Multidimensional STSA Estimators for Speech Enhancement With Correlated Spectral Components

Eric Plourde, *Member, IEEE*, and Benoît Champagne, *Senior Member, IEEE*

***Abstract*—Speech enhancement algorithms are used to remove background noise in a speech signal. 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. In this paper, we investigate a multidimensional Bayesian STSA estimator that assumes correlated spectral components. Since the closed-form solution of this optimum estimator is not readily available, we alternatively derive closed-form expressions for an upper and a lower bound on the desired estimator. Using these bounds, we propose a new family of speech enhancement estimators that are 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. An appropriate estimator for the correlation matrix of the clean speech is further derived. Evaluation results from both objective and subjective speech quality measures show that at moderate to high SNR values, where spectral correlation of speech is most noticeable, the proposed estimators can achieve significant improvements over the traditional STSA and Wiener filter estimators.**

***Index Terms*—Bayesian estimators, correlated spectral components, noise reduction, short-time spectral amplitude, speech enhancement.**

I. Introduction

SPEECH enhancement algorithms are used to remove background noise from speech signal acquired under imperfect conditions [1], [2]. They are present in many common devices such as cell phones and hearing aids. Among the various existing approaches for single channel speech enhancement, the ones based on frequency-domain processing are usually favored in applications for several reasons: low computational complexity via the use of Fast Fourier transform (FFT) algorithms, natural resemblance to the auditory processes taking place within the inner ear and especially the tonotopic mapping along the basilar membrane, and existence of efficient windowing techniques for the time-domain synthesis of the spectrally modified speech via overlap-add. Within the frequency-domain class, the Bayesian approach is particularly attractive due to its superior performance [3]. In this approach, 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.

The spectral amplitude has been found to be more perceptually relevant than the phase [4] in speech enhancement. Several Bayesian estimators of the short-time spectral amplitude (STSA), instead of the short-time Fourier transform (STFT) complex coefficients, have thus been proposed. These include the minimum mean-square error (MMSE) estimator of the STSA, known as MMSE STSA [5], as well as other related techniques such as e.g., [6]–[11]. Estimators of the STSA have shown some advantages over estimators of the STFT such as the well-known Wiener filter [3]. In fact, one desirable feature of Bayesian STSA estimators is to produce a residual background noise that is whiter than the residual musical noise produced by the Wiener estimator [12].

In Bayesian STSA estimation approaches, it is always assumed that the different spectral components of the noisy speech STFT are uncorrelated so that they can be processed independently. This assumption is however inexact as there are some well-known sources of correlation between the spectral components of speech signals in practice [13], [14]. Firstly, the fundamental frequency of voiced speech has harmonics that are inherently correlated. This correlation, which results from the periodic impulsive nature of the excitation source of voiced speech, is not resolved by short-time processing [13] and is most noticeable at higher SNR. Secondly, the finite temporal extension of the analysis window used in short-time processing introduces some correlation between adjacent frequencies [15].

Based on these considerations, a multidimensional MMSE estimator of the complex STFT coefficients that assumes correlated spectral components has been studied in [13], where the focus is on obtaining an accurate estimation of the nondiagonal clean speech correlation matrix which is required in the solution of the underlying MMSE estimation problem. The resulting estimator is shown to be advantageous, particularly at higher SNR values, over several existing estimators including a Wiener filter that assumes uncorrelated spectral components. Additional work that considers spectral correlation in speech enhancement using different frameworks can also be found in [16] and [17].

On the one hand, speech enhancement algorithms derived from Bayesian estimators of the clean speech STSA have been found to outperform algorithms based on complex STFT estimation. On the other hand, STFT estimators considering correlated spectral components yield better enhancement results than estimators not considering such correlation. Therefore, it appears that the consideration of correlated spectral components

Manuscript received June 22, 2010; revised October 22, 2010 and January 31, 2011; accepted March 21, 2011. Date of publication April 05, 2011; date of current version June 15, 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Maciej Niedzwiecki. This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

E. Plourde was with the Department of Electrical and Computer Engineering, McGill University, Montreal, QC H3A 2A7, Canada. He is now with the Neuroscience Statistics Research Laboratory, Harvard Medical School, Massachusetts General Hospital, Boston, MA 02114 USA, and also with the Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: eric.plourde@mail.mcgill.ca).

B. Champagne is with the Department of Electrical and Computer Engineering, McGill University, Montreal, QC H3A 2A7, Canada (e-mail: benoit.champagne@mcgill.ca).

Digital Object Identifier 10.1109/TSP.2011.2138697

in Bayesian STSA estimation might lead to even superior performance. However, this avenue has apparently not been previously considered in the speech and audio literature.

In this paper, we first investigate a multidimensional 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 alternatively develop closed-form expressions for a lower and an upper bound on the desired estimator. Based on those bounds, we propose a family of speech enhancement estimators being 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. We also show that the proposed bounds are tight at high SNR, confirming that the proposed scheme is optimal under this condition. Knowledge of the clean speech and noise correlation matrices is needed to implement the new 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.

We compare the proposed estimators with conventional Wiener and MMSE STSA, i.e., which both consider uncorrelated spectral components, as well as with an MMSE estimator of the complex STFT coefficients that assumes correlated spectral components. Both objective [wideband perceptual evaluation of speech quality (PESQ), log-likelihood ratio (LLR)] and subjective [multi-stimulus test with hidden reference and anchor (MUSHRA)] measures show that the proposed estimators achieve better performance than the benchmark estimators at moderate to high SNRs especially for colored noises. The proposed approach for STSA estimation, that considers correlated spectral components, opens new avenues for further developments since many improvements that have been made over the years for the uncorrelated spectral components case can be extended to the newly proposed formalism. Parts of this work have been previously reported in conferences [18], [19]. The present paper constitutes a substantial extension of these publications.

The paper is organized as follows. In Section II, we briefly review existing Bayesian estimators and discuss the limitations of the current modeling assumptions. Section III formulates the multidimensional STSA estimation problem, develops the above mentioned lower and upper bounds, and presents the proposed family of estimators. Section IV studies the proximity between the upper and lower bounds, and Section V addresses the estimation of the associated correlation matrices. Section VI presents the experimental results and discussion and Section VII concludes the work.

The following notation is used in this paper. 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.

II. Background

A. Traditional Approaches to Bayesian STSA Estimation

Let frame i of an observed noisy speech signal be

$$y_i[n] = x_i[n] + w_i[n], \quad 0 \leq n \leq N-1 \tag{1}$$

where n is the discrete-time index, $x_i[n]$ is the clean speech, $w_i[n]$ is the additive noise and N is the frame length. Let

$$Y_{k,i} \triangleq \sum_{n=0}^{N-1} y_i[n]h[n]e^{-j\frac{2\pi}{N}kn} \tag{2}$$

denote the k^{th} STFT coefficient of the noisy speech for the ith frame, where $h[n]$ is the analysis window and $k \in \{0, 1, \ldots, N-1\}$ is the frequency index. With $X_{k,i}$ and $W_{k,i}$ denoting the STFT of the clean speech and noise respectively, (1) thus becomes

$$Y_{k,i} = X_{k,i} + W_{k,i}. \tag{3}$$

To simplify the notation, we will often omit the subscript i.

In traditional Bayesian estimation for speech enhancement, the STFT coefficients are assumed to be uncorrelated and each frequency is processed independently. Let

$$X_k = \mathcal{X}_k e^{j\alpha_k} \tag{4}$$

where $\mathcal{X}_k > 0$ is the STSA and $\alpha_k \in [-\pi, \pi)$ is the associated phase. Since the spectral amplitude $\mathcal{X}_k$ has been found to be more perceptually relevant than the phase α_k [4], researchers have sought estimators of $\mathcal{X}_k$ instead of X_k. In the Bayesian STSA estimation approach, the goal is then to obtain the estimator[1] $\hat{\mathcal{X}}_k^{\mathrm{tr}}$, as a function of the noisy observation Y_k, which minimizes the expectation of a given cost function $C(\mathcal{X}_k, \hat{\mathcal{X}}_k)$:

$$\hat{\mathcal{X}}_k^{\mathrm{tr}} = \arg\min_{\hat{\mathcal{X}}_k} E\{C(\mathcal{X}_k, \hat{\mathcal{X}}_k)\} \tag{5}$$

where E denotes statistical expectation. This estimator is then combined with the phase of the noisy speech, $\measuredangle Y_k$, to yield the estimator of the complex spectrum of the clean speech:

$$\hat{X}_k^{\mathrm{tr}} = \hat{\mathcal{X}}_k^{\mathrm{tr}} e^{j\measuredangle Y_k}. \tag{6}$$

The corresponding time domain estimate of the clean speech, i.e., $\hat{x}[n]$, is obtained by performing an inverse Fourier transform of $\hat{X}_k^{\mathrm{tr}}$ for each frame, which are then combined using the overlap-add method [20].

Choosing the cost function as

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = (\mathcal{X}_k - \hat{\mathcal{X}}_k)^2 \tag{7}$$

along with a Gaussian statistical model for the clean speech and noise that assumes uncorrelated spectral components, yields the well-known MMSE STSA estimator which is proposed and studied in [5]. Similar estimators, but using other forms of cost functions, along with the Gaussian statistical model, have also been proposed. For example, a cost function considering the logarithms of the estimated and clean speech STSA is proposed in [21]. In [6], the MMSE STSA cost function is weighted by the STSA of the clean speech to obtain a more perceptually significant estimator while, in [7], power laws are applied to the estimated and actual clean speech STSA. Generalizations of these approaches are presented in [10] and [11].

[1] We use the superscript tr to distinguish the traditional estimator which does not assume correlated spectral components from the estimators proposed later.

B. Model Limitations

In recent years, the above modeling assumptions of uncorrelated Gaussian STFT coefficients have been challenged on different fronts, leading to new opportunities in speech signal processing.

A first set of issues revolve around the very use of a Gaussian model. Indeed, early studies demonstrate that speech signal samples in the time-domain are better modeled by Laplace or gamma distributions [22]. The use of a Gaussian statistical model in the above estimators is often motivated on the basis of the central limit theorem, since each Fourier expansion coefficient can be seen as a weighted sum of a large number of random variables resulting from the observed samples [5]. However, this assumption may not be fully justified in practice due to the relatively short integration windows used in the calculation of the STFT coefficients [23]. Experimental results from different sources [24]–[26] suggest that a better fit with the observed STFT data can be obtained with so-called super-Gaussian distributions, which include the Laplace and gamma distributions as special cases. Accordingly, many alternative distributions have been investigated for the real and imaginary parts of the STFT coefficients [24], [26], the STSA coefficients [25]–[28] and the complex STFT coefficients [29], [30], and for each of these, different estimators were developed that can lead to better enhancement performance under certain conditions.

Yet, a number of questions remain surrounding the use of super-Gaussian distributions. The process of fitting a parametric distribution to empirical speech STFT data is subject to interpretation as it is done under different SNR conditions. Also due to variability among studies in the experimental approaches and results, it is not yet firmly established which of these non-Gaussian distributions is susceptible to offer the best performance in a given enhancement task. For instance, some of the proposed non-Gaussian models, which assume independence of the real and imaginary STFT components, lead to noncircularly symmetric distribution in the polar domain (i.e., nonuniform phase), which seems to contradict experimental observations. More importantly from the perspective of this work, the super-Gaussian distributions are much less mathematically tractable, and therefore basic estimation approaches that are amenable to closed-form solutions under the Gaussian assumption often do not have an analytical counterpart when using other distributions [2].

A second set of issues, and which constitute the main motivation for this work, relate to the assumption of uncorrelated frequency components. In practice, and in contrast with the traditional assumptions used in the development of the estimators presented in Section II-A, there is evidence of correlation between the STFT coefficients of a speech signal corresponding to different frequencies [31]. This correlation is due to different factors, including [14]:

Use of window in frame-based processing: Indeed, the use of a finite window function $h[n]$ in the computation of the STFT in (2) introduces some correlation between adjacent spectral components. This is due to the spectral smearing phenomenon which is a known effect of the windowing process [20].

Harmonic structure of voiced speech: Voiced speech is characterized by the vibration of the vocal chords at a fundamental frequency F0 and has several harmonics at multiples of F0 [32]. This harmonic structure, which results from the periodic impulsive nature of the excitation source of voiced speech, introduces inherent correlation between STFT components corresponding to different multiple of F0. This correlation is not resolved by short-time processing and is most noticeable at higher SNR [13]. Therefore, correlations between adjacent frequencies result mainly from the windowing process while correlations between nonadjacent frequencies are mainly due to voiced speech.

In this work, our main interest lies in exploiting the correlation that exists between STFT coefficients at different frequencies, and especially the correlation in amplitude between different STSAs, to develop multidimensional MMSE STSA estimators with improved performance. To make our task tractable, and considering the above mentioned difficulties with the super-Gaussian distributions, we shall focus our analysis on the well-established Gaussian model. Through mathematical analysis, this will enable us to derive new multidimensional MMSE STSA estimators that exhibit a superior noise enhancement performance when compared to the traditional MMSE STSA estimator (i.e., assuming uncorrelated Gaussian frequency components), and to obtain fundamental insight into the desired structural attributes of multidimensional STSA estimators.

III. Multidimensional STSA Estimators

In this section, we proceed to obtain a multidimensional clean speech STSA estimator that assumes correlated spectral components. Defining $\mathbf{Y} = [Y_0\, Y_1\, \cdots\, Y_{N-1}]^T$, it follows from (3) that

$$\mathbf{Y} = \mathbf{X} + \mathbf{W} \tag{8}$$

where $\mathbf{X} = [X_0\ X_1\ \cdots\ X_{N-1}]^T$ and $\mathbf{W} = [W_0\, W_1\, \cdots\, W_{N-1}]^T$ are respectively the clean speech vector and the noise vector of the corresponding STFT coefficients. As in (4), we let $X_k = \mathcal{X}_k e^{j\alpha_k}$ and we also define the STSA vector $\boldsymbol{\mathcal{X}} = [\mathcal{X}_0\ \mathcal{X}_1\ \cdots\ \mathcal{X}_{N-1}]^T$ and the phase vector $\boldsymbol{\alpha} = [\alpha_0\ \alpha_1\ \cdots\ \alpha_{N-1}]^T$. We assume that $\mathbf{X}$ and $\mathbf{W}$ are independent, zero-mean circularly symmetric Gaussian random vectors with probability density functions:

$$f_{\mathbf{X}}(\mathbf{X}) = \frac{1}{\pi^N \det(\mathbf{R_X})} e^{-\mathbf{X}^H \mathbf{R_X}^{-1} \mathbf{X}} \tag{9}$$

$$f_{\mathbf{W}}(\mathbf{W}) = \frac{1}{\pi^N \det(\mathbf{R_W})} e^{-\mathbf{W}^H \mathbf{R_W}^{-1} \mathbf{W}}. \tag{10}$$

In these expressions, $\mathbf{R_X} = E\{\mathbf{X}\mathbf{X}^H\}$ and $\mathbf{R_W} = E\{\mathbf{W}\mathbf{W}^H\}$ are the correlation matrices of the clean speech and of the noise respectively and the superscript H indicates the conjugate transpose. We assume that these matrices are positive definite ($\mathbf{R_X} > 0, \mathbf{R_W} > 0$), so that the inverse matrices are well-defined. In practice, each spectral component of random vectors $\mathbf{X}$ and $\mathbf{W}$ has nonzero energy, although the eigenvalue spread of $\mathbf{R_X}$ and $\mathbf{R_W}$ may be large (e.g., voice sound). Traditional Bayesian STSA estimation approaches (e.g., [5]) assume that $\mathbf{R_X}$ and $\mathbf{R_W}$ are diagonal matrices, i.e., the spectral components are uncorrelated. In this work, we do not enforce such diagonality constraint: Our model considers possible frequency correlations in the clean speech and noise.

We want to evaluate the MMSE estimator of $\boldsymbol{\mathcal{X}}$:

$$\hat{\boldsymbol{\mathcal{X}}}^{o} = \arg\min_{\hat{\boldsymbol{\mathcal{X}}}} E\{\|\boldsymbol{\mathcal{X}} - \hat{\boldsymbol{\mathcal{X}}}\|^2\} \tag{11}$$

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 (11), i.e., $C(\boldsymbol{\mathcal{X}}, \hat{\boldsymbol{\mathcal{X}}}) \triangleq \|\boldsymbol{\mathcal{X}} - \hat{\boldsymbol{\mathcal{X}}}\|^2$, considers all the STSA spectral components jointly. Using matrix calculus, we can show that (11) leads to

$$\hat{\boldsymbol{\mathcal{X}}}^{o} = E\{\boldsymbol{\mathcal{X}}|\mathbf{Y}\} \tag{12}$$

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 $\mathbf{X}$:

$$\hat{\mathbf{X}}^{o} = [\hat{\mathcal{X}}_0^{o} e^{j\measuredangle Y_0}, \cdots, \hat{\mathcal{X}}_{N-1}^{o} e^{j\measuredangle Y_{N-1}}]^T. \tag{13}$$

In contrast to the scalar case under Gaussian assumptions [6], [7], a closed-form expression for (12) is not readily available. However, since the $\hat{\mathcal{X}}_k^o$ are positive real quantities, the Gaussian assumptions allow us to approach the problem of finding a realizable approximation to (12) by first obtaining tractable upper and lower bounds, $\hat{\mathcal{X}}_{U,k}^o$ and $\hat{\mathcal{X}}_{L,k}^o$ respectively, such that $\hat{\mathcal{X}}_{L,k}^o < \hat{\mathcal{X}}_k^o < \hat{\mathcal{X}}_{U,k}^o$. Based on the obtained bounds, we will then propose a parameterized family of estimators.

A. Lower Bound

Using the triangle inequality for integration [33], we can show that:

$$|E\{X_k|\mathbf{Y}\}| \leq E\{\mathcal{X}_k|\mathbf{Y}\}. \tag{14}$$

As a lower bound on the desired estimator (12), we therefore propose $\hat{\mathcal{X}}_{L,k}^o = |E\{X_k|\mathbf{Y}\}|$ or equivalently (using the notation introduced in the last paragraph of Section I for the absolute value of a matrix):

$$\hat{\boldsymbol{\mathcal{X}}}_L^{o} = |E\{\mathbf{X}|\mathbf{Y}\}|. \tag{15}$$

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 [13]

$$E\{\mathbf{X}|\mathbf{Y}\} = \hat{\mathbf{X}}_{\text{STFT}} = \mathbf{G}_{\text{MMSE}}\mathbf{Y} \tag{16}$$

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

$$\mathbf{G}_{\text{MMSE}} \triangleq \mathbf{R_X}(\mathbf{R_X} + \mathbf{R_W})^{-1}. \tag{17}$$

For future reference, it is also convenient to express $\mathbf{G}_{\text{MMSE}}$ in the following form, which can be obtained by application of the matrix inversion lemma [34]

$$\mathbf{G}_{\text{MMSE}} \triangleq (\mathbf{R_X}^{-1} + \mathbf{R_W}^{-1})^{-1}\mathbf{R_W}^{-1}. \tag{18}$$

A lower bound on the desired estimator is therefore

$$\hat{\boldsymbol{\mathcal{X}}}_L^{o} = \left|\mathbf{G}_{\text{MMSE}}\mathbf{Y}\right|. \tag{19}$$

Note that in the special case of uncorrelated spectral components (i.e., the traditional framework), $\mathbf{R_X}$ and $\mathbf{R_W}$ in (17) are diagonal matrices. Then, combining (19) with the phase of the noisy speech yields

$$\hat{X}_{L,k} = \frac{S_{X,k}}{S_{X,k} + S_{W,k}} Y_k \tag{20}$$

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

B. Upper Bound

Using Jensen's inequality [35], we have for a real convex function φ

$$\varphi(E\{\mathcal{X}_k|\mathbf{Y}\}) \leq E\{\varphi(\mathcal{X}_k)|\mathbf{Y}\}. \tag{21}$$

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

$$E\{\mathcal{X}_k|\mathbf{Y}\} \leq \sqrt{E\{\mathcal{X}_k^2|\mathbf{Y}\}} \tag{22}$$

which is also a special case of Lyapunov's inequality [36]. As an upper bound on the desired estimator (12), we therefore propose $\hat{\mathcal{X}}_{U,k}^o = \sqrt{E\{\mathcal{X}_k^2|\mathbf{Y}\}}$ or equivalently

$$\hat{\boldsymbol{\mathcal{X}}}_U^{o} = E\{\boldsymbol{\mathcal{X}}^{[2]}|\mathbf{Y}\}^{[\frac{1}{2}]}. \tag{23}$$

We next derive a closed-form expression for $E\{\mathcal{X}_k^2|\mathbf{Y}\}$.

Using a Bayesian formalism, we have

$$E\{\mathcal{X}_k^2|\mathbf{Y}\} = \frac{\int\cdots\int |X_k|^2 f_{\mathbf{Y}}(\mathbf{Y}|\mathbf{X}) f_{\mathbf{X}}(\mathbf{X}) d\mathbf{X}}{\int\cdots\int f_{\mathbf{Y}}(\mathbf{Y}|\mathbf{X}) f_{\mathbf{X}}(\mathbf{X}) d\mathbf{X}}. \tag{24}$$

We observe from (8) that

$$f_{\mathbf{Y}}(\mathbf{Y}|\mathbf{X}) = f_{\mathbf{W}}(\mathbf{Y} - \mathbf{X}). \tag{25}$$

Using (9), (10), and (25) in (24), we obtain (26), shown at the bottom of the page.

$$E\{\mathcal{X}_k^2|\mathbf{Y}\} = \frac{\int\cdots\int |X_k|^2 e^{\{\mathbf{Y}^H\mathbf{R_W}^{-1}\mathbf{X} + \mathbf{X}^H\mathbf{R_W}^{-1}\mathbf{Y} - \mathbf{X}^H(\mathbf{R_W}^{-1} + \mathbf{R_X}^{-1})\mathbf{X}\}} d\mathbf{X}}{\int\cdots\int e^{\{\mathbf{Y}^H\mathbf{R_W}^{-1}\mathbf{X} + \mathbf{X}^H\mathbf{R_W}^{-1}\mathbf{Y} - \mathbf{X}^H(\mathbf{R_W}^{-1} + \mathbf{R_X}^{-1})\mathbf{X}\}} d\mathbf{X}}. \tag{26}$$

To evaluate (26), we need to transform the multiple integrals into products of single integrals. To do so, we make use of the following eigenvalue decomposition:

$$\mathbf{U}\mathbf{\Lambda}\mathbf{U}^H = {\mathbf{R}_\mathbf{W}}^{-1} + {\mathbf{R}_\mathbf{X}}^{-1} \tag{27}$$

where $\mathbf{U}$ is the unitary matrix of eigenvectors, i.e., $\mathbf{U}^H\mathbf{U} = \mathbf{I}_N$, and $\mathbf{\Lambda}$ is the diagonal matrix containing the corresponding eigenvalues. Furthermore, we perform the following change of variables: $\mathbf{V} = \mathbf{U}^H\mathbf{X}$. Since $\mathbf{U}$ is unitary, the associated Jacobian is equal to 1 and (26) thus becomes

$$E\{\mathcal{X}_k^2|\mathbf{Y}\} = \frac{\int\cdots\int |\mathbf{U}_k\mathbf{V}|^2 e^{\{\tilde{\mathbf{Y}}^H\mathbf{V}+\mathbf{V}^H\tilde{\mathbf{Y}}-\mathbf{V}^H\mathbf{\Lambda}\mathbf{V}\}}d\mathbf{V}}{\int\cdots\int e^{\{\tilde{\mathbf{Y}}^H\mathbf{V}+\mathbf{V}^H\tilde{\mathbf{Y}}-\mathbf{V}^H\mathbf{\Lambda}\mathbf{V}\}}d\mathbf{V}} \tag{28}$$

where we define $\mathbf{U}_k$ as the k^{th} row of $\mathbf{U}$ and

$$\tilde{\mathbf{Y}} \triangleq \mathbf{U}^H{\mathbf{R}_\mathbf{W}}^{-1}\mathbf{Y}. \tag{29}$$

The product $\mathbf{U}_k\mathbf{V}$ (a scalar) can be expressed as

$$\mathbf{U}_k\mathbf{V} = \sum_{r=0}^{N-1} U_{kr}V_r \tag{30}$$

where U_{kr} is the $(k,r)^{\text{th}}$ entry of matrix $\mathbf{U}$ and V_r is the r^{th} entry of vector $\mathbf{V}$. Using (30), we can now write (28) in a form comprising only scalars:

$$E\{\mathcal{X}_k^2|\mathbf{Y}\} = \frac{\sum\limits_{r=0}^{N-1}\sum\limits_{t=0}^{N-1} U_{kt}^*U_{kr}\int\cdots\int V_t^*V_r\prod\limits_{m=0}^{N-1} g(V_m)dV_m}{\prod\limits_{m=0}^{N-1}\int g(V_m)dV_m} \tag{31}$$

where we define the positive real scalar function $g(V_m) = e^{\tilde{Y}_m^*V_m + V_m^*\tilde{Y}_m - |V_m|^2\lambda_m}$ for compactness and λ_m is the mth diagonal element of matrix $\mathbf{\Lambda}$.

Using (6.631.1), (8.411.1) and (9.212.1) from [37], we can evaluate the integrals in (31) and obtain (see Appendix A):

$$E\{\mathcal{X}_k^2|\mathbf{Y}\} = \sum_{r=0}^{N-1}\sum_{t=0}^{N-1} U_{kt}^*U_{kr}\frac{\tilde{Y}_t^*\tilde{Y}_r}{\lambda_t\lambda_r} + \sum_{p=0}^{N-1}\frac{|U_{kp}|^2}{\lambda_p}. \tag{32}$$

This last equation can also be written as

$$E\{\mathcal{X}_k^2|\mathbf{Y}\} = \mathbf{U}_k\mathbf{\Lambda}^{-1}\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^H\mathbf{\Lambda}^{-1}\mathbf{U}_k^H + \mathbf{U}_k\mathbf{\Lambda}^{-1}\mathbf{U}_k^H. \tag{33}$$

which, in turn, using the notation introduced previously, can be expressed in a more compact form as

$$E\{\boldsymbol{\mathcal{X}}^{[2]}|\mathbf{Y}\} = \mathrm{diag}\{\mathbf{U}\mathbf{\Lambda}^{-1}\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^H\mathbf{\Lambda}^{-1}\mathbf{U}^H + \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^H\}. \tag{34}$$

Using (27) and (29) along with the fact that $\mathrm{diag}\{\mathbf{A}\mathbf{A}^H\} = |\mathbf{A}|^{[2]}$ for any $\mathbf{A} \in \mathbb{C}^{N\times 1}$, we have

$$\begin{aligned} E\{\boldsymbol{\mathcal{X}}^{[2]}|\mathbf{Y}\} = {} & |({\mathbf{R}_\mathbf{W}}^{-1} + {\mathbf{R}_\mathbf{X}}^{-1})^{-1}{\mathbf{R}_\mathbf{W}}^{-1}\mathbf{Y}|^{[2]} \\ & + \mathrm{diag}\{({\mathbf{R}_\mathbf{W}}^{-1} + {\mathbf{R}_\mathbf{X}}^{-1})^{-1}\}. \end{aligned} \tag{35}$$

In light of (18), we notice that the entries of the first term in (35) are equal to the squared magnitudes of the entries of $\hat{\mathbf{X}}_{\mathrm{STFT}}$ in (16) and that the second term is simply $\mathrm{diag}\{\mathbf{G}_{\mathrm{MMSE}}\mathbf{R}_\mathbf{W}\}$. Finally, using (23), the desired upper bound is obtained as the following simple expression:

$$\hat{\boldsymbol{\mathcal{X}}}_U^o = (|\mathbf{G}_{\mathrm{MMSE}}\mathbf{Y}|^{[2]} + \mathrm{diag}\{\mathbf{G}_{\mathrm{MMSE}}\mathbf{R}_\mathbf{W}\})^{[\frac{1}{2}]}. \tag{36}$$

Since the upper bound includes the lower bound (19) and an additional positive term, it will obviously be greater than the lower bound.

It is interesting to note that in the special case of uncorrelated spectral components (i.e., the traditional framework), the estimator in (23) is the so-called β-order MMSE STSA estimator [7] with its parameter β having a value of $\beta = 2$. In fact, by considering diagonal $\mathbf{R}_\mathbf{X}$ and $\mathbf{R}_\mathbf{W}$ in (36) and the fact that $M(-1,1;-z) = z+1$, where $M(a,b;z)$ is the confluent hypergeometric function, one obtains the expression for the β-order MMSE STSA estimator with $\beta = 2$ as given in [7].

C. Proposed Family of Estimators

The desired estimator $\hat{\mathcal{X}}_k^o$ satisfies $\hat{\mathcal{X}}_{L,k}^o \leq \hat{\mathcal{X}}_k^o \leq \hat{\mathcal{X}}_{U,k}^o$. Based on the expressions of the derived bounds $\hat{\mathcal{X}}_{L,k}^o$ in (19) and $\hat{\mathcal{X}}_{U,k}^o$ in (36) we therefore propose the following family of estimators:

$$\hat{\boldsymbol{\mathcal{X}}}_\gamma^o = (|\mathbf{G}_{\mathrm{MMSE}}\mathbf{Y}|^{[2]} + \gamma\mathrm{diag}\{\mathbf{G}_{\mathrm{MMSE}}\mathbf{R}_\mathbf{W}\})^{[\frac{1}{2}]} \tag{37}$$

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:

$$\hat{\boldsymbol{\mathcal{X}}}_\gamma^o = \begin{cases} \hat{\boldsymbol{\mathcal{X}}}_U^o, & \text{if } \gamma = 1 \\ \hat{\boldsymbol{\mathcal{X}}}_L^o, & \text{if } \gamma = 0. \end{cases} \tag{38}$$

In the special case of uncorrelated spectral components, in light of the comments made at the end of Section III-A and Section III-B, the proposed estimator (38) will thus be equivalent to the Wiener filter for a value of $\gamma = 0$ and to the β-Order MMSE STSA estimator with $\beta = 2$ for $\gamma = 1$. Moreover, the proposed estimator will then correspond for some intermediate γ values to the MMSE STSA estimator, which is equivalent to the β-Order MMSE STSA estimator with $\beta = 1$, and also to the Log-MMSE STSA estimator [21], which is equivalent to the β-Order MMSE STSA estimator with $\beta \rightarrow 0$ [10].

As in (13), the spectral amplitude estimators $\hat{\boldsymbol{\mathcal{X}}}_L^o$, $\hat{\boldsymbol{\mathcal{X}}}_U^o$ and $\hat{\boldsymbol{\mathcal{X}}}_\gamma^o$ are 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.

IV. Upper and Lower Bound Proximity Analysis

In this section, in order to provide further mathematical justification for the multidimensional STSA estimators proposed in Section III-C, we study the proximity between the lower bound $\left|\mathbf{G}_{\mathrm{MMSE}}\mathbf{Y}\right|$ in (19) and upper bound $(|\mathbf{G}_{\mathrm{MMSE}}\mathbf{Y}|^{[2]} + \mathrm{diag}\{\mathbf{G}_{\mathrm{MMSE}}\mathbf{R}_\mathbf{W}\})^{[\frac{1}{2}]}$ in (36). Since $\hat{\mathcal{X}}_{U,k}^o$ and $\hat{\mathcal{X}}_{L,k}^o$ are both positive and $\hat{\mathcal{X}}_{U,k}^o > \hat{\mathcal{X}}_{L,k}^o$, we consider the vector

$$\mathbf{\Delta} = \left({\hat{\boldsymbol{\mathcal{X}}}_U^o}^{[2]} - {\hat{\boldsymbol{\mathcal{X}}}_L^o}^{[2]}\right)./\mathrm{diag}\{\mathbf{R}_\mathbf{X}\} \tag{39}$$

as a proximity indicator where ./ denotes an element-wise division. Each element Δ_k of vector $\boldsymbol{\Delta}$ is therefore a difference of squared values normalized by $S_{X,k} = E\{\mathcal{X}_k^2\}$. From (17), (19), and (36), we have

$$\begin{aligned}\boldsymbol{\Delta} &= \mathrm{diag}\{\mathbf{G}_{\mathrm{MMSE}}\mathbf{R}_{\mathbf{W}}\}./\mathrm{diag}\{\mathbf{R}_{\mathbf{X}}\} &(40)\\ &= \mathrm{diag}\{\mathbf{R}_{\mathbf{X}}(\mathbf{R}_{\mathbf{X}}+\mathbf{R}_{\mathbf{W}})^{-1}\mathbf{R}_{\mathbf{W}}\}./\mathrm{diag}\{\mathbf{R}_{\mathbf{X}}\}. &(41)\end{aligned}$$

Therefore, the second term in (36) dictates how tight the bounds are. Interestingly, this term does not depend on $\mathbf{Y}$ (however, in practice, the estimation of $\mathbf{R}_{\mathbf{X}}$ does).

1) Uncorrelated Frequencies: To gain some insight into the behavior of the proximity vector $\boldsymbol{\Delta}$, let us first consider uncorrelated spectral components. In that case, the k^{th} entry of $\boldsymbol{\Delta}$ reduces to:

$$\Delta_k = \frac{S_{W,k}}{S_{X,k}+S_{W,k}} = \frac{1}{1+\mathrm{SNR}_k} \tag{42}$$

where $\mathrm{SNR}_k = \frac{S_{X,k}}{S_{W,k}}$. For a high SNR_k, we have $\Delta_k \rightarrow 0$, while for a low SNR_k, $\Delta_k \rightarrow 1$. Since the proximity indicator is normalized by $\mathrm{diag}\{\mathbf{R}_{\mathbf{X}}\}$, a value of Δ_k close to 1 will thus indicate a distance between the bounds close to $S_{X,k}$ while a value of 0 will indicate that the bounds are identical. Therefore, the bounds will be tighter as the SNR_k gets higher. However, since $S_{X,k}$ will be small compared to $S_{W,k}$ for a low SNR_k, the distance between the bounds will also be small in that case but with respect to the noise level.

2) Correlated Frequencies: We next consider the case of correlated spectral components. $\boldsymbol{\Delta}$ can be written in a form apparented to that of (42):

$$\boldsymbol{\Delta} = \mathrm{diag}\{\mathbf{R}_{\mathbf{X}}(\mathbf{I}_N + \mathbf{R}_{\mathbf{W}}{}^{-1}\mathbf{R}_{\mathbf{X}})^{-1}\}./\mathrm{diag}\{\mathbf{R}_{\mathbf{X}}\}. \tag{43}$$

Let $\lambda_W^{\min}$, $\lambda_W^{\max}$, $\lambda_X^{\min}$ and $\lambda_X^{\max}$ denote the minimum and maximum eigenvalues of $\mathbf{R}_{\mathbf{W}}$ and $\mathbf{R}_{\mathbf{X}}$, respectively. On the one hand, it can be verified that if $\frac{\lambda_X^{\min}}{\lambda_W^{\max}} \gg 1$ (i.e., high SNR), then $\boldsymbol{\Delta} \rightarrow \mathrm{diag}\{\mathbf{R}_{\mathbf{W}}\}./\mathrm{diag}\{\mathbf{R}_{\mathbf{X}}\}$ which is equivalent to $\Delta_k \rightarrow \frac{[\mathbf{R}_{\mathbf{W}}]_{kk}}{[\mathbf{R}_{\mathbf{X}}]_{kk}} < \frac{\lambda_W^{\max}}{\lambda_X^{\min}}$. For a large SNR, Δ_k will thus be small indicating a small difference between the bounds. In fact, the lower and upper bounds will get asymptotically tighter (and thus converge to the desired estimator) in the limit of large SNR. On the other hand, if $\frac{\lambda_X^{\max}}{\lambda_W^{\min}} \ll 1$ (low SNR), then $\Delta_k \rightarrow 1$, showing that the proximity remains bounded.

V. Estimating $\mathbf{R}_{\mathbf{X}}$ and $\mathbf{R}_{\mathbf{W}}$

To compute $\hat{\boldsymbol{\mathcal{X}}}_\gamma^o$ (37), and therefore also $\hat{\boldsymbol{\mathcal{X}}}_L^o$ (19) and $\hat{\boldsymbol{\mathcal{X}}}_U^o$ (36), one needs an estimation of matrices $\mathbf{R}_{\mathbf{X}}$ and $\mathbf{R}_{\mathbf{W}}$. We shall denote the estimates of $\mathbf{R}_{\mathbf{X}}$, $\mathbf{R}_{\mathbf{W}}$ and $\mathbf{R}_{\mathbf{Y}}$ 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.

In this work, we use a decision-directed type of approach to estimate $\mathbf{R}_{\mathbf{X}}$ similar to the one used in [5]. Since $\mathbf{R}_{\mathbf{X}} = E\{\mathbf{X}\mathbf{X}^H\}$ and $\mathbf{R}_{\mathbf{X}} = \mathbf{R}_{\mathbf{Y}} - \mathbf{R}_{\mathbf{W}}$ for uncorrelated $\mathbf{X}$ and $\mathbf{W}$, we have for frame i:

$$\hat{\mathbf{R}}_{\mathbf{X},i} = \alpha\hat{\mathbf{X}}_{i-1}^o\hat{\mathbf{X}}_{i-1}^{oH} + (1-\alpha)\rho(\hat{\mathbf{R}}_{\mathbf{Y},i} - \hat{\mathbf{R}}_{\mathbf{W},i}) \tag{44}$$

where $\hat{\mathbf{X}}_{i-1}^o$ is given by (13) for frame $i-1$, $0 \le \alpha \le 1$ is a forgetting factor and $\rho(\cdot)$ is a thresholding function of its matrix argument. We note that the diagonal entries of $\hat{\mathbf{R}}_{\mathbf{X},i}$ should be positive. For an $N \times N$ matrix $\mathbf{A}$, we therefore define the lm^{th} element of $\rho(\mathbf{A})$ as

$$[\rho(\mathbf{A})]_{lm} = \begin{cases}\max([\mathbf{A}]_{lm}, 0), & \text{if } l = m\\ [\mathbf{A}]_{lm}, & \text{else.}\end{cases} \tag{45}$$

The $\max(\cdot,\cdot)$ operator is therefore applied only on the main diagonal of matrix $\hat{\mathbf{R}}_{\mathbf{Y},i} - \hat{\mathbf{R}}_{\mathbf{W},i}$ in (44). This approach may result, in practice, in a nonnegative definite $\hat{\mathbf{R}}_{\mathbf{X},i}$. A more formal approach based on eigenvalue decomposition, where the eigenvalues are forced to be positive, was also experimented to enforce a nonnegative definite constraint. However, it was observed that this latter approach gives similar results to the proposed simplified approach (44)–(45), yet at a much higher computational cost. Equation (44) will be used in the experimental results of Section VI to estimate $\mathbf{R}_{\mathbf{X}}$.

In addition to (44), we also use a modified structure of the estimator $\hat{\mathbf{R}}_{\mathbf{X},i}$ in Section VI to take into account the nature of the current frame, i.e., voiced versus unvoiced. Indeed, since the correlation due to the harmonics of the fundamental frequency is only present in the voiced parts of speech, it is 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 adopted in [13] where a hard threshold was used to distinguish between voiced and unvoiced speech sections. Here, we propose a soft threshold approach in which the constrained estimator of $\mathbf{R}_{\mathbf{X}}$, denoted $\hat{\mathbf{R}}_{\mathbf{X},i}^\delta$, is computed as

$$\hat{\mathbf{R}}_{\mathbf{X},i}^\delta = \delta_i\hat{\mathbf{R}}_{\mathbf{X},i} + (1-\delta_i)\mathrm{diag}\{\hat{\mathbf{R}}_{\mathbf{X},i}\} \tag{46}$$

where $\hat{\mathbf{R}}_{\mathbf{X},i}$ is given by (44) and $0 \le \delta_i \le 1$ is a soft threshold parameter accounting for voiced or unvoiced frames. In this work, we employ the zero-crossing rate (ZCR) in the noisy time-domain speech samples $y_i[n]$ to distinguish between voiced and unvoiced parts of speech. This approach, which is justified on the basis that voiced parts have prominently low frequencies while unvoiced parts have a broader spectral content [32], is used mainly to validate the proposed multidimensional approach for STSA estimation. Clearly, more elaborate classifiers which may be less sensitive to SNR could be implemented instead, such as those in [38]–[40].

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 computed as follows:

$$\delta_i = \begin{cases}1, & \mathrm{ZCR} \le t_v\\ \frac{t_u - \mathrm{ZCR}}{t_u - t_v}, & t_v < \mathrm{ZCR} < t_u\\ 0, & \mathrm{ZCR} \ge t_u.\end{cases} \tag{47}$$

The clean speech estimators using $\hat{\mathbf{R}}_{\mathbf{X},i}^\delta$ (46) to estimate $\mathbf{R}_{\mathbf{X}}$ is denoted by the additional superscript δ, i.e., $\hat{\mathbf{X}}_\gamma^\delta$, otherwise, the estimators use $\hat{\mathbf{R}}_{\mathbf{X},i}$ (44) and are denoted as $\hat{\mathbf{X}}_{\mathrm{STFT}}$ and $\hat{\mathbf{X}}_\gamma^o$. We refer to $\hat{\mathbf{R}}_{\mathbf{X},i}^\delta$ as the soft threshold structured estimator as opposed to the unstructured $\hat{\mathbf{R}}_{\mathbf{X},i}$. We note that the estimators that do not consider correlated spectral components, such as the MMSE STSA or Wiener estimators, use a traditional decision directed approach which always considers diagonal $\hat{\mathbf{R}}_{\mathbf{X},i}$ therefore corresponding to the case $\delta_i = 0$ in (46).

To compute $\hat{\boldsymbol{\mathcal{X}}}^o_\gamma$(37), we also need to estimate $\mathbf{R}_\mathbf{W}$. To do so, we first obtain a time-domain correlation matrix, $\hat{\mathbf{R}}_{\mathbf{w},i} = \mathbf{M}^H_{\mathbf{w},i}\mathbf{M}_{\mathbf{w},i}$ where $\mathbf{M}_{\mathbf{w},i}$ is a matrix whose columns are shifted versions of the time-domain noise data vector of the i^{th} frame (see (8.20) of [41]). Using the $N \times N$ Fourier transform matrix $\mathbf{F}$, we then obtain

$$\hat{\mathbf{R}}_{\mathbf{W},i} = \mathbf{F}\hat{\mathbf{R}}_{\mathbf{w},i}\mathbf{F}^H. \tag{48}$$

$\mathbf{R}_{\mathbf{Y},i}$ is estimated similarly.

VI. Experimental Results

In this section, we evaluate the proposed estimators and compare them to the traditional (i.e., one-dimensional) MMSE STSA [5] and Wiener (20) estimators, denoted respectively by $\hat{X}^{\text{tr}}_{\text{STSA}}$ and $\hat{X}^{\text{tr}}_{\text{Wiener}}$, as well as to the multidimensional MMSE STFT estimator, $\hat{\mathbf{X}}_{\text{STFT}}$ (16), using both objective (wideband PESQ, LLR) and subjective (MUSHRA) speech quality measures.

A. Methodology

Four types of noises from the Noisex database [42] were used in the experiments: white, pink, f16 and buccaneer-2 noises. We chose not to use babble noise and other highly nonstationary noises since the performance of the speech enhancement algorithms would then highly rely on the estimation of the time-varying noise statistics, which is not the main topic of this paper. Noisy speech signals were created according to ITU-T standard P.56 [43]. Since spectral correlations are mostly noticeable at moderate to high SNRs, we considered SNRs ranging from 5 to 20 dB. Thirty sentences (15 from three different female speakers and 15 from three different male speakers) were used in the evaluations. Zeros were padded at the beginning and end of each sentence to simulate silence periods of 0.75 s; the total sentences lengths were approximately 4 s. All speech signals were sampled at 16 kHz and a raised-cosine window [44] was employed ($N = 512$ samples, 32 ms) in the STSA computation. A 75% overlap was used in the overlap-add synthesis method as in [5]; however, experiments performed with 50% overlap (results not shown) led to the same conclusions as the ones presented below.

Following informal listening experiments, we found that a value of $\gamma = 0.5$ offered the best compromise between the amount of residual background noise, which increases as $\gamma \longrightarrow 1$, and the degree to which this background noise was musical,[2] which increases as $\gamma \longrightarrow 0$. The estimator $\hat{\mathbf{X}}^\delta_\gamma$ used the structured (soft threshold) clean speech correlation matrix estimation, $\hat{\mathbf{R}}^\delta_{\mathbf{X},i}$ (46), while $\hat{\mathbf{X}}_{\text{STFT}}$ and $\hat{\mathbf{X}}^o_\gamma$ used the unstructured estimator $\hat{\mathbf{R}}_{\mathbf{X},i}$ (44). The $\hat{X}^{\text{tr}}_{\text{STSA}}$ and $\hat{X}^{\text{tr}}_{\text{Wiener}}$ estimators both used the standard decision-directed approach of [5]. For the sake of simplicity, $\mathbf{R}_\mathbf{W}$ was estimated from the first five frames of the noisy speech signal, which did not contain any speech, and its value was kept constant for all subsequent frames. This approach is adequate here since the noises used are stationary and thus their statistics do not change over time. Other approaches, e.g., the minimum statistics approach [45], could be used for more complex noises.

We identified through experimentation the following ZCR thresholds to be used in (47): $t_v = 3500$ crossings/s and $t_u = 6000$ crossings/s. These values of t_v and t_u provide effective voiced/unvoiced signal classification over the SNR range under consideration in this study, except possibly for the lower SNR values where the ZCR is affected by noise. To avoid misclassifying voiced and unvoiced frames, $\hat{\mathbf{R}}^\delta_{\mathbf{X},i}$ (46) was only used if the power of the current frame was at least 1.5 times the estimated power of the noise, otherwise we used $\hat{\mathbf{R}}_{\mathbf{X},i}$ (44) which does not take into account voiced and unvoiced frames. We also set the forgetting factor in (44) to $\alpha = 0.98$.

B. Objective Measures

Many objective measures can be used to assess speech enhancement algorithms. They are more or less correlated with subjective measures such as the mean opinion score (MOS). A study of the correlation between MOS and objective measures was presented in [46]. Two of the objective measures that were found to have the best correlation with MOS were the PESQ and LLR measures with correlation coefficients of 0.89 and 0.85 respectively.[3] While PESQ was found to be well correlated with both signal and background distortions, the LLR was much more correlated with the signal distortion than with the background distortion. The widely used segmental SNR measure was found to have only a correlation of 0.36. We will use the PESQ and LLR measures in the following.

The PESQ measure was not originally intended to assess speech enhancement algorithms [47], however, it has been used in the past years in several speech enhancement studies, see e.g., [7], [48], [49]. PESQ generally attempts to predict MOS scores and yields a result from 1 to 4.5, the higher score being the best result. Here we will use the PESQ extension for 16 kHz sampled signals, termed wideband PESQ [50]. The clean speech files will be used as references in the PESQ evaluations.

The LLR measure for a particular frame i is defined as [51]–[53]

$$d_{\text{LLR}i} = \log\left(\frac{\hat{\mathbf{a}}_i^T \mathbf{R}_{\mathbf{X},i}\hat{\mathbf{a}}_i}{\mathbf{a}_i^T \mathbf{R}_{\mathbf{X},i}\mathbf{a}_i}\right) \tag{49}$$

where $\mathbf{a}_i$ is the linear predictive coding (LPC) vector of the original clean speech signal frame, $\hat{\mathbf{a}}_i$ is the LPC vector of the enhanced speech frame, and $\mathbf{R}_{\mathbf{X},i}$ is the autocorrelation matrix of the original clean speech signal. The mean d_{LLR} for all frames is evaluated from the different $d_{\text{LLR}i}$. To remove unrealistically high distortion levels, the frames with $d_{\text{LLR}i}$ greater than five times the standard deviation of all $d_{\text{LLR}i}$ are ignored in the averaging process [54] (typically less than 1% of the frames). A low d_{LLR} value can be interpreted as a close agreement between the spectral magnitudes of the original and enhanced speech signals. In particular, $d_{\text{LLR}} = 0$ indicates that the spectral magnitudes are identical while a large d_{LLR} implies that they are significantly different [55]. A lower LLR score thus indicates a better performance.

[2] By musical noise is meant sinusoidal components with random frequencies that come and go in each short-time frame [12].

[3] It is to be noted that the correlation results in [46] were obtained for a sampling rate of 8 kHz while we use a sampling rate of 16 kHz in the following experiments.

TABLE I
WIDEBAND PESQ VALUES FOR WHITE, PINK, F16 AND BUCCANEER-2 NOISES AT SEVERAL SNRS (5, 10, 15, AND 20 dB)

		Noisy	$\hat{X}^{tr}_{\text{STSA}}$	$\hat{X}^{tr}_{\text{Wiener}}$	$\hat{\mathbf{X}}_{\text{STFT}}$	$\hat{\mathbf{X}}^{o}_{\gamma=0}$	$\hat{\mathbf{X}}^{o}_{\gamma=0.5}$	$\hat{\mathbf{X}}^{o}_{\gamma=1}$	$\hat{\mathbf{X}}^{\delta}_{\gamma=0.5}$
		speech	[5]	(20)	(16)	$(\hat{\mathbf{X}}^{o}_{L})$		$(\hat{\mathbf{X}}^{o}_{U})$	
White	5 dB	1.04	1.14	1.29	**1.30**	**1.30**	1.24	1.21	1.24
	10 dB	1.08	1.35	1.53	1.57	1.57	1.52	1.46	**1.59**
	15 dB	1.20	1.70	1.90	1.94	1.98	1.98	1.93	**2.11**
	20 dB	1.47	2.25	2.45	2.39	2.44	2.52	2.53	**2.65**
Pink	5 dB	1.05	1.22	1.35	1.39	**1.40**	1.37	1.34	1.37
	10 dB	1.13	1.47	1.58	1.70	1.74	1.74	1.70	**1.77**
	15 dB	1.32	1.90	1.95	2.05	2.10	2.20	2.21	**2.23**
	20 dB	1.72	2.48	2.48	2.49	2.53	2.66	**2.72**	**2.72**
f16	5 dB	1.07	1.28	1.38	1.52	**1.53**	1.49	1.45	1.47
	10 dB	1.15	1.55	1.60	1.81	**1.85**	1.84	1.81	1.82
	15 dB	1.36	1.98	1.99	2.17	2.21	2.28	**2.29**	2.25
	20 dB	1.81	2.53	2.52	2.57	2.60	2.70	**2.75**	2.71
Buccaneer-2	5 dB	1.04	1.14	1.26	**1.30**	1.29	1.25	1.23	1.27
	10 dB	1.08	1.36	1.50	1.57	1.57	1.53	1.49	**1.61**
	15 dB	1.22	1.72	1.89	1.95	1.98	1.99	1.96	**2.10**
	20 dB	1.54	2.27	2.46	2.43	2.47	2.54	2.55	**2.65**

Table I presents wideband PESQ results for $\hat{X}^{\text{tr}}_{\text{STSA}}$, $\hat{X}^{\text{tr}}_{\text{Wiener}}$, $\hat{\mathbf{X}}_{\text{STFT}}$ and the proposed estimators. As can be observed, the proposed estimators always give the best results at moderate to high SNRs (> 5 dB). The improvements over $\hat{X}^{\text{tr}}_{\text{STSA}}$ are quite impressive for that SNR range, with an average improvement of 0.3. The advantages over $\hat{\mathbf{X}}_{\text{STFT}}$ are also quite significant at 15 and 20 dB, with an excess PESQ averaging 0.2 for the eight cases considered here. Considering that both the proposed and $\hat{\mathbf{X}}_{\text{STFT}}$ have similar computational complexity, there is a net advantage in opting for one of the newly proposed methods. Furthermore, at low SNR, $\hat{\mathbf{X}}^{o}_{\gamma=0}$ $(\hat{\mathbf{X}}^{o}_{L})$ is always better than $\hat{\mathbf{X}}^{o}_{\gamma=1}$ $(\hat{\mathbf{X}}^{o}_{U})$ while the inverse is true at high SNR; $\hat{\mathbf{X}}^{o}_{\gamma=0.5}$ being a good compromise across the SNR range considered here. Moreover, the algorithm that used $\hat{\mathbf{R}}^{\delta}_{\mathbf{X},i}$, i.e., $\hat{\mathbf{X}}^{\delta}_{\gamma=0.5}$, mostly gave better results than the one using the unstructured $\hat{\mathbf{R}}_{\mathbf{X},i}$, i.e., $\hat{\mathbf{X}}^{o}_{\gamma=0.5}$.

Fig. 1 presents LLR results for representative estimators identified from Table I, i.e., $\hat{X}^{\text{tr}}_{\text{Wiener}}$, $\hat{\mathbf{X}}_{\text{STFT}}$, $\hat{\mathbf{X}}^{o}_{\gamma=0.5}$, and $\hat{\mathbf{X}}^{\delta}_{\gamma=0.5}$ for the white, pink, f16, and buccaneer-2 noises. As noted before, the LLR measure is mostly correlated with the speech distortion and not the background noise, therefore, Fig. 1 mainly compares the estimators in terms of the speech distortion they produce. For white and buccaneer-2 noises, the proposed estimators produce less speech distortion than $\hat{X}^{\text{tr}}_{\text{Wiener}}$ and $\hat{\mathbf{X}}_{\text{STFT}}$ at higher SNR values, while at lower SNR, the situation is reversed. For the pink and f16 noises, $\hat{\mathbf{X}}^{\delta}_{\gamma=0.5}$ has less speech distortions than both $\hat{X}^{\text{tr}}_{\text{Wiener}}$ and $\hat{\mathbf{X}}_{\text{STFT}}$ over the entire SNR range considered here.

C. Subjective Measure

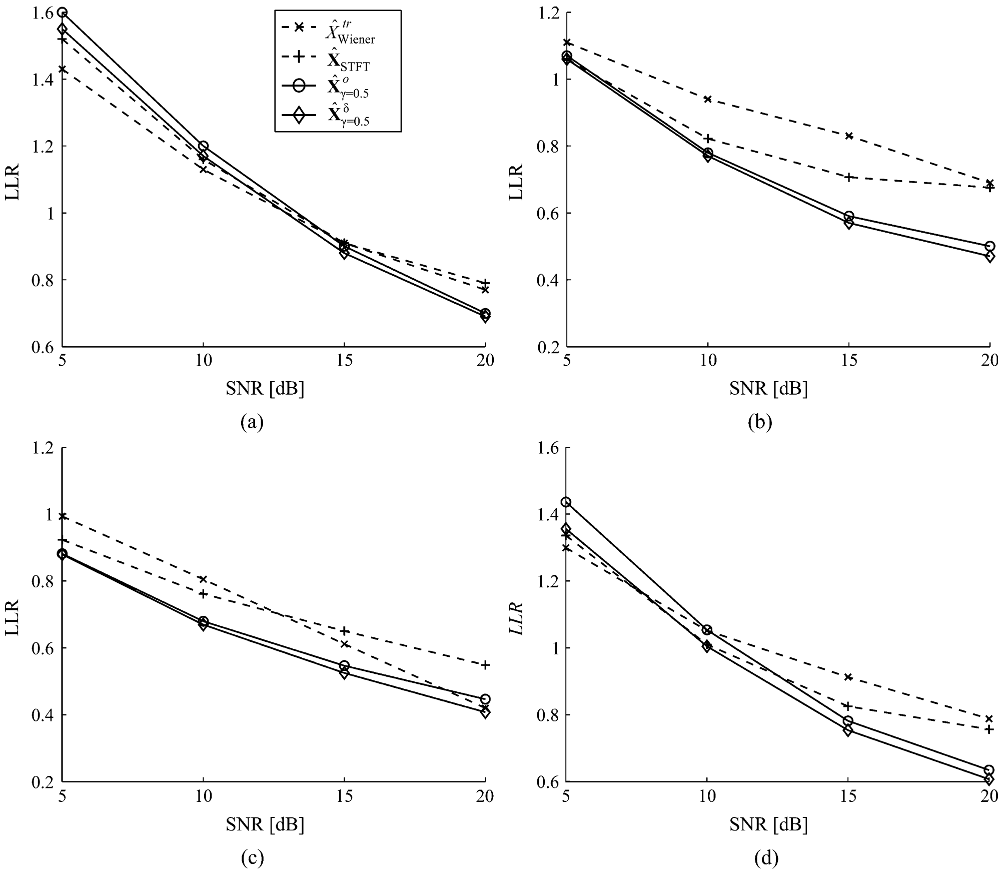
As a subjective measure, we used the MUSHRA (ITU-R Recommendation BS.1534–1) [56] test as implemented in [57]. In MUSHRA, the subjects are provided with the test utterances plus one reference and one hidden anchor and are asked to rate the different signals on a scale of 0 to 100, 100 being the best score. As the hidden anchor, which serves as a baseline for the evaluation of the enhanced sentences, we used a signal having an SNR of 5 dB less than the noisy signal to be enhanced. The clean speech signal was used as the reference. The listeners were allowed to listen to each sentence several times and always had access to the clean signal reference.

In order to limit the length of the listening test, we chose to use two noises from the four used in the previous objective evaluation: the white noise along with one of the colored noises, the f16 noise. The sentences were corrupted by the noises with an SNR of 15 dB. Furthermore, five estimators were included in the test: $\hat{X}^{\text{tr}}_{\text{STSA}}$, $\hat{X}^{\text{tr}}_{\text{Wiener}}$, $\hat{\mathbf{X}}^{o}_{\gamma=0}$, $\hat{\mathbf{X}}^{o}_{\gamma=0.5}$ and $\hat{\mathbf{X}}^{\delta}_{\gamma=0.5}$. The $\hat{\mathbf{X}}_{\text{STFT}}$ estimator was not included since, in fact, only the phase differs from the $\hat{\mathbf{X}}^{o}_{\gamma=0}$ estimator and, in practice, the two sound extremely similar. A total of 14 listeners (10 males and 4 females with a background in either speech processing or telecommunications) were involved in the test. Each participant had to rate, for each of the two noise types, two different utterances (one by a male speaker, one by a female speaker) for the 5 estimators, 1 reference and 1 anchor thus totaling 28 sentences per listener. The same sentences were used for all subjects. Tests were performed in an isolated acoustic room using beyerdynamic DT880 headphones.

Fig. 2 shows the results of the MUSHRA test, *Overall* indicating the mean of the two noises tested. As can be observed, the proposed estimators $\hat{\mathbf{X}}^{o}_{\gamma=0.5}$ and $\hat{\mathbf{X}}^{\delta}_{\gamma=0.5}$ were judged to have a similar performance but to be superior to $\hat{X}^{\text{tr}}_{\text{STSA}}$, $\hat{X}^{\text{tr}}_{\text{Wiener}}$ and $\hat{\mathbf{X}}^{o}_{\gamma=0}$. In particular, for the white noise, the advantage of the proposed estimators over the compared ones was found to be quite significant. For the colored f16 noise, $\hat{\mathbf{X}}^{o}_{\gamma=0.5}$ and $\hat{\mathbf{X}}^{\delta}_{\gamma=0.5}$ were found to be much better than $\hat{X}^{\text{tr}}_{\text{STSA}}$ and $\hat{X}^{\text{tr}}_{\text{Wiener}}$ and slightly superior to $\hat{\mathbf{X}}^{o}_{\gamma=0}$. Overall, the difference between each of $\hat{\mathbf{X}}^{o}_{\gamma=0}$, $\hat{\mathbf{X}}^{o}_{\gamma=0.5}$, $\hat{\mathbf{X}}^{\delta}_{\gamma=0.5}$ and both $\hat{X}^{\text{tr}}_{\text{STSA}}$ and $\hat{X}^{\text{tr}}_{\text{Wiener}}$ were all found to be statistically significant (pairwise t-test, $p < 0.05$).

The noisy speech enhanced by $\hat{\mathbf{X}}^{o}_{\gamma=0.5}$ sounded a little bit muffled. By allowing a better model for the unvoiced speech parts, the estimator $\hat{\mathbf{X}}^{\delta}_{\gamma=0.5}$ better preserved the fricatives than the $\hat{\mathbf{X}}^{o}_{\gamma=0.5}$ estimator and sounded clearer. Nevertheless, this aspect was not judged to have a strong impact by the participants since both $\hat{\mathbf{X}}^{o}_{\gamma=0.5}$ and $\hat{\mathbf{X}}^{\delta}_{\gamma=0.5}$ were rated similarly.

There was significantly more background noise in $\hat{X}^{\text{tr}}_{\text{STSA}}$ than in any of the other estimators tested which greatly explains its poor rating. Moreover, $\hat{X}^{\text{tr}}_{\text{Wiener}}$ had strong musical residual noise which most participant did not appreciate. The $\hat{\mathbf{X}}^{o}_{\gamma=0.5}$ and $\hat{\mathbf{X}}^{\delta}_{\gamma=0.5}$ were thus preferred since they had much less background noise than $\hat{X}^{\text{tr}}_{\text{STSA}}$ and much less musical noise than $\hat{X}^{\text{tr}}_{\text{Wiener}}$ or $\hat{\mathbf{X}}^{o}_{\gamma=0}$ which also exhibited significant musical noise.



(a) (b) (c) (d)

Fig. 1. LLR values versus SNR for (a) white, (b) pink, (c) f16, and (d) buccaneer-2 noises.

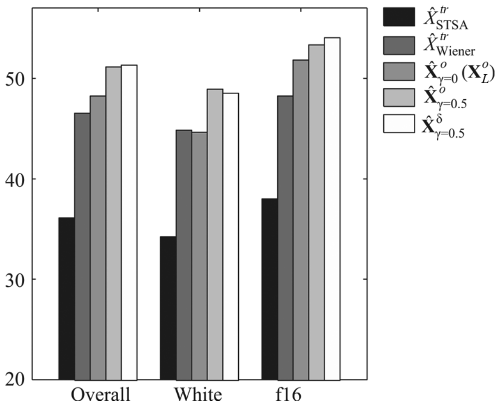


Fig. 2. Comparative subjective results (MUSHRA) for white and f16 noises as well as overall mean results (15 dB).

D. Discussion

Firstly, as mentioned previously, only the phase differs between $\hat{\mathbf{X}}_{\text{STFT}}$ and $\hat{\mathbf{X}}^{o}_{\gamma=0}$ $(\hat{\mathbf{X}}^{o}_{L})$. While $\hat{\mathbf{X}}_{\text{STFT}}$ has an optimal phase in the sense of the MMSE estimator of $\mathbf{X}$, $\hat{\mathbf{X}}^{o}_{\gamma=0}$ uses the phase of the noisy speech, which was found to be an optimal estimator of the clean speech phase for the uncorrelated spectral components case [5]. We computed for several utterances the differences between the phase of $\hat{\mathbf{X}}_{\text{STFT}}$ and the clean speech phase and also between the phase of $\hat{\mathbf{X}}^{o}_{\gamma=0}$ (which is the phase of the noisy speech) and the clean speech phase. We found that on average the phase of the noisy speech is closer to the phase of the clean speech than the phase of $\hat{\mathbf{X}}_{\text{STFT}}$. This was found to be true for the entire SNR range considered in this study.

Secondly, the proposed estimators showed much more advantage at higher SNR values than at lower SNRs as can be observed from the different objective results presented in this section. This could be due to the fact that at lower SNR values, the spectral correlations become masked by the noise and their inclusion in the estimation framework has less advantage.

Thirdly, while our discussion of the correlation between different spectral components of speech in Sections I and II was focused on the clean speech STFTs, some correlation may also exist between the noise STFTs. Of the four types of noises used in our experiments, white noise is not correlated, however the pink, f16 and buccaneer-2 noises exhibit some correlations, specifically: between adjacent frequencies for pink noise and between specific components for f16 and buccaneer-2 noises. The consideration of the speech correlations in the proposed algorithms does yield some improvements as observed for the white noise case. However, the consideration of noise correlations seems to further add some more improvements. In fact, results are found to be better when speech is corrupted by spectrally correlated noises, such as pink, f16 and buccaneer-2 noises, than when it is corrupted by white noise. This difference between the performance in white and colored noises could be due to the fact that the correlation present in the colored noises are considered in the noise correlation matrices of the proposed estimators but not in traditional approaches such as $\hat{X}^{\text{tr}}_{\text{STSA}}$ and $\hat{X}^{\text{tr}}_{\text{Wiener}}$.

Finally, only the case of fixed gamma values, i.e., $\gamma = 0$, $\gamma = 0.5$ and $\gamma = 1$, was considered in this work. One interesting avenue for future work would be to choose γ adaptively for each frame, e.g., based on SNR considerations, to obtain an estimator closer to the desired one as given by (12). It may also be possible to identify optimal γ values for different types of noises.

VII. Conclusion

In this paper we considered a multidimensional Bayesian STSA estimator for speech enhancement that assumes correlated spectral components. Since its closed-form solution is not readily available, we approached the problem of finding approximations to that estimator from a bounding perspective. We obtained convenient upper and lower bounds and proposed a new family of estimators based on these bounds. An analysis of the proximity between the bounds as a function of the speech signal's SNR was performed. Furthermore, an appropriate estimator for the correlation matrix of the clean speech, $\mathbf{R_X}$, was derived. Results of objective (wideband PESQ, LLR) and subjective (MUSHRA) measures demonstrate noticeable advantages of the proposed estimators over existing ones especially for colored noises and at moderate to high SNR values. In particular, $\hat{\mathbf{X}}^{o}_{\gamma=0.5}$ and $\hat{\mathbf{X}}^{\delta}_{\gamma=0.5}$ offer a good compromise between speech quality and background noise level and whiteness. Also of importance, the derivation of Bayesian STSA estimators considering correlation between spectral components can lead the way to many further developments such as those that have been proposed over the years for the traditional estimators, including extensions to various weighted cost functions and super-Gaussian distributions.

Appendix

In this appendix, we evaluate (31) and show that it yields (32). We start by solving for the numerator of (31), which we denote by N_k. The latter is written as a product and sum of single integrals in (50) at the bottom of the page. We need to evaluate four different integrals in (50). In order to integrate on real variables instead of complex ones, we will perform the following change of variables : $V_l = v_l e^{j\beta_l}$. The Jacobian associated with that change of variable is $J = v_l$. Let us evaluate the first integral in (50):

$$\begin{aligned}
&\int g(V_m)dV_m \\
&\quad= \int e^{\{\tilde{Y}_m^* V_m + V_m^* \tilde{Y}_m - |V_m|^2\lambda_m\}} dV_m \\
&\quad= \int_0^\infty v_m e^{-v_m^2\lambda_m} \int_{-\pi}^{\pi} e^{2\tilde{y}_m v_m \cos(\beta_m - \measuredangle \tilde{Y}_m)} d\beta_m dv_m \\
&\quad= 2\pi \int_0^\infty v_m e^{-v_m^2\lambda_m} J_0(-2j\tilde{y}_m v_m) dv_m \\
&\quad= \pi\lambda_m^{-1} M\left(1,1;\frac{\tilde{y}_m^2}{\lambda_m}\right)
\end{aligned} \tag{51}$$

where $\tilde{Y}_m = \tilde{y}_m e^{j\measuredangle\tilde{Y}_m}$, $J_n(\cdot)$ is a Bessel function of the first kind, $M(a,b;c)$ is the confluent hypergeometric function [37] and (6.631.1) of [37] was used in the last line of (51). The second integral can be evaluated similarly as

$$\begin{aligned}
&\int V_t^* g(V_t)dV_t \\
&\quad= \int V_t^* e^{\{\tilde{Y}_t^* V_t + V_t^* \tilde{Y}_t - |V_t|^2\lambda_t\}} dV_t \\
&\quad= \int_0^\infty v_t^2 e^{-v_t^2\lambda_t} \int_{-\pi}^{\pi} e^{-j\beta_t + 2\tilde{y}_t v_t \cos(\beta_t - \measuredangle \tilde{Y}_t)} d\beta_t dv_t \\
&\quad= 2\pi j e^{-j\measuredangle\tilde{Y}_t} \int_0^\infty v_t^2 e^{-v_t^2\lambda_t} J_1(-2j\tilde{y}_t v_t) dv_t \\
&\quad= \pi\tilde{Y}_t^* \lambda_t^{-2} M\left(2,2;\frac{\tilde{y}_t^2}{\lambda_t}\right).
\end{aligned} \tag{52}$$

The third integral is the complex conjugate of (52) and can be similarly shown to be

$$\int V_r g(V_r)dV_r = \pi\tilde{Y}_r \lambda_r^{-2} M\left(2,2;\frac{\tilde{y}_r^2}{\lambda_r}\right). \tag{53}$$

The first integral of the second summation in (50) is already given by (51) while the second integral of the second summation can be evaluated as

$$\begin{aligned}
&\int |V_p|^2 e^{\{\tilde{Y}_p^* V_p + V_p^* \tilde{Y}_p - |V_p|^2\lambda_p\}} dV_p \\
&\quad= \int_0^\infty v_p^3 e^{-v_p^2\lambda_p} \int_{-\pi}^{\pi} e^{2\tilde{y}_p v_p \cos(\beta_p - \measuredangle \tilde{Y}_p)} d\beta_p dv_p \\
&\quad= 2\pi \int_0^\infty v_p^3 e^{-v_p^2\lambda_p} J_0(-2j\tilde{y}_p v_p) dv_p \\
&\quad= \pi\lambda_p^{-2} M\left(2,1;\frac{\tilde{y}_p^2}{\lambda_p}\right).
\end{aligned} \tag{54}$$

We can therefore replace (51)–(54) in (50) to obtain (55), shown at the top of the following page. Using (9.212.1) from [37], i.e.,

$$M(\alpha,\gamma;z) = e^z M(\gamma-\alpha,\gamma;-z) \tag{56}$$

$$\begin{aligned}
N_k = &\sum_{\substack{r=0\\ r\neq t}}^{N-1} \sum_{t=0}^{N-1} U_{kt}^* U_{kr} \left(\prod_{\substack{m=0\\ m\neq r,t}}^{N-1} \int g(V_m)dV_m \right) \int V_t^* g(V_t)dV_t \int V_r g(V_r)dV_r \\
&+ \sum_{p=0}^{N-1} |U_{kp}|^2 \left(\prod_{\substack{m=0\\ m\neq p}}^{N-1} \int g(V_m)dV_m \right) \int |V_p|^2 g(V_p)dV_p.
\end{aligned} \tag{50}$$

$$\begin{aligned} N_k =& \pi^N \sum_{\substack{r=0\\r\neq t}}^{N-1}\sum_{t=0}^{N-1} U_{kt}^* U_{kr} \left(\prod_{\substack{m=0\\m\neq r,t}}^{N-1} \lambda_m^{-1} M\left(1,1;\frac{\tilde{y}_m^2}{\lambda_m}\right)\right) \frac{\tilde{Y}_t^* \tilde{Y}_r}{\lambda_t^2\lambda_r^2} M\left(2,2;\frac{\tilde{y}_t^2}{\lambda_t}\right) M\left(2,2;\frac{\tilde{y}_r^2}{\lambda_r}\right) \\ &+ \pi^N \sum_{p=0}^{N-1} |U_{kp}|^2 \left(\prod_{\substack{m=0\\m\neq p}}^{N-1}\lambda_m^{-1} M\left(1,1;\frac{\tilde{y}_m^2}{\lambda_m}\right)\right)\lambda_p^{-2} M\left(2,1;\frac{\tilde{y}_p^2}{\lambda_p}\right). \end{aligned} \tag{55}$$

$$N_k = \pi^N \left(\prod_{m=0}^{N-1}\lambda_m^{-1} M\left(1,1;\frac{\tilde{y}_m^2}{\lambda_m}\right)\right)\left(\sum_{\substack{r=0\\r\neq t}}^{N-1}\sum_{t=0}^{N-1} U_{kt}^* U_{kr}\frac{\tilde{Y}_t^* \tilde{Y}_r}{\lambda_t\lambda_r} + \sum_{p=0}^{N-1}\frac{|U_{kp}|^2}{\lambda_p} M\left(-1,1;-\frac{\tilde{y}_p^2}{\lambda_p}\right)\right). \tag{57}$$

$$N_k = \pi^N \left(\prod_{m=0}^{N-1}\lambda_m^{-1} M\left(1,1;\frac{\tilde{y}_m^2}{\lambda_m}\right)\right)\left(\sum_{r=0}^{N-1}\sum_{t=0}^{N-1} U_{kt}^* U_{kr}\frac{\tilde{Y}_t^* \tilde{Y}_r}{\lambda_t\lambda_r} + \sum_{p=0}^{N-1}\frac{|U_{kp}|^2}{\lambda_p}\right). \tag{58}$$

and the fact that $M(0,\gamma;z) = 1$, we get (57) at the top of the page. Now using the fact that $M\left(-1,1;-\frac{\tilde{y}_p^2}{\lambda_p}\right) = \frac{\tilde{y}_p^2}{\lambda_p} + 1$ in (57) yields (58), shown at the top of the page.

Next let us evaluate the denominator in (31). We notice that the integral in the latter is identical to (51), therefore

$$\prod_{m=0}^{N-1}\int g(V_m)dV_m = \pi^N \prod_{m=0}^{N-1}\lambda_m^{-1} M\left(1,1;\frac{\tilde{y}_m^2}{\lambda_m}\right). \tag{59}$$

Combining (58) and (59) in (31), we obtain the following:

$$E\{\mathcal{X}_k^2|\mathbf{Y}\} = \sum_{r=0}^{N-1}\sum_{t=0}^{N-1} U_{kt}^* U_{kr}\frac{\tilde{Y}_t^* \tilde{Y}_r}{\lambda_t\lambda_r} + \sum_{p=0}^{N-1}\frac{|U_{kp}|^2}{\lambda_p}$$

which is (32).

Acknowledgment

The authors would like to thank the participants in the listening tests for their precious help as well as the anonymous reviewers for their helpful comments.

References

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC Press, 2007.

[2] *Handbook of Speech Processing*, J. Benesty, M. Sondhi, and Y. E. Huang, Eds. New York: Springer, 2008.

[3] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun.*, vol. 49, pp. 588–601, 2007.

[4] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, no. 4, pp. 679–681, Aug. 1982.

[5] 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.

[6] P. C. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 857–869, Sep. 2005.

[7] C. H. You, S. N. Koh, and S. Rahardja, "β-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 475–486, Jul. 2005.

[8] J. H. L. Hansen, V. Radhakrishnan, and K. H. Arehart, "Speech enhancement based on generalized minimum mean square error estimators and masking properties of the auditory system," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 14, no. 6, pp. 2049–2063, Nov. 2006.

[9] C. Breithaupt, M. Krawczyk, and R. Martin, "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 4037–4040.

[10] E. Plourde and B. Champagne, "Auditory based spectral amplitude estimators for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 16, no. 8, pp. 1614–1623, Nov. 2008.

[11] E. Plourde and B. Champagne, "Generalized Bayesian estimators of the spectral amplitude for speech enhancement," *IEEE Signal Process. Lett.*, vol. 16, no. 6, pp. 485–488, Jun. 2009.

[12] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Process*, vol. 2, no. 2, pp. 345–349, Apr. 1994.

[13] 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.

[14] T. Fingscheidt, C. Beaugeant, and S. Suhadi, "Overcoming the statistical independence assumption w.r.t. frequency in speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2005, pp. 1081–1084.

[15] T. F. Quatieri and R. B. Dunn, "Speech enhancement based on auditory spectral change," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Orlando, FL, 2002, pp. 257–260.

[16] T. Esch, F. Heese, B. Geiser, and P. Vary, "Wideband noise suppression supported by artificial bandwidth extension techniques," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 4790–4793.

[17] M. Krini and G. Schmidt, *Speech and Audio Processing in Adverse Environments, Chapter Model-Based Speech Enhancement*. Berlin, Germany: Springer, 2008, pp. 89–134.

[18] E. Plourde and B. Champagne, "Bayesian spectral amplitude estimation for speech enhancement with correlated spectral components," in *Proc. IEEE Workshop Stat. Signal Process.*, Cardiff, U.K., 2009, pp. 397–400.

[19] E. Plourde and B. Champagne, "A family of Bayesian STSA estimators for the enhancement of speech with correlated frequency components," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Dallas, TX, 2010, pp. 4766–4769.

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

[21] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[22] W. B. Davenport, "An experimental study of speech wave probability distributions," *J. Acoust. Soc. Amer.*, vol. 24, pp. 390–399, 1952.

[23] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Process. Lett.*, vol. 10, pp. 204–207, 2003.

[24] R. Martin, "Speech enhancement based on minimum mean-square error estimation and super-Gaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 845–856, Sep. 2005.

[25] T. Lotter and P. Vary, "Speech enhancement by map spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Appl. Signal Process.*, pp. 1110–1126, 2005.

[26] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.

[27] T. H. Dat, K. Takeda, and F. Itakura, "Generalized gamma modeling of speech and its online estimation for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2005, pp. 181–184.

[28] I. Andrianakis and P. R. White, "Speech spectral amplitude estimators using optimally shaped gamma and chi priors," *Speech Commun.*, vol. 51, no. 1, pp. 1–14, Jan. 2009.

[29] S. Gazor and W. Zhang, "Speech enhancement employing Laplacian–Gaussian mixture," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 896–904, Sep. 2005.

[30] B. Chen and P. C. Loizou, "A laplacian-based MMSE estimator for speech enhancement," *Speech Commun.*, vol. 49, pp. 134–143, 2007.

[31] E. Plourde, "Bayesian short-time spectral amplitude estimators for single-channel speech enhancement," Ph.D. dissertation, McGill Univ., Montreal, QC, Canada, 2009.

[32] D. O'Shaughnessy, *Speech Communications: Human and Machine*, 2nd ed. New York: IEEE Press, 2000.

[33] D. Sarason, *Complex Function Theory*, 2nd ed. Providence, RI: Amer. Math. Soc., 2007.

[34] D. S. Bernstein, *Matrix Mathematics: Theory, Facts, and Formulas with Application to Linear Systems Theory*. Princeton, NJ: Princeton Univ. Press, 2005.

[35] W. Rudin, *Real and Complex Analysis*, 3rd ed. New York: McGraw-Hill, 1987.

[36] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Process*, 4th ed. New York: McGraw-Hill, 2002.

[37] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 6th ed. New York: Academic, 2000.

[38] B. Atal and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 3, pp. 201–212, 1976.

[39] Y. Qi and B. R. Hunt, "Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 2, pp. 250–255, 1993.

[40] E. Fisher, J. Tabrikian, and S. Dubnov, "Generalized likelihood ratio test for voiced-unvoiced decision in noisy speech using the harmonic model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 502–510, 2006.

[41] S. L. Marple, *Digital Spectral Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1987.

[42] Signal Processing Information Base: Noise Data, Rice Univ., Houston, TX [Online]. Available: http://spib.rice.edu/spib/select_noise.html

[43] *Objective Measurement of Active Speech Level*, ITU-T, Recommendation P.56, 1993.

[44] P. Kabal, "Windows for Transform Processing," McGill Univ., Montreal, QC, Canada, 2005.

[45] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[46] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[47] *Perceptual Evaluation of Speech Quality (PESQ), An Objective Method for End-to-End Speech Quality Assessment of NarrowBand Telephone Networks and Speech Codecs*, ITU-T, Recommendation P.862, 2001.

[48] N. Ma, M. Bouchard, and R. A. Goubran, "Speech enhancement using a masking threshold constrained Kalman filter and its heuristic implementations," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 19–32, Jan. 2006.

[49] V. Grancharov, J. Samuelsson, and B. Kleijn, "On causal algorithms for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 14, no. 3, pp. 764–773, May 2006.

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

[51] J. M. Tribolet, L. R. Rabiner, and M. M. Sondhi, "Statistical properties of an LPC distance measure," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 5, pp. 550–558, 1979.

[52] R. E. Crochiere, J. E. Tribolet, and L. R. Rabiner, "An interpretation of the log likelihood ratio as a measure of waveform coder performance," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 3, pp. 318–323, 1980.

[53] S. R. Quackenbush, T. P. Barnwell, III, and M. A. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[54] J. H. L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. 5th Int. Conf. Spoken Language Processing*, Sydney, Australia, 1998, pp. 2819–2822.

[55] R. E. Crochiere, "A weighted overlap-add method of short-time Fourier analysis/synthesis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 1, pp. 99–102, Feb. 1980.

[56] *Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems*, ITU-R, Recommendation BS.1534-1, 2001.

[57] E. Vincent, MUSHRAM: A MATLAB Interface for MUSHRA Listening Tests [Online]. Available: http://www.elec.qmul.ac.uk/people/emmanuelv/mushram/

Eric Plourde (S'05–M'08) received both the B.Ing. degree in electrical engineering and the M.Sc.A. degree in biomedical engineering from the Ecole Polytechnique de Montréal, QC, Canada, and the Ph.D. degree in electrical engineering from McGill University, Montreal, QC, Canada.

He is currently a Postdoctoral Fellow in the Neuroscience Statistics Research Laboratory with joint appointments at the Massachusetts General Hospital, Harvard Medical School and the Massachusetts Institute of Technology. His research interests include neural signal processing as well as speech processing with emphasis on speech enhancement and perceptually/biologically inspired processing.

Benoît Champagne (S'87–M'89–SM'03) was born in Joliette, Québec, Canada, in 1961. He received the B.Ing. degree in engineering physics from the Ecole Polytechnique de Montréal, QC, Canada, in 1983, the M.Sc. degree in physics from the Université de Montréal, QC, Canada, in 1985, and the Ph.D. degree in electrical engineering from the University of Toronto, ON, Canada, in 1990.

From 1990 to 1999, he was an Assistant and then Associate Professor at INRS-Telecommunications, Université du Québec, Montréal. In September 1999, he joined McGill University, Montreal, QC, Canada, as an Associate Professor within the Department of Electrical and Computer Engineering, where he served as Associate Chairman of Graduate Studies from January 2004 to August 2007. His research interests lie in the general area of statistical signal processing, including signal/parameter estimation, sensor array processing, and adaptive filtering and applications thereof to digital communications systems.

Dr. Champagne has served as Associate Editor of the IEEE SIGNAL PROCESSING LETTERS and the *EURASIP Journal on Applied Signal Processing* and, currently, the IEEE TRANSACTIONS ON SIGNAL PROCESSING.