IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

A postfiltering approach for dual-microphone smartphones

Juan M. Martín-Doñas[1], *Iván López-Espejo*[2], *Angel M. Gomez*[1], *Antonio M. Peinado*[1]

[1]Dept. de Teoría de la Señal, Telemática y Comunicaciones, Universidad de Granada, Spain
[2]VeriDas | das-Nano, Spain

{mdjuamart,amgg,amp}@ugr.es, ilopez@das-nano.com

Abstract

Although beamforming is a powerful tool for microphone array speech enhancement, its performance with small arrays, such as the case of a dual-microphone smartphone, is quite limited. The goal of this paper is to study different postfiltering approaches that allow for further noise reduction. These postfilters are applied to our previously proposed extended Kalman filter framework for relative transfer function estimation in the context of minimum variance distortionless response beamforming. We study two different postfilters based on Wiener filtering and non-linear estimation of the speech amplitude. We also propose several estimators of the clean speech power spectral density which exploit the speaker position with respect to the device. The proposals are evaluated when applying speech enhancement on a dual-microphone smartphone in different noisy acoustic environments, in terms of both perceptual quality and speech intelligibility. Experimental results show that our proposals achieve further noise reduction in comparison with other related approaches from the literature.

Index Terms: Postfiltering, Extended Kalman filter, Minimum variance distortionless response, Dual-microphone speech, Smartphone

1. Introduction

Multi-channel speech processing is widely employed in devices with several microphones to improve the noise reduction performance, yielding better speech quality and/or intelligibility. The most common techniques for multi-channel processing are the beamforming algorithms, which apply a spatial filtering to the existing sound field [1, 2, 3]. Nevertheless, the performance of beamforming can be insufficient if a small number of microphones is considered, as in the case of dual-microphone smartphones [4, 5]. Thus, to improve the noise reduction performance, the enhanced signal at the output of the beamformer might be further processed by using postfiltering methods [6].

Several postfilters have been proposed in the recent years mainly based on multi-channel Wiener filtering, which can be decomposed into minimum variance distortionless response (MVDR) beamforming plus single-channel Wiener filtering [3]. For multi-channel Wiener filtering, Zelinski [7] assumed a spatially-white noise to estimate the speech and noise statistics. Marro *et al.* [8] further improved the postfilter architecture and also considered acoustic echo and reverberation. The previous spatially-white noise assumption was substituted in [9] by assuming a diffuse noise field. The postfilter presented in [10] took into account the noise reduction provided by the beamformer to obtain more accurate statistics. That work also explores non-linear postfilters based on minimum mean square error (MMSE) estimation of the speech amplitude. Gannot *et al.* [11] modified the beamformer using a generalized sidelobe canceler (GSC) structure and an optimal modified log-spectral amplitude (OMLSA) estimator [12] as a postfilter. Apart from the above, a statistical analysis of dual-channel postfilters in isotropic noise fields is presented in [13].

In the case of dual-microphone smartphones, other works have exploited the information of the secondary microphone to enhance the speech signal from the reference microphone. For example, Jeub *et al.* [14] proposed an estimator of the noise power spectral density (PSD) along with a modified single-channel Wiener filter that explicitly exploits the power level difference (PLD) of the speech signal between the microphones in close-talk (CT) conditions (when the loudspeaker of the smartphone is placed at the ear of the user). Nelke *et al.* [15] developed an alternative noise PSD estimator in far-talk (FT) conditions (when the user holds the device at a distance from her/his face). This method combines a single-channel speech presence probability estimator and the coherence properties of the dual-channel target signal and background noise. The noise PSD is employed to estimate the gain function to be applied on the reference channel. Such an algorithm was extended for multi-microphone devices in [16]. Nevertheless, all of these techniques make assumptions about the noise field properties that may not be accurate in practice, thereby leading to a limited performance.

Recently, we proposed an estimator of the relative transfer function (RTF) between microphones based on an extended Kalman filter (eKF) framework [17]. Our method is capable of tracking the RTF evolution using prior information on the channel and noise statistics without making any assumption on the clean speech signal statistics. In that work we evaluated the performance of our estimator over MVDR beamforming applied on a dual-microphone smartphone configuration in CT and FT conditions, showing improvements in estimation accuracy with respect to other relevant methods from the literature. Despite this, the speech enhancement performance is still limited as a result of using beamforming with only two microphones.

In this paper we evaluate the use of postfiltering techniques to overcome shortcomings of our eKF-based MVDR approach for dual-microphone smartphones. We compare different algorithms and modify them to adapt them to our eKF-based method. Also, we propose different clean speech PSD estimators and make use of the available information about the RTF and noise to obtain the needed statistics for postfiltering. Our proposals are evaluated on a dual-microphone smartphone under several noisy acoustic environments in CT and FT condi-

This work has been supported by the Spanish MINECO/FEDER Project TEC2016-80141-P and the Spanish Ministry of Education through the National Program FPU (grant reference FPU15/04161).

10.21437/IberSPEECH.2018-30

tions, achieving improvements in terms of noise reduction performance in comparison with other state-of-the-art approaches.

The remainder of this paper is organized as follows. In Section 2 we briefly revisit the eKF-based RTF estimation and its application to MVDR beamforming. Section 3 describes the proposed postfiltering approaches and clean speech PSD estimators for CT and FT conditions. Then, in Section 4 the experimental framework is presented along with our perceptual quality and speech intelligibility results. Finally, conclusions are summarized in Section 5.

2. Beamforming for dual-microphone smartphones

Before presenting the postfiltering approaches, it is worthwhile to review the RTF estimation and beamforming for dual-microphone smartphones that we proposed in [17]. First, we introduce formulation proposed for the eKF-based RTF estimation. Next, we describe the MVDR beamforming approach for processing the dual-channel noisy speech signals using the noise statistics and RTF estimations.

2.1. Extended Kalman filter-based RTF estimation

Let us consider the following additive distortion model for the noisy speech signal in the short-time Fourier transform (STFT) domain,

$$Y_m(f,t) = X_m(f,t) + N_m(f,t), \tag{1}$$

where $Y_m(f,t)$, $X_m(f,t)$ and $N_m(f,t)$ represent, respectively, noisy speech, clean speech and noise STFT coefficients at the m-th microphone ($m = 1, 2$), f is the frequency bin and t the frame index. Using the relative transfer function (RTF) $A_{21}(f,t) = \frac{X_2(f,t)}{X_1(f,t)}$ between both microphones, we can write the speech distortion model for the secondary microphone ($m = 2$) in terms of the reference microphone ($m = 1$) as

$$Y_2(f,t) = A_{21}(f,t)\,(Y_1(f,t) - N_1(f,t)) + N_2(f,t). \tag{2}$$

We can also rewrite the previous complex variables as vectors stacking their real and imaginary parts, yielding $\mathbf{y}_m^{(t)}$, $\mathbf{a}_{21}^{(t)}$ and $\mathbf{n}_m^{(t)}$ (index f is omitted for clarity). For example, we define the noisy speech vector for the m-th microphone as

$$\mathbf{y}_m^{(t)} = \begin{bmatrix} Re(Y_m(t)), & Im(Y_m(t)) \end{bmatrix}^\top, \tag{3}$$

where $[\cdot]^\top$ indicates transpose. Then, we set a dynamic model for $\mathbf{a}_{21}^{(t)}$ as follows,

$$\mathbf{a}_{21}^{(t)} = \mathbf{a}_{21}^{(t-1)} + \mathbf{w}^{(t)}, \tag{4}$$

where $\mathbf{w}^{(t)}$ models the variability of the RTF between consecutive frames. Also, we redefine (2), using the previous vector notation, as

$$\begin{aligned} \mathbf{y}_2^{(t)} &= \mathbf{h}\left(\mathbf{a}_{21}^{(t)}, \mathbf{n}_1^{(t)}; \mathbf{y}_1^{(t)}\right) + \mathbf{n}_2^{(t)} \\ &= \left[\mathbf{C}\left(\mathbf{y}_1^{(t)} - \mathbf{n}_1^{(t)}\right), \quad \mathbf{D}\left(\mathbf{y}_1^{(t)} - \mathbf{n}_1^{(t)}\right)\right] \mathbf{a}_{21}^{(t)} + \mathbf{n}_2^{(t)}, \end{aligned} \tag{5}$$

where $\mathbf{C} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $\mathbf{D} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$.

Assuming multivariate Gaussian variables and using the models in (4) and (5), in [17] we proposed an MMSE estimator of $\mathbf{a}_{21}^{(t)}$ using an extended Kalman filter (eKF) framework, which tracks the RTF using the observable noisy speech, estimated noise statistics and a priori information on the RTF statistics [17].

2.2. MVDR beamforming

Once the RTF is estimated, both noisy signals are combined using MVDR beamforming, whose weights can be expressed (omitting indices f and t for clarity) as [3]

$$\mathbf{F} = \frac{\mathbf{\Phi}_{nn}^{-1}\mathbf{d}}{\mathbf{d}^H \mathbf{\Phi}_{nn}^{-1}\mathbf{d}}, \tag{6}$$

where $(\cdot)^H$ indicates Hermitian transpose, $\mathbf{d} = \begin{bmatrix} 1, & A_{21} \end{bmatrix}^\top$ is the steering vector and $\mathbf{\Phi}_{nn}$ is a noise spatial correlation matrix (i.e., obtained from $\mathbf{N} = \begin{bmatrix} N_1, & N_2 \end{bmatrix}^\top$). We can also express the PSD of the noise at the output of the beamformer as [10]

$$\phi_v = \left(\mathbf{d}^H \mathbf{\Phi}_{nn}^{-1} \mathbf{d}\right)^{-1}. \tag{7}$$

Finally, the enhanced signal for the reference microphone is estimated as $\widehat{X}_{1,\text{mvdr}} = \mathbf{F}^H \mathbf{Y}$, with $\mathbf{Y} = \begin{bmatrix} Y_1, & Y_2 \end{bmatrix}^\top$.

3. Postfiltering for dual-microphone smartphones

The performance of MVDR beamforming when applied on smartphones with only two microphones is quite limited due to the reduced spatial information and the particular placement of the microphones [4]. Therefore, we analyze the use of postfiltering for enhancing the signal at the output of the beamformer and further improve the noise reduction performance. We propose two different postfilters based on Wiener filtering and the optimal modified log-spectral amplitude (OMLSA) estimator [12], and also address the estimation of the clean speech PSD at the reference microphone needed by the postfilters. In addition, the postfiltering gains are further processed using the musical noise reduction algorithm proposed in [18]. This postprocessing is applied to frequencies above 1 kHz. For ease of notation, we drop the indices f and t henceforth.

3.1. Wiener filtering

The multi-channel Wiener filter can be decomposed into an MVDR beamformer followed by a single-channel Wiener filter defined as

$$G_{\text{wf}} = \frac{\xi}{1+\xi}, \tag{8}$$

where $\xi = \phi_{x_1}/\phi_v$ is the a priori signal-to-noise ratio (SNR) and ϕ_{x_1} is the clean speech PSD at the reference microphone. This Wiener filter is partially obtained from the enhanced signal at the output of the beamformer. Better performance can be achieved if the Wiener filter is fully calculated from the reference noisy signal when an overestimated noise is considered. Thus, we propose the following improved Wiener filter

$$G_{\text{iwf}} = \frac{\widehat{\phi}_{x_1}}{\widehat{\phi}_{x_1} + \mu\widehat{\phi}_v}, \tag{9}$$

where $\widehat{\phi}_{x_1}$ is an estimate of the clean speech PSD (discussed in Subsection 3.3), $\widehat{\phi}_v$ is an estimate of the noise PSD, taken as ϕ_{n_1} (first element of the diagonal of $\mathbf{\Phi}_{nn}$) in order to use an overestimated version of the noise, and μ is a factor which provides an increased overestimation. As a result, the clean speech signal is estimated as $\widehat{X}_{1,\text{iwf}} = G_{\text{iwf}}\widehat{X}_{1,\text{mvdr}}$.

3.2. Optimal modified log-spectral amplitude estimator

The OMLSA estimator proposed in [12] computes the postfilter gains as

$$G_{\text{omlsa}} = G_{\text{H1}}{}^{p_{\text{SPP}}} G_{\text{H0}}{}^{1-p_{\text{SPP}}}, \tag{10}$$

where p_{SPP} is the speech presence probability, G_{H0} is a constant gain when speech is absent and G_{H1} is the gain when speech is present, computed as

$$G_{\text{H1}} = G_{\text{wf}} \exp\left(\frac{1}{2} \int_{\frac{\xi}{1+\xi}\gamma}^{\infty} \frac{e^{-t}}{t} dt \right), \tag{11}$$

where $\gamma = |\widehat{X}_{1,\text{mvdr}}|^2/\phi_v$ is the a posteriori SNR and G_{wf} was defined in (8). We modify this gain by substituting G_{wf} by G_{iwf}, defined in (9), which yields the improved OMLSA gain G_{iomlsa}. Finally, the clean speech signal is estimated as $\widehat{X}_{1,\text{iomlsa}} = G_{\text{iomlsa}} \widehat{X}_{1,\text{mvdr}}$.

3.3. Clean speech PSD estimators

The previous postfilters require an estimation of the clean speech PSD, ϕ_{x_1}. We propose two different estimators for CT and FT conditions, respectively, based on the noisy speech and noise statistics and the estimated RTF between microphones. Therefore, these estimators take advantage of the more accurate RTFs obtained by our eKF-based approach.

For close-talk (CT) conditions, the estimator is based on the PLD between microphones [14], which can be computed as

$$\Delta\widehat{\phi}_{\text{PLD}} = \max\left(\phi_{y_1} - \phi_{y_2}, 0\right), \tag{12}$$

where ϕ_{y_1} and ϕ_{y_2} are the noisy speech PSDs at the reference and secondary microphones, respectively. This estimator takes advantage of the more attenuated clean speech component at the secondary microphone. Assuming that the noise PSD is similar at both microphones so that its difference can be neglected compared to $\Delta\widehat{\phi}_{\text{PLD}}$, it can be easily shown that the clean speech PSD can be approximated as [14]

$$\widehat{\phi}_{x_1}^{(\text{CT})} = \frac{\Delta\widehat{\phi}_{\text{PLD}}}{1 - |A_{21}|^2}. \tag{13}$$

Unlike CT conditions, in far-talk (FT) conditions speech power is similar at both microphones and the previous assumptions are inaccurate (i.e., noise PSD difference cannot be neglected compared to PLD between microphones) [14]. Therefore, a better estimator is obtained by considering the distortionless properties of MVDR beamforming, which imply that the clean speech PSD at the reference microphone is the same as the one at the beamformer output. Thus, we estimate the clean speech PSD at both the beamformer output and the reference microphone as

$$\widehat{\phi}_{x_1}^{(\text{FT})} = \mathbf{F}^H \left(\mathbf{\Phi}_{yy} - \mathbf{\Phi}_{nn}\right) \mathbf{F}, \tag{14}$$

where $\mathbf{\Phi}_{yy}$ is a noisy speech spatial correlation matrix (i.e., calculated from $\mathbf{Y}$), whose diagonal elements are the noisy speech PSDs ϕ_{y_1} and ϕ_{y_2}. Although the clean speech estimate could be obtained from a direct subtraction of the first diagonal elements of matrices $\mathbf{\Phi}_{yy}$ and $\mathbf{\Phi}_{nn}$, the estimator defined in (14) has the advantage of using all the channels in the estimation.

4. Experimental evaluation

4.1. Experimental framework

To evaluate the proposed techniques, we simulated dual-channel noisy speech recordings on a Motorola Moto G smartphone. We consider two different modes of use: close-talk (CT) and far-talk (FT). These modes can be easily identified using the proximity sensor included in the smartphone. Clean speech signals were obtained from 18 speakers of the VCTK database [19] downsampled to 16 kHz. We simulated recordings at eight different noisy environments with different reverberations: car, street, babble, mall, bus, cafe, pedestrian street and bus station. The noise signals were added at six different SNR levels from -5 dB to 20 dB. Further details about this database can be found in [17].

For STFT computation, we choose a 25 ms square-root Hann window with 75% overlap. The noisy speech spatial correlation matrix $\mathbf{\Phi}_{yy}$ is estimated by a first-order recursive averaging with an averaging constant of 0.9. The noise spatial correlation matrix $\mathbf{\Phi}_{nn}$ is estimated by recursive averaging during time-frequency bins where speech is absent. Thus, we compute the speech presence probability p_{SPP} at each bin by using the *Multi-Channel Speech Presence Probability* (MC-SPP) noise tracking algorithm proposed in [20]. Finally, we use an overestimation factor $\mu = 4$ and a speech absent gain $G_{\text{H0}} = 0.05$ for postfiltering implementation.

4.2. Results

The two proposed postfilters, Wiener filtering (eKF-WF) and OMLSA estimator (eKF-OMLSA), are evaluated in combination with MVDR beamforming, both using the eKF-based RTF estimator outlined in Subsection 2.1. The obtained results are compared with those achieved by the noisy speech at the reference microphone, and MVDR beamforming with eKF and no postfiltering (eKF-MVDR) [17]. Also, we evaluate two other state-of-the-art enhancement algorithms for dual-microphone smartphones, that is, the PLD-based Wiener filtering for close-talk conditions of [14] and the speech presence probability and coherence-based (SPPC) Wiener filtering for far-talk position of [15]. The musical noise suppressor of [18] is also applied to PLD and SPPC gains.

The resulting enhanced signals are assessed in terms of perceptual quality and speech intelligibility by means of the perceptual evaluation of the speech quality (PESQ) [21] and short-time objective intelligibility (STOI) [22] metrics, respectively. Clean speech at the reference microphone is taken as a reference for these performance metrics. The results for close-talk (CT) and far-talk (FT) conditions are shown in Tables 1 and 2, respectively.

In close-talk conditions, it is shown that the proposed postfilters outperform the other methods in terms of perceptual quality, with both Wiener filtering and OMLSA approaches obtaining similar performance on average. While the Wiener filtering approach achieves better PESQ results at low and medium SNRs, the OMLSA approach yields higher PESQ scores at high SNRs. It can also be seen that PLD is a better choice than eKF-MVDR, but the addition of postfiltering after beamforming leads to better PESQ results. On the other hand, intelligibility scores among the different evaluated techniques are similar, but MVDR beamforming without postfiltering obtains slightly better ones. That means that the superior perceptual quality achieved by postfiltering involves some speech distortion that slightly reduces intelligibility. In general, PLD and Wiener

Table 1: *PESQ and STOI scores obtained for noisy and enhanced speech when using different dual-microphone enhancement techniques in close-talk (CT) conditions.*

Metric	**Method**	**SNR (dB)**						
		-5	**0**	**5**	**10**	**15**	**20**	**Avg.**
PESQ	Noisy	1.10	1.12	1.18	1.31	1.52	1.80	1.34
	PLD	1.15	1.26	1.44	1.67	1.94	2.24	1.62
	eKF-MVDR	1.11	1.16	1.25	1.40	1.62	1.92	1.41
	eKF-WF	**1.19**	**1.30**	**1.49**	**1.73**	2.02	2.34	**1.68**
	eKF-OMLSA	**1.19**	1.29	1.46	1.72	**2.03**	**2.36**	**1.68**
STOI	Noisy	0.53	0.63	0.72	0.79	0.84	0.88	0.73
	PLD	0.54	0.64	0.73	**0.80**	**0.85**	**0.89**	0.74
	eKF-MVDR	**0.56**	**0.65**	**0.74**	**0.80**	**0.85**	**0.89**	**0.75**
	eKF-WF	0.53	0.63	0.72	**0.80**	**0.85**	**0.89**	0.74
	eKF-OMLSA	0.50	0.59	0.70	0.79	**0.85**	**0.89**	0.72

Table 2: *PESQ and STOI scores obtained for noisy and enhanced speech when using different dual-microphone enhancement techniques in far-talk (FT) conditions.*

Metric	**Method**	**SNR (dB)**						
		-5	**0**	**5**	**10**	**15**	**20**	**Avg.**
PESQ	Noisy	1.12	1.13	1.21	1.37	1.61	1.94	1.40
	SPPC	1.18	1.20	1.34	1.55	1.80	2.04	1.52
	eKF-MVDR	1.12	1.17	1.28	1.47	1.74	2.09	1.48
	eKF-WF	**1.25**	1.33	1.51	**1.80**	**2.17**	**2.56**	**1.77**
	eKF-OMLSA	1.24	**1.34**	**1.53**	**1.80**	2.14	2.49	1.76
STOI	Noisy	0.52	0.62	0.73	0.81	0.87	0.91	0.74
	SPPC	0.41	0.52	0.63	0.72	0.78	0.82	0.65
	eKF-MVDR	**0.53**	**0.63**	**0.74**	**0.82**	**0.88**	**0.92**	**0.75**
	eKF-WF	0.44	0.56	0.70	0.81	**0.88**	0.91	0.72
	eKF-OMLSA	0.49	0.59	0.71	0.81	0.87	0.91	0.73

postfiltering have a similar performance, while OMLSA shows slightly worse results, especially at low SNRs. Thus, eKF-WF seems the preferred strategy for close-talk conditions on average.

Regarding far-talk conditions, likewise, the proposed postfilters obtain the best results in terms of perceptual quality, with Wiener filtering being the best strategy for noise reduction, especially at high SNRs. The SPPC method outperforms MVDR beamforming with no postfiltering, but it does not achieve any improvements compared to eKF-WF and eKF-OMLSA. Moreover, SPPC introduces more speech distortion, yielding a poor performance in terms of speech intelligibility. MVDR beamforming with no postfiltering achieves the best STOI scores, as in CT conditions. On the other hand, the postfiltering approaches obtain similar results on average, although their performance is worse at low SNRs. The comparison of both postfilters indicates that eKF-OMLSA achieves better intelligibility on average and, especially, at low SNRs. To sum up, both eKF-WF and eKF-OMLSA perform similarly, with Wiener filtering achieving best perceptual quality and OMLSA better speech intelligibility in FT conditions.

5. Conclusions

In this paper we have proposed a postfiltering approach to our RTF extended Kalman filter framework for dual-microphone smartphones. Our proposals make use of the more accurate estimated RTFs and noise statistics in order to obtain the gain function for noise reduction. We evaluated two different postfilters based on Wiener filtering and the OMLSA estimator, and also proposed different clean speech PSD estimators for CT and FT conditions in order to compute the needed statistics. The proposed approaches were evaluated in terms of perceptual quality and speech intelligibility when they are used for enhancing noisy speech signals from a dual-microphone smartphone in adverse acoustic environments. Our results show improvements in terms of both perceptual quality and noise reduction of the enhanced signal while low speech distortion is introduced in comparison to a standalone MVDR beamformer. As future work, we will extend this study on postfiltering to general multi-microphone devices through our extended Kalman filter approach.

6. References

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Springer, 2008, vol. 1.

[2] K. Kumatani, J. McDonough, and B. Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 127–140, 2012.

[3] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[4] I. Tashev, S. Mihov, T. Gleghorn, and A. Acero, "Sound capture system and spatial filter for small devices," in *Proc. Interspeech*, 2008, pp. 435–438.

[5] I. López-Espejo, A. M. Peinado, A. M. Gomez, and J. A. González, "Dual-channel spectral weighting for robust speech recognition in mobile devices," *Digital Signal Processing*, vol. 75, pp. 13–24, 2018.

[6] M. Parchami, W. P. Zhu, B. Champagne, and E. Plourde, "Recent developments in speech enhancement in the short-time Fourier transform domain," *IEEE Circuits and Systems Magazine*, vol. 16, no. 3, pp. 45–77, 2016.

[7] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. ICASSP*, 1988, pp. 2578–2581.

[8] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 240–259, 1998.

[9] I. A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 709–716, 2003.

[10] S. Lefkimmiatis and P. Maragos, "A generalized estimation approach for linear and nonlinear microphone array post-filters," *Speech Communication*, vol. 49, no. 7-8, pp. 657–666, 2007.

[11] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 6, pp. 561–571, 2004.

[12] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[13] C. Zheng, H. Liu, R. Peng, and X. Li, "A statistical analysis of two-channel post-filter estimators in isotropic noise fields," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 2, pp. 336–342, 2013.

[14] M. Jeub, C. Herglotz, C. Nelke, C. Beaugeant, and P. Vary, "Noise reduction for dual-microphone mobile phones exploiting power level differences," in *Proc. ICASSP*, 2012, pp. 1693–1696.

[15] C. M. Nelke, C. Beaugeant, and P. Vary, “Dual microphone noise PSD estimation for mobile phones in hands-free position exploiting the coherence and speech presence probability,” in *Proc. ICASSP*, 2013, pp. 7279–7283.

[16] W. Jin, M. J. Taghizadeh, K. Chen, and W. Xiao, “Multi-channel noise reduction for hands-free voice communication on mobile phones,” in *Proc. ICASSP*, 2017, pp. 506–510.

[17] J. M. Martín-Doñas, I. López-Espejo, A. M. Gomez, and A. M. Peinado, “An extended Kalman filter for RTF estimation in dual-microphone smartphones,” in *Proc. Eusipco*, 2018, pp. 2488–2492.

[18] T. Esch and P. Vary, “Efficient musical noise suppression for speech enhancement systems,” in *Proc. ICASSP*, 2009, pp. 4409–4412.

[19] J. Yamagishi. (2012) English multi-speaker corpus for CSTR voice cloning toolkit. [Online]. Available: http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html

[20] M. Souden, J. Benesty, S. Affes, and J. Chen, “An integrated solution for online multichannel noise tracking and reduction,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2159–2169, 2011.

[21] “P.862.2: Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codec,” ITU-T Std. P.862.2, 2007.

[22] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.