BAYESIAN SPECTRAL AMPLITUDE ESTIMATION FOR SPEECH ENHANCEMENT WITH CORRELATED SPECTRAL 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 to be uncorrelated. However, this assumption is not exact since some correlation is present in practice. In this paper, we propose a STSA estimator with correlated frequency components. Since its closed-form solution is not readily available, we alternatively derive closed-form expressions for corresponding upper and lower bounds. Three new speech enhancement estimators are proposed based on those bounds: one for each bound and one that is a combination of both. Results of PESQ and informal listening experiments indicate that the proposed estimators give better performances than earlier estimators.

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

1. INTRODUCTION

Speech enhancement algorithms are used to remove background noise from a noisy speech [1]. They are present in many common devices such as cell phones and hearing aids. In the Bayesian approach for single channel speech enhancement, an estimate of the clean speech is derived by minimizing the statistical expectation of a cost function that penalizes errors in the clean speech estimate.

Several Bayesian estimators of the short-time spectral amplitude (STSA) of the clean speech have been proposed over the years [2, 3]. In those approaches, it is always assumed that the spectral components of the clean speech are uncorrelated. This assumption is however not exact as there are two main causes of correlations between STSA components [4]. Firstly, voiced speech has inherent harmonics that are necessarily correlated. Secondly, the finite dimension of the analysis window used in short-time processing introduces some correlation between adjacent frequencies.

A Bayesian estimator of the complex spectrum assuming correlated frequency components has been studied in [4]. While the correlation between STSA components has been noted in some papers, it has not yet been considered in Bayesian STSA estimators, apparently to simplify the estimators' derivations [2, 5].

In this paper, we first present an estimator that considers the STSA components to be correlated. Since a closed-form solution for such an estimator is not readily available, we alternatively find closed-form expressions for corresponding lower and upper bounds. We propose three new speech enhancement estimators: one for each bound and a last one that is the arithmetic mean of the two bounds. We compare the proposed estimators to the minimum mean-square error (MMSE) STSA estimator [2], the MMSE estimator of the complex spectrum and a Wiener estimator. Results show superior performance of the proposed estimators both in terms of the Perceptual Evaluation of Speech Quality (PESQ) and informal listening experiments.

The paper is organized as follows. Section 2 presents the desired STSA estimator with correlated frequency components while Section 3 presents the derivation of the bounds on that estimator. The bounds are characterized by correlation matrices for which an estimation approach is presented in Section 4. Section 5 presents some experimental results while Section 6 concludes the work.

2. SPECTRAL AMPLITUDE ESTIMATOR WITH CORRELATED FREQUENCY COMPONENTS

Let $\mathbf{Y}_i = \mathbf{X}_i + \mathbf{W}_i$ be an N-dimensional column vector representing the short-time Fourier coefficients of noisy speech observations for time frame i. $\mathbf{X}_i$ and $\mathbf{W}_i$ are respectively the clean speech vector and the noise vector of the corresponding short-time Fourier coefficients. 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 $\boldsymbol{\mathcal{X}} = [\mathcal{X}_1 \; \mathcal{X}_2 \; \cdots \; \mathcal{X}_N]^T$

This work was supported by a grant from the *Natural Sciences and Engineering Research Council of Canada*.

978-1-4244-2710-9/09/$25.00 © 2009 IEEE

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 Gaussian with pdfs:

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

$$f_{\mathbf{W}}(\mathbf{W}) = \quad \pi^{-N} \det(\mathbf{R_W})^{-1} e^{-\mathbf{W}^H \mathbf{R_W}^{-1} \mathbf{W}} \tag{2}$$

where $\mathbf{R_X} = E\{\mathbf{X}\mathbf{X}^H\}$ and $\mathbf{R_W} = E\{\mathbf{W}\mathbf{W}^H\}$ are the correlation matrices of the clean speech and noise respectively and $\mathbf{R_W} > 0$; superscript H indicates the conjugate transpose and E denotes statistical expectation. Traditional approaches (e.g. [2]) assume that $\mathbf{R_X}$ and $\mathbf{R_W}$ are diagonal (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 following estimator:

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

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

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

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

$$\hat{\mathbf{X}} = [\hat{\mathcal{X}}_1^o e^{j\measuredangle Y_1},\ \cdots, \hat{\mathcal{X}}_N^o e^{j\measuredangle Y_N}]^T. \tag{5}$$

The corresponding time domain estimator can be obtained by performing an inverse Fourier transform for each frame which are then combined using the overlap-add method.

Unfortunately and in contrast to the scalar case, a closed-form expression for (4) is not readily available. In the next section we approach the problem of finding a realizable solution to (4) by obtaining tractable upper and lower bounds instead of an elusive exact solution.

3. BOUNDS ON THE N-DIMENSIONAL CONDITIONAL EXPECTATION

The fact that the $\hat{\mathcal{X}}_k$ are positive real quantities makes it possible to approach the problem of finding approximations to (4) from a bounding perspective. Specifically, we derive below convenient upper and lower bounds, $\hat{\mathcal{X}}_{U,k}$ and $\hat{\mathcal{X}}_{L,k}$ respectively, such that $\hat{\mathcal{X}}_{L,k} < \hat{\mathcal{X}}_k < \hat{\mathcal{X}}_{U,k}$.

3.1. Lower bound

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

$$|E\{X_k|\mathbf{Y}\}| \leq E\{\mathcal{X}_k|\mathbf{Y}\}. \tag{6}$$

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

$$\hat{\boldsymbol{\mathcal{X}}}_L^o = |E\{\mathbf{X}|\mathbf{Y}\}| \tag{7}$$

where for any vector $\mathbf{A} = [a_k] \in \mathbb{C}^{N\text{x}1}$ we define $|\mathbf{A}| = [|a_k|] \in \mathbb{R}^{N\text{x}1}$. 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 [4]:

$$E\{\mathbf{X}|\mathbf{Y}\} = \hat{\mathbf{X}}_{\text{MMSE}} = \mathbf{R_X}(\mathbf{R_X} + \mathbf{R_W})^{-1}\mathbf{Y}. \tag{8}$$

A lower bound on the desired estimator is therefore:

$$\hat{\boldsymbol{\mathcal{X}}}_L^o = |\hat{\mathbf{X}}_{\text{MMSE}}| = \left|\mathbf{R_X}(\mathbf{R_X} + \mathbf{R_W})^{-1}\mathbf{Y}\right|. \tag{9}$$

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

$$\hat{X}_k = \frac{S_{X,k}}{S_{X,k} + S_{W,k}} Y_k \tag{10}$$

where $S_{X,k} = [\mathbf{R_X}]_{kk} = E\{|X_k|^2\}$ 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.

3.2. Upper bound

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

$$\varphi(E\{\mathcal{X}_k|\mathbf{Y}\}) \leq E\{\varphi(\mathcal{X}_k)|\mathbf{Y}\}. \tag{11}$$

If we set $\varphi(a) = a^2$, we have:

$$E\{\mathcal{X}_k|\mathbf{Y}\} \leq \sqrt{E\{\mathcal{X}_k^2|\mathbf{Y}\}} \tag{12}$$

since $\mathcal{X}_k > 0$. As an upper bound on the desired estimator (4), we therefore propose $\hat{\mathcal{X}}_{k,U}^o = \sqrt{E\{\mathcal{X}_k^2|\mathbf{Y}\}}$ or equivalently:

$$\hat{\boldsymbol{\mathcal{X}}}_U^o = \sqrt{E\{\boldsymbol{\mathcal{X}}^2|\mathbf{Y}\}} \tag{13}$$

where we define:

$$E\{\boldsymbol{\mathcal{X}}^2|\mathbf{Y}\} = [E\{\mathcal{X}_1^2|\mathbf{Y}\},\ \cdots\ , E\{\mathcal{X}_N^2|\mathbf{Y}\}]^T \tag{14}$$

and consider the square root as taken element-wise. We next derive a closed-form expression for $E\{\mathcal{X}_k^2|\mathbf{Y}\}$.

Using a Bayesian formalism we have:

$$E\{\mathcal{X}_k^2|\mathbf{Y}\} = \frac{\int\cdots\int |X_k|^2 f_{\mathbf{Y}}(\mathbf{Y}|\mathbf{X}) f_{\mathbf{X}}(\mathbf{X}) d\mathbf{X}}{\int\cdots\int f_{\mathbf{Y}}(\mathbf{Y}|\mathbf{X}) f_{\mathbf{X}}(\mathbf{X}) d\mathbf{X}}. \tag{15}$$

We observe that:

$$f_{\mathbf{Y}}(\mathbf{Y}|\mathbf{X}) = f_{\mathbf{W}}(\mathbf{Y} - \mathbf{X}). \tag{16}$$

Using (1), (2) and (16) in (15) we get (17) (bottom of this page).

To evaluate (17), we need to transform the multiple integrals into products of single integrals. To do so, we make use of the following eigenvalue decomposition:

$$\mathbf{U}\mathbf{\Lambda}\mathbf{U}^H = \mathbf{R_W}^{-1} + \mathbf{R_X}^{-1} \tag{18}$$

where $\mathbf{U}$ is the unitary matrix of eigenvectors, i.e. $\mathbf{U}^H\mathbf{U} = \mathbf{I}$, and $\mathbf{\Lambda}$ is the diagonal matrix containing the corresponding eigenvalues. Furthermore, we also perform the following change of variables: $\mathbf{V} = \mathbf{U}^H\mathbf{X}$. Since $\mathbf{U}$ is unitary, the associated Jacobian is $J = 1$ and (17) thus becomes:

$$E\{\mathcal{X}_k^2|\mathbf{Y}\} = \frac{\int\cdots\int |\mathbf{U}_k\mathbf{V}|^2 e^{\{\tilde{\mathbf{Y}}^H\mathbf{V}+\mathbf{V}^H\tilde{\mathbf{Y}}-\mathbf{V}^H\mathbf{\Lambda}\mathbf{V}\}}d\mathbf{V}}{\int\cdots\int e^{\{\tilde{\mathbf{Y}}^H\mathbf{V}+\mathbf{V}^H\tilde{\mathbf{Y}}-\mathbf{V}^H\mathbf{\Lambda}\mathbf{V}\}}d\mathbf{V}} \tag{19}$$

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

$$\tilde{\mathbf{Y}} \triangleq \mathbf{U}^H\mathbf{R_W}^{-1}\mathbf{Y}. \tag{20}$$

Since $\mathbf{U}_k\mathbf{V}$ is a scalar, we have:

$$\mathbf{U}_k\mathbf{V} = \sum_{r=1}^{N} U_{kr}V_r \tag{21}$$

where U_{kr} is the $(k,r)^{th}$ entry of matrix $\mathbf{U}$ and V_r is the r^{th} entry of vector $\mathbf{V}$. Using (21), we can now write (19) in a form comprising only scalars:

$$E\{\mathcal{X}_k^2|\mathbf{Y}\} \\ = \frac{\sum_{r=1}^{N}\sum_{t=1}^{N} U_{kt}^*U_{kr}\int\cdots\int V_t^*V_r\prod_{m=1}^{N}[h(V_m)dV_m]}{\prod_{m=1}^{N}\int h(V_m)dV_m} \tag{22}$$

where we define the positive real scalar function $h(V_m) = e^{\tilde{Y}_m^*V_m+V_m^*\tilde{Y}_m-|V_m|^2\lambda_m}$ 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 [8], we can evaluate the integrals in (22) and obtain:

$$E\{\mathcal{X}_k^2|\mathbf{Y}\} = \sum_{r=1}^{N}\sum_{t=1}^{N} U_{kt}^*U_{kr}\frac{\tilde{Y}_t^*\tilde{Y}_r}{\lambda_t\lambda_r} + \sum_{p=1}^{N}\frac{|U_{kp}|^2}{\lambda_p}. \tag{23}$$

This last equation can also be written as:

$$E\{\mathcal{X}_k^2|\mathbf{Y}\} = \mathbf{U}_k\mathbf{\Lambda}^{-1}\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^H\mathbf{\Lambda}^{-1}\mathbf{U}_k^H + \mathbf{U}_k\mathbf{\Lambda}^{-1}\mathbf{U}_k^H. \tag{24}$$

Defining $\mathrm{diag}\{\mathbf{A}\}$ as the column vector containing the diagonal elements of matrix $\mathbf{A}$, we can also write:

$$E\{\boldsymbol{\mathcal{X}}^2|\mathbf{Y}\} = \mathrm{diag}\{\mathbf{U}\mathbf{\Lambda}^{-1}\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^H\mathbf{\Lambda}^{-1}\mathbf{U}^H + \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^H\}. \tag{25}$$

Using (18) and (20) along with the fact that $\mathrm{diag}\{\mathbf{A}\mathbf{A}^H\} = |\mathbf{A}|^2$, where we consider the absolute value and squaring operators as taken element-wise, we have:

$$\begin{aligned} E\{\boldsymbol{\mathcal{X}}^2|\mathbf{Y}\} &= |(\mathbf{R_W}^{-1}+\mathbf{R_X}^{-1})^{-1}\mathbf{R_W}^{-1}\mathbf{Y}|^2 \\ &\quad + \mathrm{diag}\{(\mathbf{R_W}^{-1}+\mathbf{R_X}^{-1})^{-1}\}. \end{aligned} \tag{26}$$

We notice that the first term in (26) is equal to the squared magnitude value of $\hat{\mathbf{X}}_{\text{MMSE}}$ in (8) and the second term can be simplified in a form similar to $\hat{\mathbf{X}}_{\text{MMSE}}$. Finally, using (13), the desired upper bound is therefore the following simple expression:

$$\hat{\boldsymbol{\mathcal{X}}}_U^o = \sqrt{|\hat{\mathbf{X}}_{\text{MMSE}}|^2 + \mathrm{diag}\{\mathbf{R_X}(\mathbf{R_X}+\mathbf{R_W})^{-1}\mathbf{R_W}\}}. \tag{27}$$

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

The lower and upper bounds on the spectral amplitude estimator, i.e. $\hat{\boldsymbol{\mathcal{X}}}_L^o$ and $\hat{\boldsymbol{\mathcal{X}}}_U^o$, are then combined with the phase of the noisy speech, as in (5), resulting in estimators $\hat{\mathbf{X}}_L^o$ and $\hat{\mathbf{X}}_U^o$ respectively. Furthermore, we propose another estimator by simply considering the arithmetic mean of $\hat{\mathbf{X}}_L^o$ and $\hat{\mathbf{X}}_U^o$:

$$\hat{\mathbf{X}}_{LU}^o = (\hat{\mathbf{X}}_L^o + \hat{\mathbf{X}}_U^o)/2 \tag{28}$$

We will evaluate those three estimators in Section 5.

4. ESTIMATING $\mathbf{R_X}$ AND $\mathbf{R_W}$

To compute $\hat{\boldsymbol{\mathcal{X}}}_L^o$ (9) and $\hat{\boldsymbol{\mathcal{X}}}_U^o$ (27), one needs an estimation of matrices $\mathbf{R_X}$ and $\mathbf{R_W}$. In this work, we use a decision-directed type of approach similar to [2] to estimate $\mathbf{R_X}$. Since $\mathbf{R_X} = E\{\mathbf{X}\mathbf{X}^H\}$ and $\mathbf{R_X} = \mathbf{R_Y} - \mathbf{R_W}$ for uncorrelated $\mathbf{X}$ and $\mathbf{W}$, we have for frame i:

$$\hat{\mathbf{R}}_{\mathbf{X},i} = \alpha\hat{\mathbf{X}}_{i-1}\hat{\mathbf{X}}_{i-1}^H + (1-\alpha)g(\mathbf{R}_{\mathbf{Y},i} - \mathbf{R}_{\mathbf{W},i}) \tag{29}$$

where $\hat{\mathbf{X}}_{i-1}$ is given by (5) applied to frame $i-1$ and $\alpha = 0.98$. The terms on the diagonal of $\mathbf{R_X}$ should be positive, we therefore define $g(\mathbf{R}_{\mathbf{Y},i} - \mathbf{R}_{\mathbf{W},i})$ element-wise as:

$$g(\mathbf{R}_{\mathbf{Y},i} - \mathbf{R}_{\mathbf{W},i})|_{(l,m)} \\ = \begin{cases} \max[(\mathbf{R}_{\mathbf{Y},i} - \mathbf{R}_{\mathbf{W},i})|_{(l,m)}, 0] & \text{if } l = m \\ (\mathbf{R}_{\mathbf{Y},i} - \mathbf{R}_{\mathbf{W},i})|_{(l,m)} & \text{else} \end{cases} \tag{30}$$

$$E\{\mathcal{X}_k^2|\mathbf{Y}\} = \frac{\int\cdots\int |X_k|^2 e^{\{\mathbf{Y}^H\mathbf{R_W}^{-1}\mathbf{X}+\mathbf{X}^H\mathbf{R_W}^{-1}\mathbf{Y}-\mathbf{X}^H(\mathbf{R_W}^{-1}+\mathbf{R_X}^{-1})\mathbf{X}\}}d\mathbf{X}}{\int\cdots\int e^{\{\mathbf{Y}^H\mathbf{R_W}^{-1}\mathbf{X}+\mathbf{X}^H\mathbf{R_W}^{-1}\mathbf{Y}-\mathbf{X}^H(\mathbf{R_W}^{-1}+\mathbf{R_X}^{-1})\mathbf{X}\}}d\mathbf{X}} \tag{17}$$

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

		MMSE STSA [2]	Wiener (10)	$\hat{\mathbf{X}}_{\text{MMSE}}$ (8)	$\hat{\mathbf{X}}^o_L$	$\hat{\mathbf{X}}^o_U$	$\hat{\mathbf{X}}^o_{\overline{LU}}$
White	10 dB	2.47	2.60	2.63	**2.65**	2.60	2.64
	15 dB	2.82	2.88	2.95	**2.98**	2.97	**2.98**
	20 dB	3.14	3.19	3.31	**3.33**	3.31	**3.33**
Pink	10 dB	2.47	2.57	2.61	2.64	2.63	**2.65**
	15 dB	2.81	2.85	2.92	2.94	**2.98**	**2.98**
	20 dB	3.13	3.16	3.25	3.27	**3.30**	3.29
Cockpit	10 dB	2.25	2.20	2.25	2.29	2.31	**2.32**
	15 dB	2.60	2.54	2.60	2.64	**2.70**	2.69
	20 dB	2.96	2.91	2.97	3.00	**3.06**	3.05

where l and m are the matrix elements indices; the $\max$ operator is therefore applied only on the main diagonal of matrix $\mathbf{R}_{\mathbf{Y},i} - \mathbf{R}_{\mathbf{W},i}$.

We also need to estimate $\mathbf{R}_{\mathbf{W},i}$ (or equivalently $\mathbf{R}_{\mathbf{W}}$ if we omit the frame index i). We first estimate the time-domain correlation matrix, $\mathbf{R}_{\mathbf{w},i}$, using a $(N-1)^{th}$ order predictive error model [9]. Using the NxN Fourier transform matrix, $\mathbf{F}$, we then obtain:

$$\mathbf{R}_{\mathbf{W},i} = \mathbf{F}\mathbf{R}_{\mathbf{w},i}\mathbf{F}^H. \qquad (31)$$

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

5. EXPERIMENTAL RESULTS

In this section we report PESQ [10] results for the MMSE STSA [2], Wiener (10), $\hat{\mathbf{X}}_{\text{MMSE}}$ (8) and the proposed estimators $\hat{\mathbf{X}}^o_L$, $\hat{\mathbf{X}}^o_U$ and $\hat{\mathbf{X}}^o_{\overline{LU}}$. We also report informal listening observations.

Three types of noises from the Noisex database [11] are used in the experiments: a white noise, a pink noise and an aircraft cockpit noise (buccaneer-1). Thirty noisy speech signals [12] were created using ITU-T standard P.56 [13]. All speech signals were sampled at 16 kHz and a raised-cosine window was used (512 samples, 32ms) in the STSA computation. A 75% overlap was used in the overlap-add synthesis method as in [2]. For value of simplicity, $\mathbf{R}_{\mathbf{W}}$ was estimated from the first five frames of the noisy signal which did not contain any speech signal and its value was kept constant for all subsequent frames.

As can be observed from the PESQ values in Table 1, either $\hat{\mathbf{X}}^o_L$, $\hat{\mathbf{X}}^o_U$ or $\hat{\mathbf{X}}^o_{\overline{LU}}$ performed better than the other estimators for all cases, with the most significant improvements being observed for colored noises (i.e. pink and cockpit).

Informal listening experiments were also conducted. On the one hand, $\hat{\mathbf{X}}_{\text{MMSE}}$, $\hat{\mathbf{X}}^o_L$, $\hat{\mathbf{X}}^o_U$ and $\hat{\mathbf{X}}^o_{\overline{LU}}$ resulted in similar speech quality that seemed to be more natural than Wiener's. On the other hand, Wiener, $\hat{\mathbf{X}}_{\text{MMSE}}$ and $\hat{\mathbf{X}}^o_L$ had less background noise than $\hat{\mathbf{X}}^o_U$ but the noise sounded more musical. The algorithm based on the mean of the two bounds, i.e. $\hat{\mathbf{X}}^o_{\overline{LU}}$, offered a good compromise between speech quality, background noise quantity and whiteness.

6. CONCLUSION

In this paper we propose a Bayesian STSA estimator for speech enhancement that considers correlated frequency components. Since its closed-form solution is not readily available, we approach the problem of finding approximations to that estimator from a bounding perspective. We obtain convenient upper and lower bounds and propose three new estimators derived from those bounds. Results of PESQ and informal listening experiments indicate that the proposed estimators give better performances than earlier estimators. In particular, $\hat{\mathbf{X}}^o_{\overline{LU}}$ offers a good compromise between speech quality, background noise quantity and whiteness.

7. REFERENCES

[1] J. Benesty, M. Sondhi, and Y. Huang, Eds., *Springer Handbook of Speech Processing*, Springer, 2008.

[2] 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. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[3] 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.

[4] 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.

[5] I. Cohen, "Relaxed statistical model for speech enhancement and a priori SNR estimation," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 870–881, Sep. 2005.

[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] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products, 6th Edition*, Academic Press, 2000.

[9] S. Haykin, *Adaptive filter theory, 4th Edition*, Prentice Hall, 2002.

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

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

[12] "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. AU-17, no. 3, Sep. 1969.

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