Microphone Array Based Speech Spectral Amplitude Estimators with Phase Estimation

Mahdi Parchami*, Wei-Ping Zhu* and Benoit Champagne†

*Department of Electrical and Computer Engineering
Concordia University, Montreal, Quebec, Canada H3G 1M8
Emails: m_parch@ece.concordia.ca, weiping@ece.concordia.ca

†Department of Electrical and Computer Engineering
McGill University, Montreal, Quebec, Canada H3A 0E9
Email: benoit.champagne@mcgill.ca

***Abstract*—Bayesian estimators of short time spectral amplitude (STSA) have received considerable attention in the field of speech enhancement. In this paper, we propose new multi-microphone extensions for the conventional Ephraim and Malah's speech spectral amplitude estimation method. Unlike the conventional estimators where the spectral phase is assumed to be uniformly distributed, the proposed extensions treat the latter as an unknown parameter to be estimated. It is shown that the proposed methods can exploit spectral phase estimates to improve the performance of the current speech STSA estimators and have the potential to provide even further improvement given a more accurate estimate of the spectral phase. Experimental results indicate the superiority of the new approaches in terms of noise reduction and speech distortion measures, in addition to the reduced computational complexity provided by the proposed minimum mean square method as compared to state-of-the-art solutions.**

***Keywords*—Bayesian STSA estimators, noise reduction, short-time spectral amplitude, speech enhancement**

I. Introduction

Speech enhancement aims at the estimation of a clean speech signal from its noisy observations. In this application, frequency domain methods have become a dominant choice for practical systems, due to their effectiveness in providing better separation of the clean speech from the noise in addition to their efficient implementation via fast Fourier transform (FFT) [1]. Among the well-known speech spectrum estimation methods, the class of short-time spectral amplitude (STSA) estimators proposed by Ephraim and Malah in [2] and [3] are favored due to their superior performance compared to the other classical approaches like spectral subtraction and Wiener filtering. These authors proposed to estimate the speech STSA by minimization of cost functions representing the error between the clean speech and the estimated speech spectral amplitude with the spectral phase treated as a nuisance variable. Several major modifications of their groundbreaking work have been later suggested in the literature, either by defining new perceptually more relevant cost functions, e.g. [4] and references therein, or by taking into account heavy-tailed non-Gaussian prior distributions in modeling the clean speech STSA [5].

It is noted that the above mentioned STSA estimators are derived by assuming a uniformly distributed speech spectral phase and then treating the problem by taking statistical expectation with respect to this unknown phase. Although being optimal in the sense of minimum mean square error (MMSE), these methods lack the use of any prior information for the phase component, and therefore, neglect the potential to improve the speech STSA estimation performance by employing the spectral phase estimate. Moreover, they lead to the appearance of complex hypergeometric or modified Bessel functions in the structure of the STSA gain functions, which are computationally expensive and numerically unstable for large input arguments.

In this paper, we propose to treat the speech spectral phase as an unknown parameter to be estimated and obtain a new class of STSA estimators which exploits the phase component in its structure. We consider a more general scenario of speech STSA estimation with microphone array and obtain the unknown spectral phase component by conventional estimation methods. We investigate the new estimators with spectral phase in terms of objective quality measures as well as computational load, and demonstrate that they outperform the conventional STSA estimators where the speech spectral phase is treated as a uniformly distributed random variable.

II. Problem Statement

We consider a multiple microphone configuration for the proposed method, leaving the single microphone scheme as a special case. Without loss of generality, we assume that a set of N microphones are used to capture the noisy observation waveforms $y_n(t)$. The latter consists of the time delayed clean speech signals $x(t-\tau_n)$ contaminated by additive spatially uncorrelated noise samples $v_n(t)$, where n is the microphone index and τ_n is the relative time delay of the speech signal in the nth microphone with respect to the reference (first) microphone. Assuming the microphone array can accurately time align the delayed speech signal components, $x(t-\tau_n)$, to compensate for the time delays, we have equivalently that

$$y_n(t) = x(t) + v_n(t), \quad n = 1, 2, ..., N \tag{1}$$

where $x(t)$ is the coherent speech signal under estimation. After sampling and using short-time Fourier transform (STFT)

978-1-4799-3432-4/14/$31.00 ©2014 IEEE

analysis, the noisy speech signal can be represented in the frequency domain as

$$Y_n(k,l) = X(k,l) + V_n(k,l), \quad n = 1,2,...,N \tag{2}$$

with k denoting the frequency bin number and l the time frame index. The speech spectral component $X(k,l)$ can be written as $X(k,l) = A(k,l)e^{j\theta(k,l)}$ with $A(k,l) \geq 0$ being the spectral amplitude and $\theta(k,l) \in [-\pi, \pi]$ the spectral phase. The goal of the proposed STSA estimator is to estimate the signal's spectral amplitude $A(k,l)$ given the set of noisy spectral observations $Y_n(k,l)$.

III. Proposed STSA Estimators

In this section, we first derive a novel solution for the MMSE STSA estimation problem in the general multiple microphone case, based on an estimate of the speech spectral phase. Next, an extension of the MMSE based algorithm using a generalized Bayesian cost function is introduced. Finally, the problem of spectral phase estimation is addressed using a multi-microphone MMSE algorithm.

A. Proposed MMSE STSA estimator

An MMSE based spectral amplitude estimator aims at minimization of the MMSE cost function, $E\{(A_k - \hat{A}_k)^2\}$, given the set of spectral observations, where A_k and $\hat{A}_k$ denote the clean and estimated speech STSA, respectively and $E\{.\}$ is the statistical expectation. As discussed in [2], the general form of an MMSE optimal STSA estimate, $\hat{A}_k^{MMSE}$, in the single microphone case, based on the assumption of independent spectral observations, is indeed the conditional expectation of the STSA given the spectral observation, i.e., $E\{A_k|Y_k\}$. In [7], it is stated that for the multiple microphone case, the conditional expectation is replaced by $E\{A_k|\mathbf{Y}_k\}$ with $\mathbf{Y}_k = [Y_{k,1}, Y_{k,2}, ..., Y_{k,N}]^T$ as the vector of spectral observations from all microphones. Note that, hereafter, we omit the time frame index l for brevity and only note the frequency bin number k. In contrast to [2], we base our STSA estimation on treating the spectral phase component $\theta(k)$ as a known parameter that will be replaced by the estimated speech phase later in this section. Using Bayesian rule for the a posteriori probability density function (pdf) of the spectral amplitudes, $p(a|\mathbf{Y}_k)$, we obtain [2]

$$\hat{A}_k^{MMSE} = \frac{\int_0^\infty a p(\mathbf{Y}_k|a,\theta_k)p(a)d_a}{\int_0^\infty p(\mathbf{Y}_k|a,\theta_k)p(a)d_a} \tag{3}$$

with $p(\mathbf{Y}_k|a,\theta_k)$ and $p(a)$ as the conditional pdf of the observations and the STSA prior distribution, respectively. To derive the resulting STSA estimator, further assumptions are required for the aforementioned distributions. In the single microphone case, based on the assumption of complex Gaussian distribution for $p(Y_k|a,\theta_k)$ resulting from the complex noise components, and also Rayleigh distribution for the STSA prior pdf [6], we have

$$p(Y_k|a,\theta_k) = \frac{1}{\pi\sigma_{v_k}^2} \exp\left(-\frac{1}{\sigma_{v_k}^2}|Y_k - ae^{j\theta_k}|^2\right) \tag{4}$$

$$p(a) = \frac{2a}{\sigma_{A_k}^2} \exp\left(-\frac{a^2}{\sigma_{A_k}^2}\right) \tag{5}$$

where $\sigma_{v_k}^2$ and $\sigma_{A_k}^2$ are the spectral noise and speech STSA variances, respectively. We consider a diffuse noise field where the noise component across all microphones is spatially uncorrelated. Using the model in (2), it is inferred that the conditional observation pdf, $p(\mathbf{Y}_k|a,\theta_k)$, can be written as the product of the individual observation pdfs across all microphones [7]. Under this assumption and by considering $Y_{k,n} = R_{k,n}e^{j\theta_{y_{k,n}}}$, we can obtain the conditional joint pdf of the observation vector as

$$p(\mathbf{Y}_k|a,\theta_k) = \prod_{n=1}^{N} p(Y_{k,n}|a,\theta_k) = \prod_{n=1}^{N} \frac{1}{\pi\sigma_{v_k,n}^2} \times \tag{6}$$
$$\exp\left(\sum_{n=1}^{N} \frac{2R_{k,n}a\cos(\theta_k - \theta_{y_{k,n}}) - a^2 - R_{k,n}^2}{\sigma_{v_k,n}^2}\right)$$

where n denotes the microphone index. Substituting (5) and (6) into (3), and using Eq. (3.462.5) and Eq. (3.462.7) in [8] to compute the resulting integrations in (3), the following MMSE STSA estimator is obtained

$$\hat{A}_k^{MMSE} = \frac{-\frac{\nu_k}{2\mu_k^2} + \left(\frac{2\nu_k^2+\mu_k}{4}\right)\sqrt{\frac{\pi}{\mu_k^5}}\exp\left(\frac{\nu_k^2}{\mu_k}\right)\left(1 - erf(\frac{\nu_k}{\sqrt{\mu_k}})\right)}{\frac{1}{2\mu_k} - \left(\frac{\nu_k}{2\mu_k}\right)\sqrt{\frac{\pi}{\mu_k}}\exp\left(\frac{\nu_k^2}{\mu_k}\right)\left(1 - erf(\frac{\nu_k}{\sqrt{\mu_k}})\right)} \tag{7}$$

where $erf(.)$ denotes the Gaussian error function and the parameters μ_k and ν_k are defined as

$$\mu_k = \frac{1}{\sigma_{A_k}^2} + \sum_{n=1}^{N} \frac{1}{\sigma_{v_k,n}^2}\ ,\ \nu_k = -\sum_{n=1}^{N} \frac{R_{k,n}}{\sigma_{v_k,n}^2}\cos(\theta_k - \theta_{y_k,n}) \tag{8}$$

It is observed that, unlike the state-of-the-art spectral amplitude estimation methods, the proposed STSA estimator in (7) does not employ hypergeometric or modified Bessel functions, and instead, exploits one error function term which has less computational load and has a fast convergence rate by using its power series expansion [9].

B. Extension to the auditory based STSA estimator

The MMSE spectral amplitude estimator exploits the most basic cost function, i.e., the expected value of the square error between the clean and estimated STSA. Yet, a few developments of such cost functions have been suggested and used in the literature, such as the auditory based (weighted β-SA) cost function introduced in [4]. Minimization of this parametric Bayesian cost function results in the following STSA estimator

$$\hat{A}_k^{Auditory} = \left(\frac{E\left\{A_k^{\beta_k - 2\alpha_k}|\mathbf{Y}_k\right\}}{E\left\{A_k^{-2\alpha_k}|\mathbf{Y}_k\right\}}\right)^{1/\beta_k} \tag{9}$$

where α_k and β_k are the frequency dependent parameters of the cost function. Hence, to extend our proposed estimator for the auditory based cost function, in light of (9), it is required to obtain the conditional expectation, $E\{A_k^\rho|\mathbf{Y}_k\}$ with ρ as an arbitrary power. In a more general case than that in (3), use of Eq. (3.462.1) in [8] to handle the resulting integration leads to

$$E\{A_k^\rho|\mathbf{Y}_k\} = \frac{\Gamma(\rho+2)D_{-(\rho+2)}\left(\sqrt{\frac{2}{\mu_k}}\nu_k\right)}{\Gamma(2)(2\mu)^{\frac{\rho}{2}}D_{-2}\left(\sqrt{\frac{2}{\mu_k}}\nu_k\right)} \tag{10}$$

with $D_{\cdot}(.)$ as the parabolic cylinder function defined by Eq. (9.24) in [8], $\Gamma(.)$ as the Gamma function, and μ_k and ν_k as given by (8). Now, by using (10) into (9), an auditory based formulation of the STSA estimator is obtained as

$$\hat{A}_k^{Auditory} = \left(\frac{\Gamma(\beta_k - 2\alpha_k + 2) D_{2\alpha_k - \beta_k - 2}\left(\sqrt{\frac{2}{\mu_k}}\nu_k\right)}{\Gamma(2-2\alpha_k)(2\mu)^{\frac{\beta_k}{2}} D_{2\alpha_k - 2}\left(\sqrt{\frac{2}{\mu_k}}\nu_k\right)} \right)^{\frac{1}{\beta_k}} \tag{11}$$

It should be noted that the frequency dependent parameters α_k and β_k are to be selected based on the properties of human auditory system, which is elaborated in [4]. Also, the above considered cost function, and hence the resulting estimator, are simplified to the MMSE estimator discussed in Subsection III-A by choosing α_k and β_k to be zero and one, respectively.

C. Estimation of the Speech Spectral Phase

As stated in Subsection III-A, the spectral phase of the speech signal, i.e., θ_k, is regarded as an unknown parameter which needs to be estimated. In [2], it is proved that an MMSE optimal estimate of the principle value of the phase is simply the noisy phase of the spectral observations, i.e., $\theta_{y_k,n}$. All typical STSA estimators, for the same reason, aim at estimation of the spectral amplitude while keeping the phase unchanged. Nevertheless, recently there has been growing interest in the investigation of the spectral phase component in the spectral estimation process [10]. Whereas in the multi-microphone scenario, averaging schemes can be done across the noisy phases of different observations, in the diffuse noise field, i.e., spatially uncorrelated noise, the following MMSE optimal phase estimator, $\hat{\theta_k}$, has been derived in [7]

$$\tan(\hat{\theta_k}) = \frac{\sum_{n=1}^{N} \left(\sqrt{\zeta_{k,n}}/\sigma_{v_k,n}\right) \Im\{Y_{k,n}\}}{\sum_{n=1}^{N} \left(\sqrt{\zeta_{k,n}}/\sigma_{v_k,n}\right) \Re\{Y_{k,n}\}} \tag{12}$$

where $\zeta_{k,n}$ is the a priori SNR defined as $\sigma^2_{A_k}/\sigma^2_{v_k,n}$, and $\Re\{.\}$ and $\Im\{.\}$ denote the real and imaginary parts. While in the single microphone case, we still use the input noisy phase as the spectral phase estimate to be combined with the estimated STSA, the aforementioned method is employed to estimate the spectral phase component in the multi-microphone case.

IV. Performance Assessment

In this section, we investigate the performance of the proposed multiple microphone speech enhancement methods in terms of objective measures. We choose clean speech sentences from TIMIT database [11] and additive babble noise at various SNR points from NOISEX-92 [12]. The sampling rate is set to 8 kHz and a time segmental length of 20 ms is chosen for the STFT. To evaluate the estimators parameters as in (8), the noise and speech variances are required. Due to the non-stationarity of the considered noise scenarios, estimates of the noise variance $\sigma^2_{v_k}$ are obtained using Cohen's soft-decision IMCRA method [13]. The spectral speech variance $\sigma^2_{x_k}$ is then obtained by the product $\zeta_k.\sigma^2_{v_k}$ where ζ_k is the a priori SNR, which in turn, is estimated by the decision-directed approach [2]. For the single channel scenario, we consider the proposed MMSE and auditory based estimators, defined respectively in (7) and (11) with $N = 1$, and compare their performance to that of the MMSE [2] and logarithmic MMSE (Log MMSE) [3] estimators. In the multi-microphone case, however, comparisons are made with the multi-microphone generalizations of the MMSE and Log MMSE methods introduced in [7].

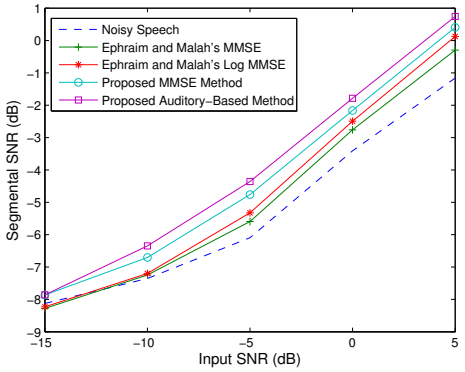

Fig. 1. Segmental SNR for the proposed and conventional STSA estimators versus the input SNR for the single-microphone case.

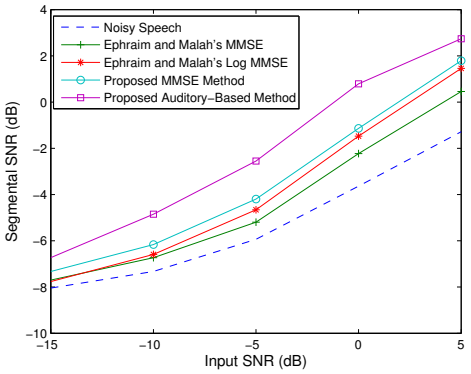

Fig. 2. Segmental SNR for the proposed and conventional STSA estimators versus the input SNR for the multi-microphone case with $N = 2$.

To assess the noise reduction performance of our methods, we investigate the segmental SNR of the enhanced speech signals for both the single and multiple microphone scenarios. Fig. 1 shows the segmental SNR measure of the output speech signal for the discussed methods versus the input SNR of the noisy speech signal for the single microphone case. The results corresponding to the two-microphone case are shown in Fig. 2. It is apparent that in the single microphone scenario, the proposed estimators reach better performance scores compared to the conventional estimators for a moderate range of the input SNR. This advantage along with the relative computational efficiency is inherent in the structure of the proposed MMSE estimator. As for the multi-microphone scenario, the same trend appears to be true while the auditory based estimation approach largely outperforms the other estimators. This proves the usefulness of the employment of more elaborate cost functions in the multi-microphone STSA estimation methods. Next, we assess the performance of the proposed STSA estimators using the log-likelihood ratio (LLR). Whereas the segmental SNR is associated with the amount of noise reduction in the enhanced speech, the LLR measure commonly corresponds to the level of distortion in output speech signal, with smaller LLR values indicating smaller speech distortion. Figs. 3 and

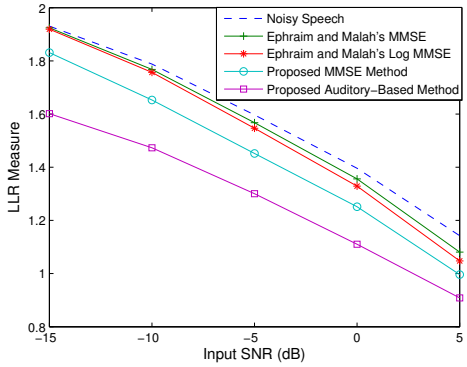

Fig. 3. LLR performance measure for the proposed and conventional STSA estimators versus the input SNR for the single-microphone case.

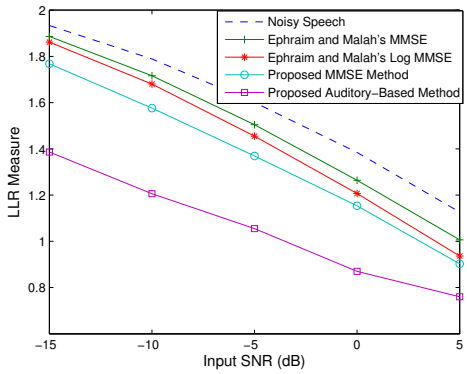

Fig. 4. LLR performance measure for the proposed and conventional STSA estimators versus the input SNR for the multi-microphone case with $N = 2$.

4 indicate the LLR measurement in the cases of the single and multiple microphones, respectively. In a fashion similar to that for the segmental SNR measure, both the MMSE and auditory based STSA estimators outperform Ephraim and Malah's estimators with the auditory based estimator showing a higher level of improvement. This leaves an open topic for further investigating the structure of the Bayesian cost functions used in the STSA estimators.

TABLE I. EXECUTION TIME PER SECOND OF THE INPUT SPEECH SIGNAL FOR DIFFERENT METHODS

Method	Normalized Execution Time
Conventional MMSE with $N = 1$	0.77
Proposed MMSE with $N = 1$	0.54
Conventional MMSE with $N = 4$	1.29
Proposed MMSE with $N = 4$	0.98

We have also experimentally evaluated the performance of the proposed MMSE method in terms of the computational load. As discussed in Section III, the proposed estimators use the Gaussian error function as opposed to the computationally lengthy hypergeometric or modified Bessel functions used in the conventional STSA estimators. Table I illustrates the execution time needed to process each second of the noisy speech files of large sizes on a dual core i7 CPU at 2.80GHz with 4 GB of RAM. We measured the time length needed for the execution of the methods in Matlab and normalized it with respect to the time length of the input noisy speech. The MMSE estimator in (7) is considered herein for comparison with the conventional MMSE estimator for $N = 1$ and $N = 4$. Smaller normalized execution times for the proposed STSA estimation methods prove their relative computational efficiency in comparison with their previous counterparts.

V. CONCLUSIONS

In this paper, we have proposed new structures for the conventional Ephraim and Malah's speech STSA estimators, by using the spectral phase estimate of the speech signal. In addition to providing superior performance in terms of objective measures, the proposed Bayesian MMSE method is found to be advantageous in terms of computational complexity over the conventional MMSE estimator. Extension of the introduced STSA estimators to the multiple microphone scheme in the case of diffuse noise fields was also studied and the corresponding results demonstrated that the use of more elaborated Bayesian cost functions, e.g., the auditory based cost function, is highly beneficial in such scenarios. Based on this study, innovation of spectral phase estimators and use of perceptually more relevant Bayesian cost functions are expected to further improve the current STSA estimation methods.

REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, Signal Processing and Communications, CRC Press, 2007.

[2] Y. Ephraim, D. Malah, *Speech Enhancement Using a Minimum Mean Square Error Short-Time Spectral Amplitude Estimator*, IEEE Trans. Acoustics, Speech and Signal Process., Vol. 32, No. 6, pp. 1109-1121, Dec. 1984.

[3] Y. Ephraim, D. Malah, *Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator*, IEEE Trans. Acoustics, Speech and Signal Process., Vol. 33, No. 2, pp. 443-445, Apr. 1985.

[4] E. Plourde, B. Champagne, *Auditory-Based Spectral Amplitude Estimators for Speech Enhancement*, IEEE Trans. Audio, Speech and Language Process., Vol. 16, No. 8, pp. 1614-1623, Nov. 2008.

[5] 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 and Language Process., Vol. 15, No. 6, pp. 1741-1752, Aug. 2007.

[6] P. C. Loizou, *Speech Enhancement Based on Perceptually Motivated Bayesian Estimators of the Magnitude Spectrum*, IEEE Trans. Speech and Audio Process., Vol. 13, No. 5, pp. 857-869, Sep. 2005.

[7] M. B. Trawicki, M.T. Johnson, *Distributed Multichannel Speech Enhancement with Minimum Mean-Square Error Short-Time Spectral Amplitude,log-Spectral Amplitude, and Spectral Phase Estimation*, Signal Processing 92 (2012), pp. 345-356, Feb. 2012.

[8] I. S. Gradshteyn and I.M. Ryzhik, *Table of Integrals, Series and Products*, 7th edition, Academic Press, 2007.

[9] J. A. C. Weideman, *Computation of the Complex Error Function*, SIAM Journal on Numerical Analysis, Vol. 31, No. 5, pp. 1497-1518, Oct. 1994.

[10] K. Paliwal, K. Wojcicki, B. Shannon, *The importance of phase in speech enhancement*, Speech Communication 53 (2011), pp. 465-494, Apr. 2011.

[11] J. S. Garofolo, *DARPA TIMIT Acoustic-Phonetic Speech Database*, Boulder, CO: NIST, 1988.

[12] NOISEX-92, *Signal processing information base: Noise data*, Rice University, Houston, TX [Online]. Available at http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html

[13] I. Cohen, *Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging*, IEEE Trans. Speech and Audio Process., Vol. 11, No. 5, pp. 466-475, Sep. 2003.