{\mathbf{W}}$ are diagonal matrices, i.e. the spectral components are uncorrelated. In this work, we do not enforce such diagonality constraint. Our model therefore considers possible frequency correlations in the clean speech and noise.

We want to evaluate the minimum mean square error (MMSE) estimator of $\boldsymbol{\mathcal{X}}$:

$$\hat{\boldsymbol{\mathcal{X}}}^o = \underset{\hat{\boldsymbol{\mathcal{X}}}}{\operatorname{argmin}}\, E\{\|\boldsymbol{\mathcal{X}} - \hat{\boldsymbol{\mathcal{X}}}\|^2\} \quad (3)$$

where the minimum is over all possible functions $\hat{\boldsymbol{\mathcal{X}}} \equiv \hat{\boldsymbol{\mathcal{X}}}(\mathbf{Y})$ of the observation vector $\mathbf{Y}$. We note that the cost function in (3), i.e. $C(\boldsymbol{\mathcal{X}}, \hat{\boldsymbol{\mathcal{X}}}) \triangleq \|\boldsymbol{\mathcal{X}} - \hat{\boldsymbol{\mathcal{X}}}\|^2$, considers all the STSA frequency components jointly. Using matrix calculus, we can show that (3) leads to:

$$\hat{\boldsymbol{\mathcal{X}}}^o = E\{\boldsymbol{\mathcal{X}}|\mathbf{Y}\} \quad (4)$$

i.e. the N-dimensional conditional expectation of $\boldsymbol{\mathcal{X}}$ given the complete vector of observations $\mathbf{Y}$. This estimator can then be combined with the phase of the noisy speech, for each frequency, to yield the estimator of $\mathbf{X}$:

$$\hat{\mathbf{X}}^o = [\hat{\mathcal{X}}_0^o e^{j\measuredangle Y_0}, \cdots, \hat{\mathcal{X}}_{N-1}^o e^{j\measuredangle Y_{N-1}}]^T. \quad (5)$$

Unfortunately a closed-form expression for (4) is not readily available. Since the $\hat{\mathcal{X}}_k^o$ are positive real quantities, we approached the problem of finding a realizable approximation to (4), in [5], by obtaining tractable upper and lower bounds, $\hat{\mathcal{X}}_{U,k}^o$ and $\hat{\mathcal{X}}_{L,k}^o$ respectively, such that $\hat{\mathcal{X}}_{L,k}^o < \hat{\mathcal{X}}_k^o < \hat{\mathcal{X}}_{U,k}^o$. In the next subsections, we briefly review these bounds and we then propose a new parameterized family of estimators that is based on these lower and upper bounds.

2.1. Lower Bound

Using the triangle inequality for integration [6], we can show that:

$$|E\{X_k|\mathbf{Y}\}| \leq E\{\mathcal{X}_k|\mathbf{Y}\}. \quad (6)$$

As a lower bound on the desired estimator (4), we therefore propose $\hat{\mathcal{X}}_{L,k}^o = |E\{X_k|\mathbf{Y}\}|$ or equivalently:

$$\hat{\boldsymbol{\mathcal{X}}}_L^o = |E\{\mathbf{X}|\mathbf{Y}\}|. \quad (7)$$

Under the Gaussian statistical model for the clean speech and noise presented previously, the term $E\{\mathbf{X}|\mathbf{Y}\}$ is the MMSE estimator of $\mathbf{X}$, which is known to be equal to [2]:

$$E\{\mathbf{X}|\mathbf{Y}\} = \hat{\mathbf{X}}_{\text{MMSE}} = \mathbf{G}_{\text{MMSE}}\mathbf{Y} \quad (8)$$

where the MMSE gain matrix $\mathbf{G}_{\text{MMSE}}$ is:

$$\mathbf{G}_{\text{MMSE}} \triangleq \mathbf{R}_{\mathbf{X}}(\mathbf{R}_{\mathbf{X}} + \mathbf{R}_{\mathbf{W}})^{-1}. \quad (9)$$

A lower bound on the desired estimator is therefore:

$$\hat{\boldsymbol{\mathcal{X}}}_L^o = \left|\mathbf{G}_{\text{MMSE}}\mathbf{Y}\right|. \quad (10)$$

Note that in the special case of uncorrelated frequency components (i.e. the traditional framework), $\mathbf{R}_{\mathbf{X}}$ and $\mathbf{R}_{\mathbf{W}}$ in (9) are diagonal matrices. Then, combining (10) with the phase of the noisy speech yields:

$$\hat{X}_k = \frac{S_{X,k}}{S_{X,k} + S_{W,k}} Y_k \quad (11)$$

where $S_{X,k} = [\mathbf{R}_{\mathbf{X}}]_{kk} = E\{\mathcal{X}_k^2\}$ and $S_{W,k} = [\mathbf{R}_{\mathbf{W}}]_{kk} = E\{|W_k|^2\}$. The processing of each frequency is therefore decoupled and the corresponding operation amounts to a standard Wiener filter.

2.2. Upper Bound

Using Jensen's inequality [7], we have for a real convex function φ:

$$\varphi(E\{\mathcal{X}_k|\mathbf{Y}\}) \leq E\{\varphi(\mathcal{X}_k)|\mathbf{Y}\}. \quad (12)$$

If we set $\varphi(a) = a^2$, we obtain $E\{\mathcal{X}_k|\mathbf{Y}\}^2 \leq E\{\mathcal{X}_k^2|\mathbf{Y}\}$ and,

$$E\{\mathcal{X}_k|\mathbf{Y}\} \leq \sqrt{E\{\mathcal{X}_k^2|\mathbf{Y}\}} \quad (13)$$

which is also a special case of Lyapunov's inequality [8]. As an upper bound on the desired estimator (4), we thus proposed $\hat{\mathcal{X}}_{U,k}^o = \sqrt{E\{\mathcal{X}_k^2|\mathbf{Y}\}}$ or equivalently:

$$\hat{\boldsymbol{\mathcal{X}}}_U^o = E\{\boldsymbol{\mathcal{X}}^{[2]}|\mathbf{Y}\}^{[1/2]}. \quad (14)$$

We showed previously [5] that this upper bound is given by:

$$\hat{\boldsymbol{\mathcal{X}}}_U^o = \left(|\mathbf{G}_{\text{MMSE}}\mathbf{Y}|^{[2]} + \text{diag}\{\mathbf{G}_{\text{MMSE}}\mathbf{R}_{\mathbf{W}}\}\right)^{[1/2]}. \quad (15)$$

Since the upper bound includes the lower bound and an additional positive term, it will obviously be greater than the lower bound.

2.3. Proposed Family of Estimators

The true estimator $\hat{\mathcal{X}}_k^o$ is smaller than $\hat{\mathcal{X}}_{U,k}^o$ and greater than $\hat{\mathcal{X}}_{L,k}^o$. Based on the expressions of the derived bounds $\hat{\mathcal{X}}_{L,k}^o$ and $\hat{\mathcal{X}}_{U,k}^o$ in [5], we propose here the following family of estimators:

$$\hat{\boldsymbol{\mathcal{X}}}_\gamma^o = \left(|\mathbf{G}_{\text{MMSE}}\mathbf{Y}|^{[2]} + \gamma\,\text{diag}\{\mathbf{G}_{\text{MMSE}}\mathbf{R}_{\mathbf{W}}\}\right)^{[1/2]} \quad (16)$$

where $0 \leq \gamma \leq 1$. We have that $\hat{\mathcal{X}}_{L,k}^o \leq \hat{\mathcal{X}}_{\gamma,k}^o \leq \hat{\mathcal{X}}_{U,k}^o$ with the limit cases:

$$\hat{\boldsymbol{\mathcal{X}}}_\gamma^o = \begin{cases} \hat{\boldsymbol{\mathcal{X}}}_U^o & \text{if } \gamma = 1 \\ \hat{\boldsymbol{\mathcal{X}}}_L^o & \text{if } \gamma = 0. \end{cases} \quad (17)$$

As in (5), the spectral amplitude estimators $\hat{\boldsymbol{\mathcal{X}}}_L^o$, $\hat{\boldsymbol{\mathcal{X}}}_U^o$ and $\hat{\boldsymbol{\mathcal{X}}}_\gamma^o$ are then combined with the phase of the noisy speech to obtain the complex spectrum estimators $\hat{\mathbf{X}}_L^o$, $\hat{\mathbf{X}}_U^o$ and $\hat{\mathbf{X}}_\gamma^o$ respectively.

3. OTHER CONSIDERATIONS

3.1. Upper and Lower Bound Proximity Analysis

In this section, we study the proximity between the lower and upper bounds. Since $\hat{\mathcal{X}}^o_{U,k}$ and $\hat{\mathcal{X}}^o_{L,k}$ are both positive terms and $\hat{\mathcal{X}}^o_{U,k} > \hat{\mathcal{X}}^o_{L,k}$, we consider the vector

$$\mathbf{B} = (\hat{\boldsymbol{\mathcal{X}}}_U^{o\,[2]} - \hat{\boldsymbol{\mathcal{X}}}_L^{o\,[2]})./\operatorname{diag}\{\mathbf{R_X}\} \tag{18}$$

as a proximity measure where ./ indicates an element-wise division. Each element B_k of vector $\mathbf{B}$ is therefore a difference of squared values normalized by $S_{X,k} = E\{\mathcal{X}_k^2\}$. From (9), (10) and (15), we have :

$$\mathbf{B} = \operatorname{diag}\{\mathbf{G}_{\text{MMSE}}\mathbf{R_W}\}./\operatorname{diag}\{\mathbf{R_X}\} \tag{19}$$

$$= \operatorname{diag}\{\mathbf{R_X}(\mathbf{R_X}+\mathbf{R_W})^{-1}\mathbf{R_W}\}./\operatorname{diag}\{\mathbf{R_X}\}. \tag{20}$$

Therefore, the second term in (15) dictates how tight are the bounds. Interestingly, this term does not depend on $\mathbf{Y}$ (however, in practice, the estimation of $\mathbf{R_X}$ does).

3.1.1. Uncorrelated Frequencies

To gain some insight into the behavior of the proximity vector $\mathbf{B}$, let us first consider uncorrelated frequency components. In that case, the k^{th} entry of $\mathbf{B}$ reduces to:

$$B_k = \frac{S_{W,k}}{S_{X,k}+S_{W,k}} = \frac{1}{1+\text{SNR}_k} \tag{21}$$

where $\text{SNR}_k = S_{X,k}/S_{W,k}$. For a high SNR_k, we have $B_k \to 0$, while for a low SNR_k $B_k \to 1$. Thereore, the bounds will be tighter as the SNR_k gets higher.

3.1.2. Correlated Frequencies

We next consider the case of correlated frequency components. $\mathbf{B}$ can be written in a form apparented to that of (21):

$$\mathbf{B} = \operatorname{diag}\{\mathbf{R_X}(\mathbf{I}_N + \mathbf{R_W}^{-1}\mathbf{R_X})^{-1}\}./\operatorname{diag}\{\mathbf{R_X}\}. \tag{22}$$

Let $\mu_{\text{max}} = \mu_N \geq \cdots \geq \mu_1 = \mu_{\text{min}}$ denote the eigenvalues of $\mathbf{R_W}^{-1/2}\mathbf{R_X}\mathbf{R_W}^{-1/2}$. It can be shown that, on the one hand, if $\mu_{\text{min}} >> 1$ (high SNR), then $\mathbf{B} \to \operatorname{diag}\{\mathbf{R_W}\}./\operatorname{diag}\{\mathbf{R_X}\}$ while on the other hand, if $\mu_{\text{max}} << 1$ (low SNR), then $\mathbf{B} \to \mathbf{1}_{N\times 1}$, where $\mathbf{1}_{N\times 1}$ denotes an N-dimensional column vector of ones. Therefore, again, the bounds will be tighter as the SNR gets higher.

3.2. Estimating $\mathbf{R_X}$ and $\mathbf{R_W}$

To compute $\hat{\boldsymbol{\mathcal{X}}}^o_L$ (10), $\hat{\boldsymbol{\mathcal{X}}}^o_U$ (15) or $\hat{\boldsymbol{\mathcal{X}}}^o_\gamma$ (16), one needs an estimation of matrices $\mathbf{R_X}$ and $\mathbf{R_W}$. We shall denote the estimates of $\mathbf{R_X}$, $\mathbf{R_W}$ and $\mathbf{R_Y}$ for the i^{th} frame by $\hat{\mathbf{R}}_{\mathbf{X},i}$, $\hat{\mathbf{R}}_{\mathbf{W},i}$ and $\hat{\mathbf{R}}_{\mathbf{Y},i}$ respectively. These are computed as in [5]. However, we also experiment in this paper with a modified structure of the estimator $\hat{\mathbf{R}}_{\mathbf{X},i}$ to take into account the nature of the current frame, i.e. voiced vs. unvoiced.

Indeed, since the correlation due to the harmonics of the fundamental frequency is only present in the voiced parts of speech, it would be appropriate to consider a diagonal $\hat{\mathbf{R}}_{\mathbf{X},i}$ in unvoiced parts and a full (i.e. unconstrained) $\hat{\mathbf{R}}_{\mathbf{X},i}$ in voiced parts. A similar approach was used in [2] where a hard threshold is used to distinguish between voiced and unvoiced speech sections. Here, we propose a soft threshold approach in which the constrained estimator of $\mathbf{R}_{\mathbf{X},i}$, denoted $\hat{\mathbf{R}}^{\delta_i}_{\mathbf{X},i}$, is computed as:

$$\hat{\mathbf{R}}^{\delta_i}_{\mathbf{X},i} = \delta_i\hat{\mathbf{R}}_{\mathbf{X},i} + (1-\delta_i)\operatorname{diag}\{\hat{\mathbf{R}}_{\mathbf{X},i}\}. \tag{23}$$

where $0 \leq \delta_i \leq 1$ is a soft threshold parameter accounting for voiced or unvoiced frames. We use the zero-crossing rates (ZCR) in the noisy speech time domain signal to distinguish between voiced and unvoiced parts since voiced parts are primarily low frequencies and unvoiced parts are primarily high frequencies [4]. A ZCR voiced threshold t_v is used, below which the frame is judged to be voiced and δ_i is set to 1. A ZCR unvoiced threshold $t_u > t_v$ is also used, above which the frame is judged to be unvoiced and δ_i is set to 0. For ZCR between t_u and t_v, intermediate values of δ_i are used. Specifically, the value of δ_i is evaluated as follows:

$$\delta_i = \begin{cases} 1 & \text{ZCR} \leq t_v \\ \frac{t_u - \text{ZCR}}{t_u - t_v} & t_v < \text{ZCR} < t_u \\ 0 & \text{ZCR} \geq t_u. \end{cases} \tag{24}$$

The clean speech estimators using $\hat{\mathbf{R}}^{\delta_i}_{\mathbf{X},i}$ (23) to estimate $\mathbf{R}_{\mathbf{X},i}$ will be denoted by the additional subscript δ, i.e. $\hat{\mathbf{X}}_{\delta\text{MMSE}}$, $\hat{\mathbf{X}}^o_{\delta L}$, $\hat{\mathbf{X}}^o_{\delta U}$ and $\hat{\mathbf{X}}^o_{\delta\gamma}$, otherwise, the estimator will use $\hat{\mathbf{R}}_{\mathbf{X},i}$. We refer to $\hat{\mathbf{R}}^{\delta_i}_{\mathbf{X},i}$ as the soft threshold structured estimator as opposed to the unstructured $\hat{\mathbf{R}}_{\mathbf{X},i}$.

4. EXPERIMENTAL RESULTS

In this section, we evaluate the proposed estimators. The value of $\gamma = 0.5$ will be considered in the $\hat{\mathbf{X}}^o_\gamma$ and $\hat{\mathbf{X}}^o_{\delta\gamma}$ estimators. Two types of noises from the Noisex database [9] are used in the experiments: a white noise and a colored (i.e. pink) noise. Other noise types were considered during the experimentation and lead to the same conclusions as the ones drawn below. Thirty noisy sentences were used in the evaluations and were all sampled at 16 kHz. A raised-cosine window [10] was used ($N = 512$ samples, 32ms) in the STSA computation and a 75% overlap was used in the overlap-add synthesis method as in [1].

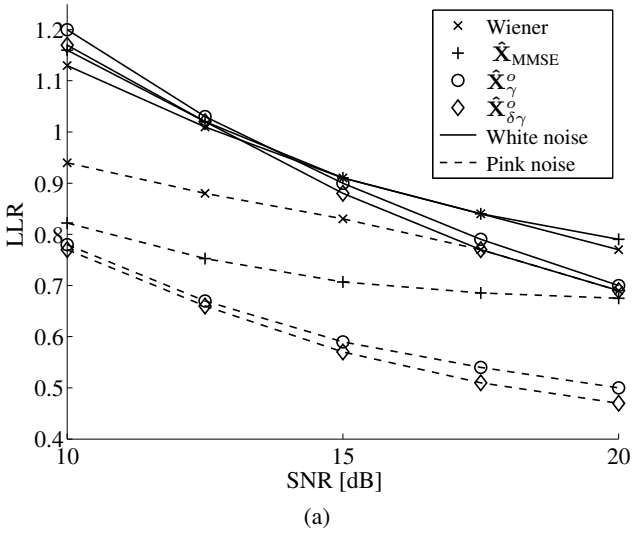
We identified through experimentation the following ZCR thresholds to be used in (24): $t_v = 3500$ crossings/sec and $t_u = 6000$ crossings/sec. Since ZCR are affected when the SNR is very low, $\hat{\mathbf{R}}^{\delta}_{\mathbf{X},i}$ (23) was only used if the power of the current frame was 1.5 times the estimated power of the noise, otherwise we used $\hat{\mathbf{R}}_{\mathbf{X},i}$.

Table 1 presents wideband PESQ [11] results. The wideband PESQ attempts to predict MOS scores and yields a result from 1 to 4.5, the higher score being the best result. As can be observed, the best results for all cases are always obtained by one of the proposed estimators. The algorithms that used the soft threshold structured estimator for the clean speech correlation matrix estimation $\hat{\mathbf{R}}^{\delta_i}_{\mathbf{X},i}$ (i.e. $\hat{\mathbf{X}}_{\delta\text{MMSE}}$, $\hat{\mathbf{X}}^o_{\delta L}$, $\hat{\mathbf{X}}^o_{\delta U}$ and $\hat{\mathbf{X}}^o_{\delta\gamma}$) gave better results than the ones using the unstructured $\hat{\mathbf{R}}_{\mathbf{X},i}$ (i.e. $\hat{\mathbf{X}}_{\text{MMSE}}$, $\hat{\mathbf{X}}^o_L$, $\hat{\mathbf{X}}^o_U$ and $\hat{\mathbf{X}}^o_\gamma$) for white noise while they were found more or less equivalent for colored noise. $\hat{\mathbf{X}}^o_{\delta\gamma}$ mostly gave better results than $\hat{\mathbf{X}}^o_{\delta L}$ and $\hat{\mathbf{X}}^o_{\delta U}$ while the advantage of $\hat{\mathbf{X}}^o_\gamma$ over $\hat{\mathbf{X}}^o_L$ and $\hat{\mathbf{X}}^o_U$ was case dependent.

Fig. 1 presents LLR results for selected representative estimators for white and pink noises. The LLR measure is evaluated as in [12]; a lower LLR score indicates a better performance. As can be observed, the comparison between the existing and the proposed algorithms were quite different between white and colored (i.e. pink) noises. In fact, for white noise with an SNR of 20 dB, the proposed

Table 1. Comparative wideband PESQ values for white and colored (pink) noises at several SNRs (10, 15 and 20 dB).

		MMSE STSA [1]	Wiener (11)	$\hat{\mathbf{X}}_{\text{MMSE}}$ (8)	$\hat{\mathbf{X}}_L^o$	$\hat{\mathbf{X}}_U^o$	$\hat{\mathbf{X}}_\gamma^o$	$\hat{\mathbf{X}}_{\delta\text{MMSE}}$	$\hat{\mathbf{X}}_{\delta L}^o$	$\hat{\mathbf{X}}_{\delta U}^o$	$\hat{\mathbf{X}}_{\delta\gamma}^o$
White	10 dB	1.35	1.53	1.57	1.57	1.46	1.52	**1.61**	**1.61**	1.52	1.59
	15 dB	1.70	1.90	1.94	1.98	1.93	1.98	1.98	2.01	2.04	**2.11**
	20 dB	2.25	2.45	2.39	2.44	2.53	2.52	2.48	2.51	2.62	**2.65**
Pink	10 dB	1.47	1.58	1.70	1.74	1.70	1.74	1.71	1.75	1.71	**1.77**
	15 dB	1.90	1.95	2.05	2.10	2.21	2.20	2.06	2.11	2.19	**2.23**
	20 dB	2.48	2.48	2.49	2.53	**2.72**	2.66	2.55	2.58	2.70	**2.72**



(a)

Fig. 1. LLR values versus SNR for white and pink noises.

estimators gave the best results while for the 0 dB case, the Wiener and $\hat{\mathbf{X}}_{\text{MMSE}}$ were slightly better than the proposed estimators. For the colored pink noise case, the proposed estimators were always better.

Only results for the value of $\gamma = 0.5$ in the $\hat{\mathbf{X}}_\gamma^o$ and $\hat{\mathbf{X}}_{\delta\gamma}^o$ estimators were reported in this paper. However, we also performed experiments with other values of γ. As expected, the results indicated that choosing values for γ closer to 0 yielded an enhanced speech closer to the one obtained with $\hat{\mathbf{X}}_L^o$ while choosing a value closer to 1 yielded an enhanced speech closer to $\hat{\mathbf{X}}_U^o$.

Informal listening experiments were also conducted to evaluate the qualitative merits of the proposed estimators. In particular, it was found that the processed speech in $\hat{\mathbf{X}}_{\text{MMSE}}$, $\hat{\mathbf{X}}_L^o$, $\hat{\mathbf{X}}_U^o$ and $\hat{\mathbf{X}}_\gamma^o$ sounded a little bit more muffled than the one obtained by Wiener or MMSE STSA. By allowing a better model for the unvoiced speech parts, the estimators $\hat{\mathbf{X}}_{\delta\text{MMSE}}$, $\hat{\mathbf{X}}_{\delta L}^o$, $\hat{\mathbf{X}}_{\delta U}^o$ and $\hat{\mathbf{X}}_{\delta\gamma}^o$ better preserve the fricatives and have less muffled speech. The best estimator overall was found to be the $\hat{\mathbf{X}}_{\delta\gamma}^o$ estimator. In fact, it has whiter background noise than Wiener's, less background noise than MMSE STSA and less speech distortions than the unconstrained full matrix equivalent $\hat{\mathbf{X}}_\gamma^o$.

5. CONCLUSION

In this paper we proposed a family of multidimensional Bayesian STSA estimators for speech enhancement that assume correlated frequency components. Results of wideband PESQ, LLR and informal listening experiments demonstrate noticeable advantages of the proposed estimators over existing ones. In particular, $\hat{\mathbf{X}}_{\delta\gamma}^o$ offers a good compromise between speech quality and background noise quantity and whiteness and is found to be the best overall estimator, especially at high SNR.

6. REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[2] C. Li and S. V. Andersen, "A block-based linear MMSE noise reduction with a high temporal resolution modeling of the speech excitation," *EURASIP J. Appl. Signal Process.*, vol. 18, pp. 2965–2978, 2005.

[3] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Prentice Hall, 2002.

[4] D. O'Shaughnessy, *Speech Communications: Human and Machine, 2nd Edition*, IEEE Press, 2000.

[5] E. Plourde and B. Champagne, "Bayesian spectral amplitude estimation for speech enhancement with correlated spectral components," in *Proc. 2009 IEEE Workshop on Statistical Signal Processing*, Cardiff, U.K., 2009.

[6] D. Sarason, *Complex Function Theory, 2nd Edition*, American Mathematical Society, 2007.

[7] W. Rudin, *Real and Complex Analysis, 3rd Edition*, McGraw-Hill, 1987.

[8] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Process, 4th Edition*, McGraw-Hill, 2002.

[9] Rice University, "Signal processing information base: Noise data," [Online] Available http://spib.rice.edu/spib/select_noise.html.

[10] P. Kabal, *Windows for Transform Processing*, Tech. Rep., McGill University, 2005, http://www-mmsp.ece.mcgill.ca/Documents/Reports/2005/KabalR2005a.pdf.

[11] ITU-T, *Recommendation P.862.2: Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*, Nov. 2005.

[12] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements, *Objective Measures of Speech Quality*, Prentice Hall, 1988.