Decision-directed speech power spectral density matrix estimation for multichannel speech enhancement

Yu Gwang Jin, Jong Won Shin and Nam Soo Kim

Citation: The Journal of the Acoustical Society of America **141**, EL228 (2017); doi: 10.1121/1.4977098

View online: https://doi.org/10.1121/1.4977098

View Table of Contents: https://asa.scitation.org/toc/jas/141/3

Published by the Acoustical Society of America

Decision-directed speech power spectral density matrix estimation for multichannel speech enhancement

Yu Gwang Jin
Corporate R&D Center, SK Telecom Co., Ltd., 65 Eulji-ro, Jung-gu, Seoul 04539, Korea
ygjin@sk.com

Jong Won Shin[a)]
School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, 123 Cheomdan-gwagiro, Buk-gu, Gwangju 61005, Korea
jwshin@gist.ac.kr

Nam Soo Kim
School of Electrical and Computer Engineering and Institute of New Media and Communications, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea
nkim@snu.ac.kr

Abstract: In this letter, a multichannel decision-directed approach to estimate the speech power spectral density (PSD) matrix for multichannel speech enhancement is proposed. There have been attempts to build multichannel speech enhancement filters which depend only on the speech and noise PSD matrices, for which the accurate estimate of the clean speech PSD matrix is crucial for a successful noise reduction. In contrast to the maximum likelihood estimator which has been applied conventionally, the proposed decision-directed method is capable of tracking the time-varying speech characteristics more robustly and improves the noise reduction performance under various noise environments.

[DDO]
Date Received: October 24, 2016 **Date Accepted:** February 7, 2017

1. Introduction

The main purpose of speech enhancement is to estimate the desired clean speech signal from the observations corrupted by unwanted interferences and additive noises.[1] During the past decades, a number of multichannel speech enhancement approaches have been proposed.[2–6] In Ref. 7, new simplified expressions for the speech distortion weighted multichannel Wiener filter (SDW-MWF), the minimum variance distortionless response (MVDR) beamformer, and the generalized sidelobe canceller (GSC) were proposed which depend only on the complex power spectral density (PSD) matrices of the signals, instead of the channel transfer functions or the location of microphones and sound sources.

Since the multichannel speech and noise PSD matrices become the only statistics required to determine the final gain function in Ref. 7 and the following works,[8–10] it is certain that an accurate estimation of these PSD matrices is the key to a successful noise reduction. In order to estimate the time-varying noise PSD matrix, the multichannel minima controlled recursive averaging (MCRA) technique[8] has been applied to the recent approaches for multichannel speech enhancement.[8–10] As for the speech PSD matrix, the maximum likelihood (ML) estimation technique which turns out to be a simple subtraction of the noise PSD matrix from the noisy input PSD matrix has been widely adopted.[5–10] However, the ML estimation approach based only on the temporally smoothed statistics of the input signal is not sufficient to track the nonstationary speech signals.

In this letter, we propose a decision-directed (DD) approach to estimate the complex clean speech PSD matrix for the multichannel speech enhancement. In a similar way to the single channel DD approach,[11] the processed output of the previous frame is combined with the estimate by the ML approach to derive the proposed speech PSD matrix estimate. Since the complex speech PSD matrix estimate could be used to obtain not only the multichannel noise suppression gain but also the

[a)]Author to whom correspondence should be addressed.

multichannel speech presence probability (SPP), the proposed method can also improve the performance of various noise estimators and speech enhancement modules which requires SPP, such as Refs. 8 and 12–14. From a number of experiments on multichannel speech enhancement, the proposed DD speech PSD matrix estimator showed better performances compared with the conventional ML estimator.

2. Multichannel speech enhancement techniques

Compared with single microphone-based techniques, multichannel speech enhancement approaches could achieve more effective noise reduction without much speech distortion as spatial diversity can also be exploited. Classical beamformers such as the MVDR beamformer[2] and the GSC (Ref. 3) that reduce interfering components by steering the array to the direction of signals of interest require the estimation of the direction of the desired speaker with respect to the microphone locations or the channel transfer functions, which may be quite difficult in real environments.

In Refs. 1 and 7, an optimal multichannel filtering technique which depends only on the statistics of the speech and noise signals at the microphones was proposed with new simplified expressions of the SDW-MWF, the MVDR filter, and the GSC. These are dependent on the channel transfer functions only through the multichannel speech and noise PSD matrices, and it was further extended to a spectro-temporal filtering to exploit temporal and spectral correlations.[9]

Let $\mathbf{y}(k,t)$, $\mathbf{x}(k,t)$, and $\mathbf{v}(k,t)$ denote the N-dimensional vectors which consist of the short-time Fourier transform coefficients of the noisy speech, clean speech, and additive noise signal, respectively, for the kth frequency bin at frame t observed from N microphones. The output signal $\hat{\mathbf{x}}(k,t)$, which is an estimate of $\mathbf{x}(k,t)$ is then obtained by applying a noise suppression gain $\mathbf{g}(k,t)$ to $\mathbf{y}(k,t)$ in the following way:

$$\hat{\mathbf{x}}(k,t) = \mathbf{g}^H(k,t)\mathbf{y}(k,t) = \mathbf{g}^H(k,t)(\mathbf{x}(k,t) + \mathbf{v}(k,t)), \tag{1}$$

where the superscript ^{H} denotes the transpose-conjugate operator. When the $N \times N$ dimensional multichannel complex PSD matrices of the noisy speech, clean speech, and noise are defined as $\Phi_{yy}(k,t) \triangleq E\{\mathbf{y}(k,t)\mathbf{y}^H(k,t)\}$, $\Phi_{xx}(k,t) \triangleq E\{\mathbf{x}(k,t)\mathbf{x}^H(k,t)\}$, and $\Phi_{vv}(k,t) \triangleq E\{\mathbf{v}(k,t)\mathbf{v}^H(k,t)\}$, respectively, the gain $\mathbf{g}(k,t)$ in Eq. (1) can be derived while depending only on the PSD matrix estimates $\hat{\Phi}_{yy}(k,t), \hat{\Phi}_{xx}(k,t)$, and $\hat{\Phi}_{vv}(k,t)$. In this work, we adopted the gain function for the SDW-MWF incorporating the SPP $p(k,\ t)$[15] given by

$$\mathbf{g}(k,t) = \left[\hat{\Phi}_{xx}(k,t) + \hat{\Phi}_{vv}(k,t)/\hat{p}(k,t)\right]^{-1}\hat{\Phi}_{xx}(k,t). \tag{2}$$

$p(k,\ t)$ can be estimated based on a Gaussian model[16] as $\hat{p}(k,t) = \Lambda(k,t)/[1+\Lambda(k,t)]$ in which

$$\Lambda(k,t) = \frac{1/\hat{q}(k,t) - 1}{1 + tr\left[\hat{\Phi}_{vv}^{-1}(k,t)\hat{\Phi}_{xx}(k,t)\right]} \exp\left\{\frac{\mathbf{y}(k,t)\hat{\Phi}_{vv}^{-1}(k,t)\hat{\Phi}_{xx}(k,t)\hat{\Phi}_{vv}^{-1}(k,t)\mathbf{y}^H(k,t)}{1 + tr\left[\hat{\Phi}_{vv}^{-1}(k,t)\hat{\Phi}_{xx}(k,t)\right]}\right\}, \tag{3}$$

where $tr[\,]$ is a trace of a matrix and the *a priori* probability of speech absence $q(k,\ t)$ is estimated as in Ref. 8.

3. DD speech PSD matrix estimation

Accurate estimation of the multichannel speech and noise PSD matrices is crucial for successful noise reduction and SPP estimation as can be seen from Eqs. (2) and (3). The estimated SPP can also be utilized for other noise tracking or speech enhancement modules. For the noise statistics estimation, it is common to recursively average past statistics of the noisy input signal depending on the SPP estimates as given by

$$\hat{\Phi}_{vv}(k,t) = \tilde{\alpha}_v(k,t)\hat{\Phi}_{vv}(k,t-1) + (1-\tilde{\alpha}_v(k,t))\left[\mathbf{y}(k,t)\mathbf{y}^H(k,t)\right], \tag{4}$$

where $\tilde{\alpha}_v(k,t) = \alpha_v + (1-\alpha_v)\hat{p}(k,t)$ is a time-varying frequency-dependent smoothing parameter which is a function of the SPP estimate $\hat{p}(k,t)$ and $0 < \alpha_v < 1$. In this work, a multichannel version[8] of the MCRA algorithm[17] which is popular for single channel speech enhancement was applied to estimate the SPP and noise PSD matrix.

For the estimation of clean speech statistics, the simple ML method has been commonly used in Refs. 7–10 as given by $\hat{\Phi}_{xx}(k,t) = \hat{\Phi}_{yy}(k,t) - \hat{\Phi}_{vv}(k,t)$, where $\hat{\Phi}_{yy}(k,t)$ can be obtained by a temporal smoothing of $\mathbf{y}(k,t)\mathbf{y}^H(k,t)$. However, the ML-based estimation techniques occasionally incur musical noises[18] when $\hat{\Phi}_{yy}(k,t)$ is not smoothed enough, and cannot track the rapidly varying speech statistics when $\hat{\Phi}_{yy}(k,t)$ is smoothed too much.

In order to alleviate this difficulty, we propose a novel estimation method for the complex speech PSD matrix, which can be considered to be an extension of the DD approach[11] to the multichannel case. For single channel speech enhancement, the DD approach was proposed[11] to estimate the *a priori* signal-to-noise ratio (SNR), which has been proven to provide the improved subjective quality of the output speech. However, the estimation of the multichannel counterpart of the *a priori* SNR $\Phi_{vv}^{-1}(k,t)\Phi_{xx}(k,t)$ may not be reliable enough since the complex noise PSD matrix becomes almost singular when one or a few strong noise sources produce highly directional noise fields. Therefore, a time-varying multichannel speech PSD matrix $\Phi_{xx}(k,t)$ is estimated instead based on the DD scheme as follows:

$$\hat{\Phi}_{xx}(k,t) = \alpha_x \hat{\mathbf{x}}(k,t-1)\hat{\mathbf{x}}^H(k,t-1) + (1-\alpha_x)\left[\mathbf{y}(k,t)\mathbf{y}^H(k,t) - \hat{\Phi}_{vv}(k,t)\right]_{\geq 0}, \tag{5}$$

where $0 < \alpha_x < 1$ is a smoothing parameter and $[\Phi]_{\geq 0}$ denotes the positive semidefinite matrix closest to Φ. In our implementation, however, we just replace $[\Phi]_{\geq 0}$ by Φ simply because we found that the matrix modification by eigendecomposition did not show any notable improvement of overall speech quality while requiring heavy computation of the eigen analysis. It is noted that the spectral amplitudes in the formulation of the single channel DD estimation approach are replaced by complex spectral vectors.

In Eq. (5), the speech PSD matrix $\Phi_{xx}(k,t)$ is estimated by a weighted sum of two different terms. It is clear that the first one $\hat{\mathbf{x}}(k,t-1)\hat{\mathbf{x}}^H(k,t-1)$ is an instantaneous estimate of $\Phi_{xx}(k,t)$ derived from the previous frame. The other term comes from the ML estimation approach except the current input power spectrum matrix is used instead of the temporally smoothed noisy PSD matrix. As a result, the proposed multichannel DD approach for the speech PSD matrix estimation reflects the enhanced speech components of the previous frame and the current input components in a more direct way, which may lead to a rapid tracking of the time-varying speech PSD matrix without introducing much artifact. Since the accurate complex speech PSD matrix estimate is helpful not only to obtain the proper noise suppression gain but also to estimate the SPP as in Ref. 16 and consequentially the noise PSD matrix in Eq. (4) more precisely, the proposed multichannel DD approach can provide enhanced speech signals with very little musical noise even in a quite noisy environment.

4. Experimental results

In order to show the effectiveness of the proposed multichannel DD approach for speech PSD matrix estimation, the quality of output speech enhanced by SDW-MWF in Eq. (2) using the speech PSD matrix estimated by the proposed or conventional method was evaluated under various noise conditions. We have recorded spoken utterances and interference signals with a commercial smartphone, Samsung Galaxy S4, SHV-E300L (Samsung Electronics Co., Ltd., Suwon, Korea) which has two microphones about 140 mm away from each other. Overall geographical placement of the sound sources and receiver is illustrated in Fig. 1. One person stood in the center of a reverberant room with size $3119 \times 3232 \times 2080\,\mathrm{mm}^3$ holding a phone with the right hand, exactly in the same way as in a usual telecommunication scenario with the handset mode. Twenty sentences spoken by the person and interference signals played by loudspeakers from eight different locations at the distance of 1000 mm were recorded individually, and then mixed with 0, 5, 10, and 15 dB SNR. The interference signals used for the experiments were destroyer, F-16, and factory noise from NOISEX-92 database. Each signal was sampled at 16 kHz and a half-overlapped Hann window of length 512 was applied. In this work, we set $\alpha_v = 0.92$, the same as in Ref. 8, and $\alpha_x = 0.95$ which was experimentally determined.

We have measured the quality of the output signals in terms of the perceptual evaluation of speech quality (PESQ) score.[19] The PESQ scores for eight different directions of noise which are averaged over all types of interferences are shown in Fig. 2 for each SNR. With any angles and levels of the interfering noise signals, the SDW-MWF utilizing the proposed multichannel DD method for the speech PSD estimation consistently outperformed that utilizing the ML method in terms of PESQ scores.

We have also measured the improvement of the speech quality by the proposed technique in terms of the SNR improvement, the difference between the input and output SNRs,[17] in dB scale. The SNR improvements for each noise type averaged over all loudspeaker positions are summarized in Table 1. From the results, we can see that the proposed multichannel DD approach for speech PSD matrix estimation

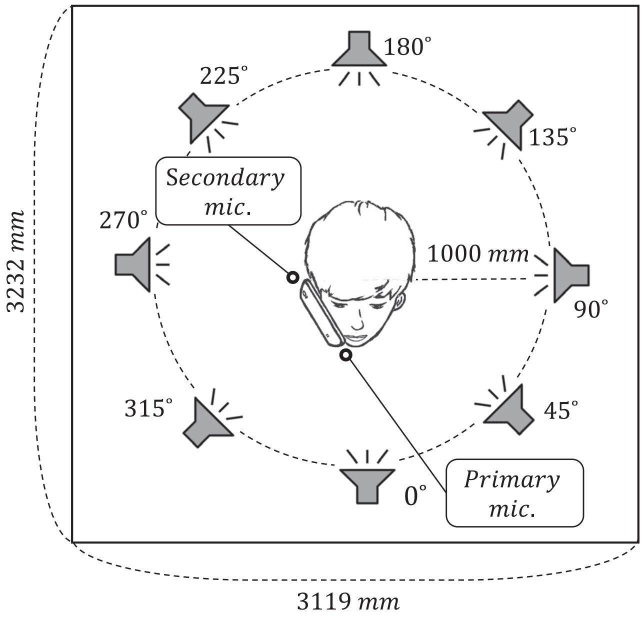

Fig. 1. The geographical placement of the noise sources and receivers.

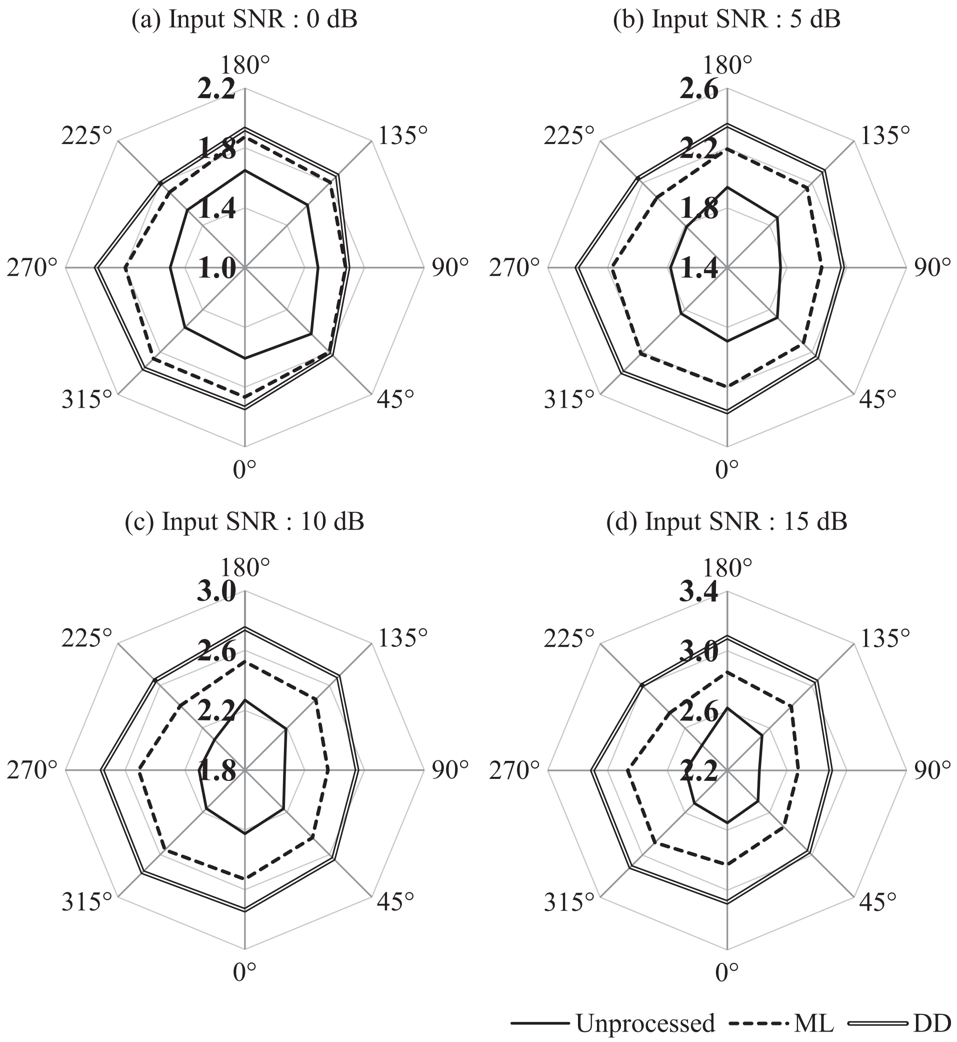

Fig. 2. The PESQ scores by the ML and the DD speech PSD matrix estimation approaches.

Table 1. The SNR improvements under various noisy conditions by the ML and the DD speech PSD matrix estimation approaches.

Noise	Destroyer		F-16		Factory		Avg.	
Input SNR	ML	DD	ML	DD	ML	DD	ML	DD
0 dB	5.74	9.18	5.68	9.61	2.65	4.99	4.69	7.93
5 dB	5.89	9.24	5.41	9.12	3.15	5.78	4.82	8.05
10 dB	5.54	8.68	4.90	8.40	3.13	5.85	4.54	7.65
15 dB	4.99	8.01	4.21	7.53	3.03	5.86	4.08	7.13

outperformed the conventional ML approach in terms of both the PESQ scores and SNR improvements for all types and locations of interference signals.

5. Conclusions

In this letter, we have proposed a multichannel DD approach to estimate the complex speech PSD matrix for multichannel speech enhancement. In contrast to the conventional ML estimation, the proposed DD method takes the processed output of the previous frame into account and interpolated it with the estimate by the ML approach, which enables an effective tracking of the time-varying speech statistics. A number of experiments have confirmed that the proposed DD estimation approach for the multichannel speech PSD matrix considerably improved the speech quality of signals when applied to the SDW-MWF compared with the ML estimation technique under various noisy environments.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) (Grant No. NRF-2016R1C1B1015291).

References and links

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing* (Springer-Verlag, Berlin, Germany, 2008).

[2] O. L. Frost III, "An algorithm for linearly constrained adaptive array processing," Proc. IEEE **60**(8), 926–935 (1972).

[3] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," IEEE Trans. Antennas Propagat. **AP-30**(1), 27–34 (1982).

[4] S. Gannot, D. Burstein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," IEEE Trans. Signal Process. **49**(8), 1614–1626 (2001).

[5] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multichannel Wiener filtering for noise reduction," Signal Process. **84**(12), 2367–2387 (2004).

[6] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction," Speech Commun. **49**, 636–656 (2007).

[7] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," IEEE Trans. Audio, Speech, Lang. Process. **18**(2), 260–276 (2010).

[8] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," IEEE Trans. Audio, Speech, Lang. Process. **19**(7), 2159–2169 (2011).

[9] Y. G. Jin, J. W. Shin, and N. S. Kim, "Spectro-temporal filtering for multichannel speech enhancement in short-time Fourier transform domain," IEEE Signal Process. Lett. **21**(3), 352–355 (2014).

[10] Y. G. Jin, J. W. Shin, C. M. Lee, S. H. Bae, and N. S. Kim, "Parametric multichannel noise reduction algorithm utilizing temporal correlations in reverberant environment," in *Proceedings of the IEEE International Conference of Acoustical, Speech, and Signal Processing* (May 2014), pp. 7099–7102.

[11] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Process. **32**(6), 1109–1121 (1984).

[12] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," IEEE Trans. Speech Audio Process. **12**(6), 561–571 (2004).

[13] M. Taseska and E. A. P. Habets, "MMSE-based blind source extraction in diffuse noise fields using a complex coherence-based a priori SAP estimator," in *Proceedings of the International Workshop on Acoustical Signal Enhancement* (September 2012).

[14] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "Spherical harmonic domain noise reduction using an MVDR beamformer and DOA-based second-order statistics estimation," in *Proceedings of the IEEE International Conference on Acoustical, Speech, and Signal Processing* (May 2013), pp. 654–658.

[15] K. Ngo, A. Spriet, M. Moonen, J. Wouters, and S. H. Jensen, "Variable speech distortion weighted multichannel Wiener filter based on soft output voice activity detection for noise reduction in hearing aids," in *Proceedings of the International Workshop on Acoustical Echo Noise Control* (July 2008).

[16] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," IEEE Trans. Audio, Speech, Lang. Process. **18**(5), 1072–1077 (2010).

[17] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," IEEE Trans. Speech Audio Process. **11**(5), 466–475 (2003).

[18] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," IEEE Trans. Speech Audio Process. **2**(2), 345–349 (1994).

[19] ITU, "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 Rec. P. 862 (2000).