IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Dual-channel eKF-RTF framework for speech enhancement with DNN-based speech presence estimation

J. M. Martín-Doñas[1], *A. M. Peinado*[2], *I. López-Espejo*[3] *and A. M. Gomez*[2]

[1]Vicomtech Foundation, Basque Research and Technology Alliance (BRTA),
Mikeletegi 57, 20009 Donostia/San Sebastian, Spain
[2]Dept. de Teoría de la Señal, Telemática y Comunicaciones, Universidad de Granada, Spain
[3]Dept. of Electronic Systems, Aalborg University, Denmark
jmmartin@vicomtech.org, {amp,amgg}@ugr.es, ivl@es.aau.dk

Abstract

This paper presents a dual-channel speech enhancement framework that effectively integrates deep neural network (DNN) mask estimators. Our framework follows a beamforming-plus-postfiltering approach intended for noise reduction on dual-microphone smartphones. An extended Kalman filter is used for the estimation of the relative acoustic channel between microphones, while the noise estimation is performed using a speech presence probability estimator. We propose the use of a DNN estimator to improve the prediction of the speech presence probabilities without making any assumption about the statistics of the signals. We evaluate and compare different dual-channel features to improve the accuracy of this estimator, including the power and phase difference between the speech signals at the two microphones. The proposed integrated scheme is evaluated in different reverberant and noisy environments when the smartphone is used in both close- and far-talk positions. The experimental results show that our approach achieves significant improvements in terms of speech quality, intelligibility, and distortion when compared to other approaches based only on statistical signal processing.

Index Terms: Dual-microphone smartphone, beamforming, extended Kalman filter, speech presence probability, deep neural network

1. Introduction

Speech-related services are ubiquitously available thanks to mobile devices such as smartphones. These devices are frequently used in reverberant and noisy environments, both in close-talk (CT) conditions (i.e., the smartphone is placed at the ear of the user) and far-talk (FT) conditions (i.e., the user holds the device at a distance from her/his face). This makes speech enhancement algorithms particularly necessary to improve speech quality and intelligibility on these challenging scenarios.

Current smartphones often embed several microphones. Particularly, a widely used layout consists of a primary microphone at the bottom and a secondary one at the top or back of the device. While beamforming techniques (i.e., spatial filtering) [1] can be used in these devices, the reduced number of microphones and their location limit the speech enhancement performance [2]. In these circumstances, postfiltering techniques can be incorporated to these devices [3, 4] to improve the noise reduction. Alternatively, other approaches employ single-channel filters exploiting dual-channel information. For example, the power level difference between channels was exploited in [5, 6] for noise estimation and reduction in CT conditions. On the other hand, in [7, 8] the coherence properties of the noise field were considered to estimate the noise statistics, needed by a Wiener filter, in FT conditions. In addition to speech enhancement, the dual-channel information has been exploited for other related speech processing tasks from a classical signal processing perspective, as feature enhancement in automatic speech recognition (ASR) systems [9, 10] and noise estimation [11]. Finally, the use of deep neural networks (DNNs) has also been explored on dual-microphone smartphones. For example, in [12, 13] a DNN-based feature enhancement approach was investigated in the context of noise-robust ASR for smartphones. On the other hand, in [14] we proposed a dual-channel DNN-based speech enhancement algorithm based on spectral mapping. Recently, this idea was evaluated in [15] for spectral masking using phase-sensitive masks and dual-channel features.

In previous works [16, 17, 18], we proposed a dual-channel speech enhancement framework, intended for smartphones, based on a beamforming-plus-postfiltering scheme. The main contribution of our approach was the estimation of the acoustic response between microphones using an extended Kalman filter (eKF) framework, which allows us to track these acoustic channels in reverberant environments. Moreover, noise estimation was performed using a speech presence probability (SPP)-based approach to update the noise statistics when speech was absent. This SPP estimation was carried out using statistical spatial models with a priori SPP information obtained from dual-channel information. On the contrary, in this work we propose the integration of DNN-based mask estimators [19, 20, 21, 22] for this task. The DNN, which is fed with dual-channel features based on power and phase differences, aims to improve the accuracy of the prediction. Our proposal is then evaluated on a dual-microphone smartphone under several noisy acoustic environments in CT and FT conditions. The results show that our approach achieves improvements in terms of speech quality, intelligibility, and distortion in comparison with other state-of-the-art approaches for dual-channel smartphones.

The remainder of this paper is organized as follows. In Section 2, we briefly review our dual-channel eKF-based framework for smartphones. Section 3 describes our DNN-based

This work has been supported by the Spanish Ministry of Science and Innovation Project No. PID2019-104206GB-I00/AEI/10.13039/501100011033 and the Spanish Ministry of Universities through the National Program FPU (grant reference FPU15/04161).

10.21437/IberSPEECH.2021-7

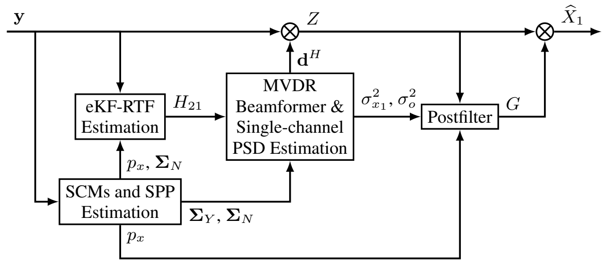


Figure 1: *Overview of the dual-channel speech enhancement algorithm for dual-microphone smartphones.*

mask estimator for SPP prediction and the dual-channel features evaluated. Then, in Section 4, the experimental framework and results are presented and analyzed. Finally, conclusions are summarized in Section 5.

2. Dual-channel speech enhancement based on an eKF-RTF framework

Let us consider the following multichannel observation model, in the short-time Fourier transform (STFT) domain, for the noisy speech signal acquired by a dual-microphone smartphone,

$$\mathbf{y}(t,f) = \mathbf{h}(t,f)X_1(t,f) + \mathbf{n}(t,f), \tag{1}$$

where $\mathbf{y}(t,f) = \begin{bmatrix} Y_1(t,f) & Y_2(t,f) \end{bmatrix}^\top$ and $\mathbf{n}(t,f) = \begin{bmatrix} N_1(t,f) & N_2(t,f) \end{bmatrix}^\top$ are the noisy speech and noise multichannel vectors (subscripts identify the primary and secondary microphones in the array), respectively, $X_1(t,f)$ is the clean speech signal at the primary microphone, $\mathbf{h}(t,f) = \begin{bmatrix} 1 & H_{21}(t,f) \end{bmatrix}^\top$ is the relative transfer function (RTF) vector, and t and f are the time frame and frequency indices, respectively. From now on, when possible, we will omit indices t and f for the sake of simplicity.

Our goal is to estimate the clean speech signal at the reference microphone, X_1, from the noisy speech observations. To do this, we apply our extended Kalman filter (eKF) dual-channel framework [18]. A diagram of the algorithm pipeline is depicted in Figure 1. The noisy speech signal is processed using a beamforming algorithm for noise reduction, as in $Z = \mathbf{d}^H\mathbf{y}$, where $\{\cdot\}^H$ stands for the Hermitian transpose operator. We use the minimum-variance distortionless response (MVDR) beamformer [1],

$$\mathbf{d} = \frac{\mathbf{\Sigma}_N^{-1}\mathbf{h}}{\mathbf{h}^H\mathbf{\Sigma}_N^{-1}\mathbf{h}}, \tag{2}$$

where $\mathbf{\Sigma}_N = E\left\{\mathbf{n}\mathbf{n}^H\right\}$ is the noise spatial covariance matrix (SCM), with $E\{\cdot\}$ representing the expectation operator over a random variable. At the beamformer output, signal Z represents the clean speech signal X_1 plus a residual noise with a power spectral density (PSD) given by $\sigma_o^2 = \left(\mathbf{h}^H\mathbf{\Sigma}_N^{-1}\mathbf{h}\right)^{-1}$.

As can be seen, MVDR needs estimates for H_{21} and $\mathbf{\Sigma}_N$. For RTF estimation, the already proposed eKF-RTF algorithm is applied [16, 18]. We first define $\mathbf{H}_{21}$ and $\mathbf{Y}_2$ as vectors that stack the real and imaginary components of H_{21} and Y_2, respectively. Then, the RTF vector is estimated using the following recursion,

$$\widehat{\mathbf{H}}_{21}(t) = \widehat{\mathbf{H}}_{21}(t-1) + \mathbf{K}(t)\left(\mathbf{Y}_2(t) - \boldsymbol{\mu}_Y(t)\right), \tag{3}$$

where $\mathbf{K}$ is the Kalman gain matrix and $\boldsymbol{\mu}_Y = E\{\mathbf{Y}_2\}$ is the expectation over the noisy speech at the secondary microphone. A detailed derivation of these terms can be found in [18]. For the noise statistics, we use a recursive estimator based on the speech presence probability (SPP) [23],

$$\widehat{\mathbf{\Sigma}}_N(t) = \alpha(t)\widehat{\mathbf{\Sigma}}_N(t-1) + (1-\alpha(t))\,\mathbf{y}(t)\mathbf{y}^H(t), \tag{4}$$

where $\alpha = \widetilde{\alpha} + (1-\widetilde{\alpha})\,p_x$ is an updating parameter that depends on the a posteriori SPP p_x, which ranges from 0 to 1, and $\widetilde{\alpha} = 0.9$ is a constant factor. Thus, the noise SCM is updated with the current noisy observation when speech is absent, while the previous value is kept when speech is present.

The speech signal Z is further processed using a postfilter which provides $\widehat{X}_1 = GZ$, where G is a single-channel gain function. In our proposal, we use the optimally-modified log spectral amplitude (OMLSA) estimator [24], which is defined as

$$G = (G_x)^{p_x}(G_n)^{1-p_x}, \tag{5}$$

in which G_n is the speech absence gain, set to -25 dB, and

$$G_x = \frac{\xi}{1+\xi}\exp\left(\frac{1}{2}\int_{\frac{\xi}{1+\xi}\gamma}^{\infty}\frac{e^{-u}}{u}du\right) \tag{6}$$

is the speech presence gain, with $\gamma = |Z|^2/\sigma_o^2$ being the a posteriori signal-to-noise ratio (SNR), $\xi = \sigma_{x_1}^2/\sigma_o^2$ the a priori SNR, and $\sigma_{x_1}^2$ the clean speech PSD at the primary microphone. This last PSD can be obtained using a maximum-likelihood estimator at the beamformer output [25],

$$\widehat{\sigma}_{x_1}^2 = \mathbf{d}^H\left(\widehat{\mathbf{\Sigma}}_Y - \widehat{\mathbf{\Sigma}}_N\right)\mathbf{d}, \tag{7}$$

where

$$\widehat{\mathbf{\Sigma}}_Y(t) = \tilde{\alpha}\widehat{\mathbf{\Sigma}}_Y(t-1) + (1-\tilde{\alpha})\,\mathbf{y}(t)\mathbf{y}^H(t) \tag{8}$$

is a recursive estimator of the noisy speech SCM. Finally, the gain function G is further processed by a musical noise reduction algorithm, as that described in [26], before applying it to the beamformed speech signal Z.

3. DNN-based a posteriori SPP estimation

As can be observed, the a posteriori SPP p_x plays a crucial role in our eKF-RTF framework. Not only that it controls the noise SCM estimation, but also the postfiltering proper performance depends on accurate SPP estimates. Besides, the RTF updating in (3) is only performed when speech presence is detected [18]. In our previous work [18], the a posteriori SPP was obtained using statistical spatial models that combine the use of multivariate Gaussian likelihoods (formulated for the noisy speech and noise signals) with a dual-channel a priori SPP estimator. The main drawback of this method lies on the assumptions made about the statistics of the signals, which can be inappropriate in realistic non-stationary environments.

In this work, we explore the use of DNN-based mask estimators [19] to directly compute the a posteriori SPP. In particular, we consider a convolutional recurrent network (CRN) [15] for the estimation of p_x. A diagram of the applied CRN architecture is depicted in Figure 2. As can be observed, the model comprises an encoder with five convolutional layers, a decoder with five deconvolutional layers, and an intermediate long short-term memory (LSTM) network. We use exponential linear units (ELUs) as non-linear functions in all the convolutional and deconvolutional layers except for the output layer, which uses the sigmoid function. A dropout layer is placed before the input to

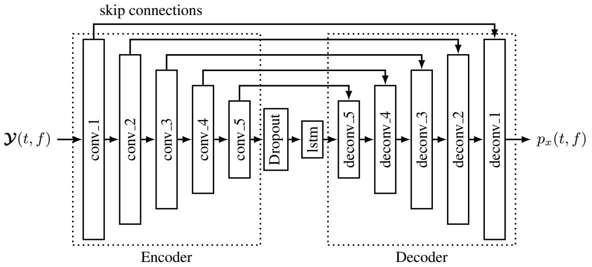


Figure 2: *Diagram of the CRN architecture used for the estimation of speech presence probability masks.*

the LSTM layer to help prevent overfitting. The convolutional and deconvolutional layers operate along the frequency dimension only, while the LSTM layer exploits the temporal dimension. Furthermore, we use skip connections that concatenate the output of each encoder layer to the input of each decoder layer. The CRN is trained using ideal binary masks from the reference channel as target features, using binary cross-entropy as loss function.

As network input features $\boldsymbol{\mathcal{Y}}(t,f)$, a set of different features exploiting spectral or spatial properties was considered in this paper. The main features of this set are the log-magnitude spectrum (LMS) of the primary channel, $\mathcal{Y}_{\text{LMS}}(t,f) = \log|Y_1(t,f)|$. An online normalization is applied to them using a time-recursive mean computation and subtraction at each frequency bin [27]. We also include additional features that make use of the inter-channel properties of the signals. In particular, we consider the spectral relation between the channels by using instantaneous power level difference (PLD) features, which are defined as

$$\mathcal{Y}_{\text{PLD}}(t,f) = \frac{|Y_1(t,f)|^2 - |Y_2(t,f)|^2}{|Y_1(t,f)|^2 + |Y_2(t,f)|^2}. \tag{9}$$

In addition, spatial properties of the signals are exploited by using inter-channel phase difference (IPD) features [28],

$$\boldsymbol{\mathcal{Y}}_{\text{IPD}}(t,f) = \begin{bmatrix} \cos\left(\theta_{y_1}(t,f) - \theta_{y_2}(t,f)\right) \\ \sin\left(\theta_{y_1}(t,f) - \theta_{y_2}(t,f)\right) \end{bmatrix}, \tag{10}$$

where θ_{y_1} and θ_{y_2} are the phases of the noisy speech signals at the reference and secondary microphones, respectively.

4. Experimental results

The TIMIT-2C-CT/FT database [18] was used to evaluate the proposed dual-channel algorithm. This database includes simulated dual-channel noisy speech recordings at 16 kHz acquired with a dual-microphone smartphone in both CT and FT conditions. Each condition (CT or FT) comprises three sets, i.e., training, validation, and test, with a total of 4800, 1200, and 2400 noisy speech samples, respectively. To simulate the noisy samples, clean speech signals from the TIMIT database [29, 30] are convolved with dual-channel acoustic responses obtained from a smartphone at different reverberant scenarios [16]. Then, the reverberated speech signals are mixed with noises at six SNRs from -5 dB to 20 dB (5 dB steps). For the training and validation sets, four reverberant and noisy environments are considered: car, bus, babble, and mall. For the test set, apart from the previous noise types, four additional environments are also evaluated: street, pedestrian street, bus station, and cafe. In addition, the reverberation level of the acoustic responses depends on the noisy environment.

Table 1: *Architecture of the CRN mask estimator. The feature size is indicated in the form feature maps × frames × frequency channels, being N_{in} the number of input features. Hyperparameters refer to kernel size, stride and output channels. For the LSTM layer, the number of hidden units is indicated.*

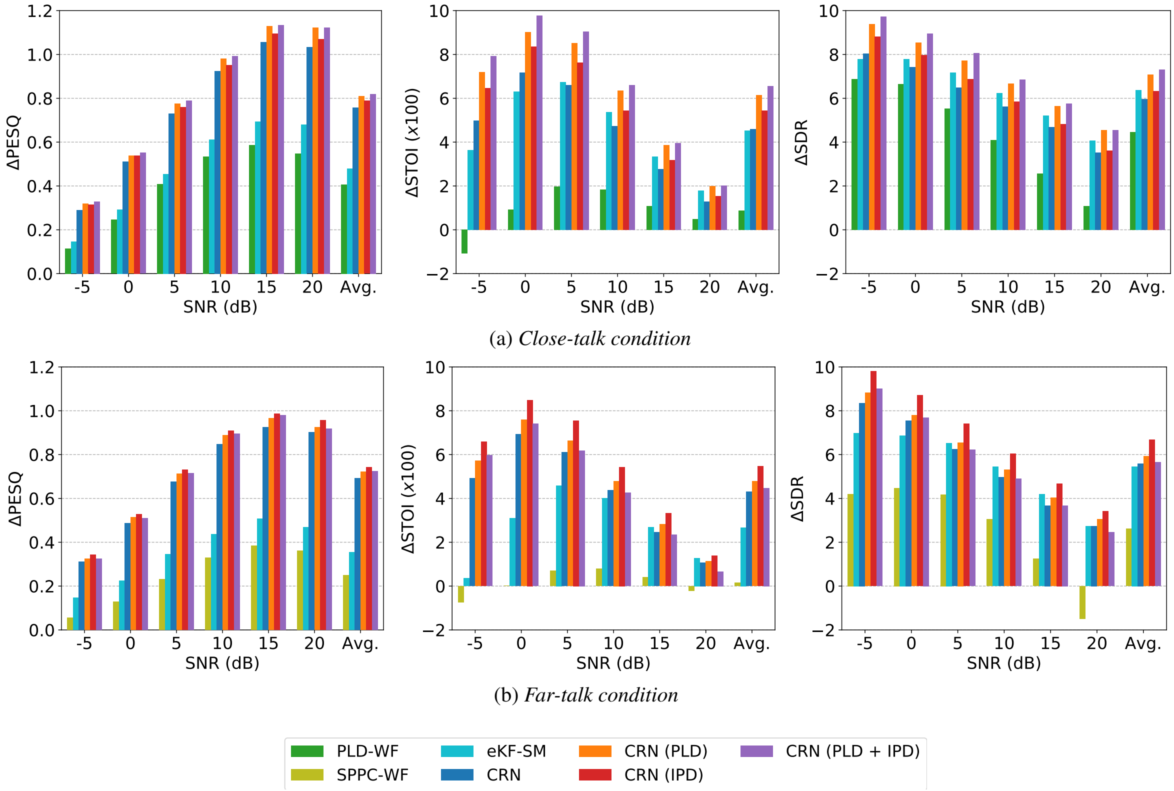
Layer name	Input size	Hyperparameters	Output size
conv_1	N_{in} × T × 257	1 × 3, (1, 2), 8	8 × T × 128
conv_2	8 × T × 128	1 × 3, (1, 2), 8	8 × T × 64
conv_3	8 × T × 64	1 × 3, (1, 2), 16	16 × T × 32
conv_4	16 × T × 32	1 × 3, (1, 2), 32	32 × T × 16
conv_5	32 × T × 16	1 × 3, (1, 2), 64	64 × T × 8
reshape_1	64 × T × 8	-	T × 512
lstm	T × 512	512	T × 512
reshape_2	T × 512	-	64 × T × 8
deconv_5	128 × T × 8	1 × 3, (1, 2), 32	32 × T × 16
deconv_4	64 × T × 16	1 × 3, (1, 2), 16	16 × T × 32
deconv_3	32 × T × 32	1 × 3, (1, 2), 8	8 × T × 64
deconv_2	16 × T × 64	1 × 3, (1, 2), 8	8 × T × 128
deconv_1	16 × T × 128	1 × 3, (1, 2), 1	1 × T × 257

Table 2: *Objective metric results for the noisy speech signals of the test set in the TIMIT-2C-CT/FT database, broken down by SNR and device use mode (CT or FT).*

Metric	Mode	SNR (dB)						Avg.
		-5	0	5	10	15	20	
PESQ	CT	1.09	1.11	1.23	1.45	1.81	2.27	1.49
	FT	1.07	1.11	1.25	1.50	1.88	2.38	1.53
STOI	CT	0.51	0.63	0.74	0.84	0.91	0.95	0.76
	FT	0.50	0.61	0.73	0.83	0.90	0.95	0.75
SDR	CT	-5.80	-0.81	4.19	9.15	14.02	18.70	6.58
	FT	-5.79	-0.80	4.19	9.15	14.03	18.70	6.58

For STFT computation, a 512-point DFT was applied using a 32 ms square-root Hann window with 50% overlap. The eKF-RTF framework implemented is the same as in [18]. The CRN network architecture used in our experiments is concisely described in Table 1. The ADAM optimizer [31] was used to train the DNN model. We used a batch size of 10 utterances, which were zero-padded to have the same number of frames. The dropout rate was set to 0.5 deactivation probability. Besides, the early-stopping procedure [32] was applied with a patience of 20 epochs.

The enhanced signal provided by our proposal is evaluated in terms of the following objective quality metrics: perceptual evaluation of the speech quality (PESQ) [33], short-time objective intelligibility (STOI) [34] and scale-invariant signal-to-distortion ratio (SDR) [35]. As a reference, Table 2 shows the results obtained in terms of these metrics when evaluating the noisy speech signals from the test set without any enhancement algorithm. For the CRN-based mask estimator, different combinations of input features were tested: using only LMS features (CRN), jointly integrating either PLD features (PLD) or IPD features (IPD), and fully integrating all the features (PLD+IPD). For comparison purposes, we also evaluated our framework with SPP estimation based on statistical models (eKF-SM) [18], and two single-channel Wiener filters relying on dual-channel information: the PLD-based filter (PLD-WF) [5] for the CT condition, and the SPP- and coherence-based filter (SPPC-WF) [7] for the FT condition. The results achieved by the tested methods (improvements obtained over the noisy speech results) are shown in Figure 3.



(a) *Close-talk condition*

(b) *Far-talk condition*

Figure 3: *PESQ, STOI and SDR differential results from the evaluation of the CRN-based SPP mask estimator with the different input features. The OMLSA postfilter with MVDR-based PSD estimation, and state-of-the art single-channel filters for smartphones, are also shown for comparison purposes. The plots show the increments obtained on the metrics with respect to noisy speech (see Table 2).*

In CT conditions, and according to Figure 3a, the CRN approach outperforms PLD-WF and eKF-SM, especially in terms of PESQ. Moreover, the CRN estimator benefits from the use of dual-channel features. PLD+IPD obtains the best results in terms of all the considered metrics. Between the dual-channel features considered, the CRN estimator mainly benefits from the PLD features, achieving similar results to those from PLD+IPD. This shows that the power difference between microphones can be a good indicator of speech presence in CT conditions. Although the CRN with IPD features improves with respect to the CRN approach, PLD+IPD slightly improves the variant including PLD features in STOI and SDR. Therefore, PLD features poses a good trade-off between performance and network complexity in CT conditions.

In FT conditions (Figure 3b), the CRN approach also achieves better results than the other evaluated methods. In this case, the variant including IPD features stands as the best choice, especially in terms of STOI and SDR. On the other hand, the utilization of PLD features slightly improves the CRN approach. Unlike in the CT scenario, PLD features seem not to provide enough information in FT conditions, as they tend to zero. Then, the phase difference is the main information source that allows to distinguish between speech and noise components. Finally, the combination of both types of dual-channel features does not yield improvements in comparison with using standalone features, either PLD or IPD ones. This can be explained by our CRN estimator not being able to deal with multiple input features in this challenging scenario. In particular, the use of additional PLD features may mislead the network, as they do not provide accurate information in this case. Thus, the IPD features stand as the best alternative.

5. Conclusions

In this paper, we have proposed a DNN-based SPP estimator that is integrated into our dual-channel eKF-RTF framework for speech enhancement on smartphones. Our approach allows for a more accurate prediction of the SPP probability thanks to the modeling capabilities of the DNN models and the use of dual-channel information. We use a convolutional recurrent neural network to exploit the spectral, spatial, and temporal properties of the speech signal. Two different dual-channel features were considered and tested: the instantaneous power level difference and the phase difference between channels. The proposed integrated scheme was compared with the same framework but using statistical spatial models for SPP prediction, as well as with other dual-channel speech enhancement algorithms from the state-of-the-art. The results show that the DNN-based mask estimator outperforms the rest of the evaluated approaches in terms of objective quality and intelligibility metrics. Among the considered spatial features, the PLD features show better performance in CT conditions, while the IPD features are more useful in FT conditions.

6. References

[1] 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.

[2] 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.

[3] E. Habets, S. Gannot, and I. Cohen, "Dual-microphone speech dereverberation in a noisy environment," in *Proc. IEEE International Symposium on Signal Processing and Information Technology*, 2006, pp. 651–655.

[4] 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.

[5] 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.

[6] V. B. Truong, D. M. Nguyen, and Q. H. Dang, "An MC-SPP approach for noise reduction in dual microphone case with power level difference," in *Proc. International Conference on Advanced Technologies for Communications*, 2014, pp. 292–297.

[7] 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.

[8] 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.

[9] I. López-Espejo, A. M. Gomez, J. A. González, and A. M. Peinado, "Feature enhancement for robust speech recognition on smartphones with dual-microphone," in *Proc. EUSIPCO*, 2014, pp. 21–25.

[10] 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.

[11] I. López-Espejo, J. M. Martín-Doñas, A. M. Gomez, and A. M. Peinado, "Unscented transform-based dual-channel noise estimation: Application to speech enhancement on smartphones," in *Proc. IEEE Telecommunications and Signal Processing*, 2018, pp. 88–91.

[12] I. López-Espejo, J. A. González, Á. M. Gómez, and A. M. Peinado, "A deep neural network approach for missing-data mask estimation on dual-microphone smartphones: Application to noise-robust speech recognition," in *Advances in Speech and Language Technologies for Iberian Languages*, 2014, pp. 119–128.

[13] I. López-Espejo, A. M. Peinado, A. M. Gomez, and J. M. Martín-Doñas, "Deep neural network-based noise estimation for robust ASR in dual-microphone smartphones," in *Proc. IberSpeech*, 2016, pp. 117–127.

[14] J. M. Martín-Doñas, A. M. Gomez, I. López-Espejo, and A. M. Peinado, "Dual-channel DNN-based speech enhancement for smartphones," in *Proc. IEEE Workshop on Multimedia Signal Processing (MMSP)*, 2017, pp. 1–6.

[15] K. Tan, X. Zhang, and D. Wang, "Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios," in *Proc. ICASSP*, 2019, pp. 5751–5755.

[16] 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.

[17] J. M. Martín-Doñas, I. López-Espejo, A. M. Gomez, and A. M. Peinado, "A postfiltering approach for dual-microphone smartphones," in *Proc. IberSpeech*, 2018, pp. 142–146.

[18] J. M. Martín-Doñas, A. M. Peinado, I. López-Espejo, and A. Gomez, "Dual-channel speech enhancement based on extended Kalman filter relative transfer function estimation," *Applied Sciences*, vol. 9, no. 12, p. 2520, 2019.

[19] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. ICASSP*, 2016, pp. 196–200.

[20] ——, "A generic neural acoustic beamforming architecture for robust multi-channel speech processing," *Computer Speech and Language*, vol. 46, pp. 374–385, 2017.

[21] Y. Liu, A. Ganguly, K. Kamath, and T. Kristjansson, "Neural network based time-frequency masking and steering vector estimation for two-channel MVDR beamforming," in *Proc. ICASSP*, 2018, pp. 6717–6721.

[22] S. Chakrabarty and E. Habets, "Time-frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 787–799, 2019.

[23] 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.

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

[25] A. Kuklasinski, S. Doclo, S. Jensen, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 9, pp. 1595–1608, 2016.

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

[27] J. Heitkaemper, J. Heymann, and R. Haeb-Umbach, "Smoothing along frequency in online neural network supported acoustic beamforming," in *Speech Communication; 13th ITG-Symposium*, 2018.

[28] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. ICASSP*, 2018, pp. 1–5.

[29] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, vol. 107, p. 16, 1988.

[30] L. Lamel, R. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Proc. of the DARPA Speech Recognition Workshop*, 1989, pp. 2161–2170.

[31] D. P. Kingma and J. L. Ba, "ADAM: A method for stochastic optimization," in *Proc. of 3rd International Conference on Learning Representations*, 2015, pp. 1–13.

[32] L. Prechelt, "Early Stopping - But When?" in *Neural Networks: Tricks of the Trade*. Springer Berlin Heidelberg, 2012, pp. 53–67.

[33] "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.

[34] 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.

[35] J. Roux, S. Wisdom, H. Erdogan, and J. Hershey, "SDR - Half-baked or Well Done?" in *Proc. ICASSP*, 2019, pp. 626–630.