REAL-TIME SPEECH ENHANCEMENT USING AN EFFICIENT CONVOLUTIONAL RECURRENT NETWORK FOR DUAL-MICROPHONE MOBILE PHONES IN CLOSE-TALK SCENARIOS

Ke Tan[1]*, Xueliang Zhang*[2]*, and DeLiang Wang*[1,3]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Department of Computer Science, Inner Mongolia University, China
[3]Center for Cognitive and Brain Sciences, The Ohio State University, USA
tan.650@osu.edu, cszxl@imu.edu.cn, wang.77@osu.edu

ABSTRACT

In mobile speech communication, the quality and intelligibility of the received speech can be severely degraded by background noise if the far-end talker is in an adverse acoustic environment. Therefore, speech enhancement algorithms are typically integrated into mobile phones to remove background noise. In this paper, we propose a novel deep learning based framework for real-time speech enhancement on dual-microphone mobile phones in a close-talk scenario. It incorporates a convolutional recurrent network (CRN) with high computational efficiency. In addition, the framework amounts to a causal system, which is necessary for real-time processing on mobile phones. We find that the proposed approach consistently outperforms a deep neural network (DNN) based method, as well as two traditional methods for speech enhancement.

Index Terms— convolutional recurrent network, close-talk scenario, dual-microphone mobile phone, real-time speech enhancement

1. INTRODUCTION

Mobile speech communication has become an increasingly important application as mobile phones are extensively used. The quality and intelligibility of the received speech can be severely degraded by background noise if the far-end talker is in an adverse acoustic environment. In order to attenuate background noise from noisy speech signals, speech enhancement algorithms have been integrated into most mobile phones. Typically, a small microphone array including from two to four microphones is deployed in mobile phones to yield better enhancement performance. In a typical dual-microphone configuration, a primary microphone is placed on the bottom of mobile phones, and a secondary microphone is deployed on the top. In this study, we focus on speech enhancement for such commonly-used dual-microphone mobile phones in a close-talk scenario, where a speech signal is picked up with small distance between the primary microphone and the human mouth.

In the last decade, various algorithms have been developed for dual-microphone speech enhancement. Yousefian *et al.* [1] designed an approach that uses the dissimilarity between the power of received signals in the two channels, *i.e.* the inter-microphone power level difference (PLD), as a criterion for noise reduction. Their results show that the approach improves speech quality. In [2], the inter-microphone PLD is utilized to estimate the power spectral density of the noise, which amounts to a spectral gain function. Subsequently, the spectral gain function is applied to the noisy spectrum to derive the enhanced spectrum. In this method, it is assumed that the diffuse noise field is homogeneous and the PLD between the clean speech signals picked up by the two microphones is sufficiently large. Another approach is to use the power level ratio (PLR) of the signal received by the primary microphone to that by the secondary microphone [3]. Based on the PLR, a spectral gain function is calculated using the sigmoid function. The experimental results show that the algorithm yields comparable performance with the methods in [1] and [2], while it is more computationally efficient. More recently, Fu *et al.* [4] utilized the inter-microphone posteriori signal-to-noise ratio difference to estimate the speech presence probability (SPP). The estimated SPP is subsequently used to derive a noise correlation matrix estimator, which yields a multichannel minimum variance distortionless response (MVDR) filter. Other related studies include [5], [6] and [7].

In recent years, speech enhancement has been formulated as supervised learning, inspired by the concept of time-frequency (T-F) masking in computational auditory scene analysis (CASA) [8]. Many supervised speech enhancement algorithms have been developed, in which the discriminative patterns within speech or noise signals are learned from training data. In 2013, Wang and Wang [9] first introduced DNNs to address supervised speech enhancement. For dual-microphone mobile phones, López-Espejo *et al.* [10] employed a DNN to perform noise reduction, where the DNN is trained to predict a binary mask from the log-Mel features of the noisy signals picked up by the two microphones. The estimated mask is used for spectral reconstruction by the truncated-Gaussian based imputation algorithm. In a more recent study [11], they trained a DNN to map from the log-Mel features of noisy speech to those of background noise. The estimated log-Mel features of the noise, along with the noisy signal at the primary channel, are used to obtain the log-Mel features of enhanced speech through a vector Taylor series feature compensation method. Subsequently, the enhanced log-Mel features are transformed into the cepstral domain, prior to being fed into a speech recognizer for evaluation. Their results show that the DNN-based approach significantly outperforms several representative traditional algorithms in terms of word accuracy.

Motivated by our recent study [12] on CRNs, we propose a novel framework for dual-microphone speech enhancement on mobile phones, where a CRN is employed to predict the phase sensitive mask (PSM) [13] [14]. The proposed model leads to a causal system, which is necessary for real-time processing. Moreover, the CRN is computationally efficient, and thus is amenable to mobile phone ap-

This work was conducted when Ke Tan was doing an internship at Elevoc Technology Co., Ltd., Shenzhen, Guangdong, China.

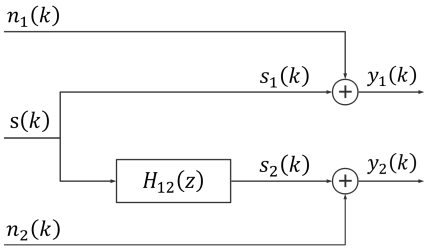

Fig. 1. Illustration of the dual-channel signal model.

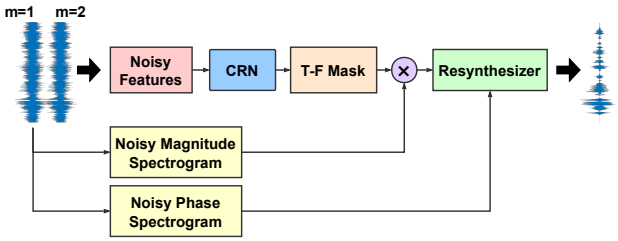

Fig. 2. Overview of the dual-channel speech enhancement system.

plications. In order to resynthesize the time-domain waveforms, we propose to use the phase of noisy signal difference between the primary channel and the secondary channel. In addition, we propose to use both intra-channel and inter-channel features as the CRN input. We find that the proposed approach substantially outperforms a DNN-based method that is similar to [10], as well as two traditional methods for speech enhancement.

The rest of this paper is organized as follows. We provide a detailed description of our proposed approach in Section 2. The experimental setup and results are presented in Section 3. Section 4 concludes this paper.

2. ALGORITHM DESCRIPTION

2.1. Problem formulation

Let $y_m(k)$, $s_m(k)$ and $n_m(k)$ denote noisy speech, clean speech and background noise, respectively, where m is the channel index. Specifically, $m = 1$ refers to the primary channel, and $m = 2$ the secondary channel. We model the signals as in [2]:

$$y_1(k) = s_1(k) + n_1(k) = s(k) + n_1(k) \quad (1)$$

$$y_2(k) = s_2(k) + n_2(k) = s(k) * h_{12}(k) + n_2(k) \quad (2)$$

where $s(k)$ denotes target clean speech, and '*' the convolution operation. The acoustic transfer function of clean speech between the two channels is represented by $H_{12}(z)$, and the corresponding impulse response by $h_{12}(k)$. In a close-talk scenario, where the distance between the primary microphone and the human mouth is small, we regard clean speech at the primary channel as target clean speech, *i.e.* $s_1(k) = s(k)$. Fig. 1 depicts the dual-channel signal model.

In this study, we treat the dual-microphone enhancement as supervised learning, as shown in Fig. 2. We first extract features from the noisy signals picked up by the two microphones, which are subsequently fed into a CRN to predict a T-F mask. The estimated T-F mask is applied to the magnitude spectrogram of noisy speech at the primary channel. The resulting enhanced magnitude spectrogram is then combined with the noisy phase to resynthesize the time-domain waveform of enhanced speech.

2.2. Intra-channel and inter-channel features

We assume that all signals are sampled at 16 kHz. A 20-ms Hamming window is utilized to segment a signal into a set of time frames, with a 50% overlap between adjacent frames. A straightforward idea is to use the noisy magnitude spectrograms at the two channels, *i.e.* $|Y_1|$ and $|Y_2|$, as input features, where Y_1 and Y_2 are 161-dimensional spectra corresponding to a 320-point short-time Fourier transform (STFT). The inter-microphone correlations, however, may not be sufficiently leveraged. In particular, the inter-microphone correlations between the phase spectra are not utilized. To alleviate this problem, we propose to additionally include the magnitude spectrum of the noisy signal difference, as well as that of the noisy signal summation, *i.e.* $|Y_1 - Y_2|$ and $|Y_1 + Y_2|$. These inter-channel features, $|Y_1 - Y_2|$ and $|Y_1 + Y_2|$, implicitly incorporate phase correlations between channels. In other words, the intra-channel features (*i.e.* $|Y_1|$ and $|Y_2|$) and the inter-channel features (*i.e.* $|Y_1 - Y_2|$ and $|Y_1 + Y_2|$) are concatenated, and are treated as four different input channels of the CRN. We find that the inclusion of inter-channel features significantly improves objective intelligibility and perceptual quality.

2.3. Training target and waveform resynthesis

In this study, we use the PSM [13] [14] as the training target, which incorporates the phase information. It is typically defined on the noisy speech spectrum and the clean speech spectrum at the primary channel as follows:

$$\begin{aligned} PSM(t,f) &= Re\Big\{\frac{|S_1(t,f)|e^{j\theta_{s_1}}}{|Y_1(t,f)|e^{j\theta_{y_1}}}\Big\} \\ &= \frac{|S_1(t,f)|}{|Y_1(t,f)|}\cos(\theta_{s_1} - \theta_{y_1}) \end{aligned} \quad (3)$$

where $|S_1(t,f)|$ and $|Y_1(t,f)|$ denote spectral magnitudes of clean speech and noisy speech within a T-F unit at time frame t and frequency channel f, respectively, and θ_{s_1} and θ_{y_1} the phases of clean speech and noisy speech within the unit, respectively. $Re\{\cdot\}$ computes the real component.

Once the PSM is estimated, we apply it to the magnitude spectrogram of the noisy speech at the primary channel. Typically, the enhanced magnitude spectrogram is combined with the phase spectrogram of the corresponding noisy speech, as shown in Fig. 2. Based on the analysis of the acoustical environment in [2], we assume that the PLD between the clean speech signals at the two channels is larger than that between the noise signals. In this case, the noisy signal difference between channels, *i.e.* $y_1 - y_2$, may have a higher signal-to-noise ratio (SNR) than y_1, and thus have a cleaner phase. In our experiments, we find that using the phase of $y_1 - y_2$ to resynthesize waveforms improves both objective intelligibility and perceptual quality over using the phase of y_1. Note that the PSM should be redefined in this case:

$$\begin{aligned} PSM(t,f) &= Re\Big\{\frac{|S_1(t,f)|e^{j\theta_{s_1}}}{|Y_1(t,f)|e^{j\theta_{y_1-y_2}}}\Big\} \\ &= \frac{|S_1(t,f)|}{|Y_1(t,f)|}\cos(\theta_{s_1} - \theta_{y_1-y_2}) \end{aligned} \quad (4)$$

where $\theta_{y_1-y_2}$ represents the phase of $y_1 - y_2$. For convenience, we refer to *PSM-1* as the PSM defined in Eq. 3, and *PSM-2* as the PSM defined in Eq. 4.

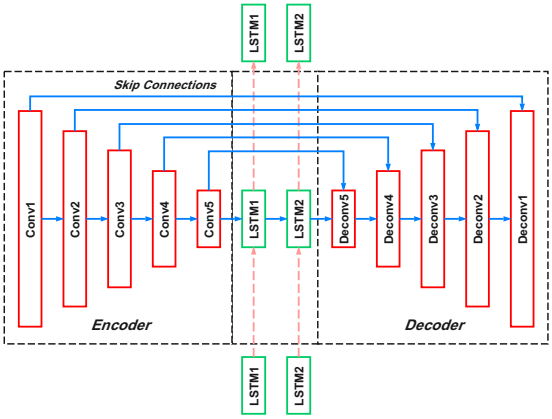

Fig. 3. Illustration of the CRN architecture.

Table 1. Architecture of our proposed CRN. Here T denotes the number of time frames in the spectrogram.

layer name	input size	hyperparameters	output size
conv2d_1	$4 \times T \times 161$	$1 \times 3, (1, 2), 8$	$8 \times T \times 80$
conv2d_2	$8 \times T \times 80$	$1 \times 3, (1, 2), 8$	$8 \times T \times 39$
conv2d_3	$8 \times T \times 39$	$1 \times 3, (1, 2), 16$	$16 \times T \times 19$
conv2d_4	$16 \times T \times 19$	$1 \times 3, (1, 2), 16$	$16 \times T \times 9$
conv2d_5	$16 \times T \times 9$	$1 \times 3, (1, 2), 16$	$16 \times T \times 4$
reshape_1	$16 \times T \times 4$	-	$T \times 64$
lstm_1	$T \times 64$	64	$T \times 64$
lstm_2	$T \times 64$	64	$T \times 64$
reshape_2	$T \times 64$	-	$32 \times T \times 4$
deconv2d_5	$32 \times T \times 4$	$1 \times 3, (1, 2), 16$	$16 \times T \times 9$
deconv2d_4	$32 \times T \times 9$	$1 \times 3, (1, 2), 16$	$16 \times T \times 19$
deconv2d_3	$32 \times T \times 19$	$1 \times 3, (1, 2), 8$	$8 \times T \times 39$
deconv2d_2	$16 \times T \times 39$	$1 \times 3, (1, 2), 8$	$8 \times T \times 80$
deconv2d_1	$16 \times T \times 80$	$1 \times 3, (1, 2), 1$	$1 \times T \times 161$

2.4. Convolutional recurrent network

In [12], we have recently designed a convolutional recurrent network, which combines convolutional layers and recurrent layers. It benefits from the feature extraction capability of convolutional neural networks (CNNs) and the temporal modeling capability of recurrent neural networks (RNNs). With an encoder-decoder architecture, the CRN encodes the input features into a higher-dimensional latent space, and then models the sequence of latent feature vectors via two long short-term memory (LSTM) layers. The output sequence of the LSTM layers is subsequently converted back to the original input size by the decoder. Specifically, the encoder comprises five convolutional layers, and the decoder five deconvolutional layers. To improve the flow of information and gradients throughout the network, skip connections are used to concatenate the output of each encoder layer to the input of the corresponding decoder layer. In the CRN, all convolutions and deconvolutions are causal, so that no future information is used for mask estimation at each time frame. Fig. 3 illustrates the CRN in [12].

For mobile phone applications, high computational efficiency is required for real-time processing with low latency. In addition, a small memory footprint is desired. In order to achieve a computationally efficient model, we prune the CRN in [12] simply by reducing the number of kernels. Additionally, unlike the 2×3 (*time* $\times$ *frequency*) kernels in [12], we use a kernel size of 1×3, without degrading the performance.

Table 1 provides a detailed description of our proposed CRN architecture. The input size and the output size of each layer are specified in *featureMaps* $\times$ *timeSteps* $\times$ *frequencyChannels* format, and the layer hyperparameters in (*kernelSize*, *strides*, *outChannels*) format. Note that the number of feature maps in each decoder layer is doubled by the skip connections. We employ exponential linear units (ELUs) [15] in all convolutional and deconvolutional layers except the output layer. In the output layer, we use the sigmoid nonlinearity for mask estimation. In this study, the PSM is clipped to between 0 and 1, to fit the range of the sigmoid function.

3. EXPERIMENTS

3.1. Experimental setup

3.1.1. Data preparation

In our experiments, we use the WSJ0 SI-84 training set [16] which includes 7138 utterances from 83 speakers (42 males and 41 females). Of these speakers, we set aside 6 speakers (3 males and 3 females) as untrained speakers for test. In other words, we train the models with the 77 remaining speakers. We consider the target clean speech the same as the clean speech signal picked up by the primary microphone, *i.e.* s_1. The clean speech at the secondary microphone, s_2, is generated by the acoustic path h_{12} from the primary channel to the secondary channel. We model the inter-channel acoustic path h_{12} as a time-invariant finite impulse response (FIR) filter, whose coefficients are trained by minimizing the mean squared error (MSE), *i.e.* $\mathrm{E}[e^2(k)]$, where

$$e(k) = s_2^{(tr)}(k) - \sum_{l=0}^{p} \hat{h}_{12}(l) s_1^{(tr)}(k-l). \tag{5}$$

Here p denotes the order of the FIR filter, and $s_1^{(tr)}$ and $s_2^{(tr)}$ the clean speech signals recorded by a dual-microphone mobile phone that is mounted on a dummy head in an anechoic environment. In our experiments, we use six different mobile phones, which amount to six different inter-channel acoustic paths. Of them, we randomly select one for the test set, and use the other five for the training set. Note that the distance between the two microphones is about 10 cm.

In order to simulate stereo noise signals n_1 and n_2, we consider two different noise fields: quasi-diffuse noise and point noise. We follow the approach in [17] to generate the quasi-diffuse noise. Specifically, we place equally strong noise sources in a reverberant room at the height of the primary microphone with azimuths between 0° and 360° spaced by 10°. Hence, the noise signals at the two microphones are generated by convolving these noise sources with a binaural room impulse response (BRIR) that is simulated through the image method [18]. Analogously, we simulate the point noise by place a noise source at an azimuth that is randomly sampled between 0° and 360° spaced by 10°. We assume that the mobile phone is placed vertically. For training, we use 10,000 noises, which are from a sound effect library (available at https://www.sound-ideas.com), as noise sources. For test, we use two highly nonstationary noises (babble and cafeteria) from an Auditec CD (available at http://www.auditec.com).

Our training set includes 320,000 mixtures, with one half created using quasi-diffuse noise and the other half using point noise. To create a training mixture, we randomly selected a training utterance to generate the clean speech signals at the two microphones. A random cut from the 10,000 training noises is treated as a noise source. The SNR at the primary channel is randomly sampled from -5 to 5 dB with a step of 1 dB. A simulated reverberant room with the size of 10 m $\times$ 7 m $\times$ 3 m is used to generate the BRIRs, where the reverberation time (T_{60}) is randomly drawn from 0.2 s to 0.3 s with

Table 2. Comparisons of different approaches in terms of STOI and PESQ for quasi-diffuse noise. The numbers represent the averages over the two test noises.

metrics	STOI (in %)				PESQ			
SNR	-5 dB	0 dB	5 dB	10 dB	-5 dB	0 dB	5 dB	10 dB
noisy	57.58	69.66	80.71	89.19	1.49	1.77	2.09	2.43
MMSE	52.88	65.45	76.67	85.74	1.48	1.81	2.15	2.45
MS	54.30	67.05	79.05	87.84	1.49	1.83	2.17	2.47
DNN	80.80	87.07	91.81	95.00	2.18	2.54	2.87	3.18
Prop.	**92.52**	**94.95**	**96.66**	**97.88**	**2.89**	**3.20**	**3.48**	**3.70**

Table 3. Comparisons of different approaches in terms of STOI and PESQ for point noise.

metrics	STOI (in %)				PESQ			
SNR	-5 dB	0 dB	5 dB	10 dB	-5 dB	0 dB	5 dB	10 dB
noisy	57.65	69.82	80.87	89.27	1.51	1.77	2.09	2.42
MMSE	53.08	65.47	76.63	85.83	1.50	1.83	2.15	2.45
MS	54.35	67.42	79.29	87.87	1.51	1.83	2.16	2.45
DNN	80.49	87.04	91.82	95.03	2.16	2.53	2.87	3.18
Prop.	**91.81**	**94.68**	**96.54**	**97.83**	**2.85**	**3.17**	**3.45**	**3.68**

a step of 0.02 s. Our test set comprises 150 mixtures created from 25×6 utterances of 6 untrained speakers for each noise. We use four SNRs for the test set, *i.e.* -5, 0, 5 and 10 dB. Moreover, the distance between the primary microphone and a noise source is sampled from 1 m, 1.5 m and 2 m for the training set, while it is set to 1.5 m for the test set. In a close-talk scenario, the direct-to-reverberant ratio (DRR) of the speech signal is high, so that the reverberation from it can be omitted.

3.1.2. Baselines and training details

In our experiments, we compare our proposed method with three other baselines, *i.e.* the MMSE-based noise estimation (MMSE) [19], the minimum statistics (MS) approach [20] and a DNN-based approach that is similar to [10]. Both MMSE and MS are single-channel methods that operate on the primary channel. In the DNN-based approach, we train a three-layer DNN to predict the PSM (*i.e. PSM-1*) from the noisy spectral magnitudes at the two channels (*i.e.* $|Y_1|$ and $|Y_2|$), which has a comparable model size with the proposed CRN. The past three feature frames and the current feature frame are concatenated into a long vector as the DNN input. From the input layer to the output layer, the DNN has (3+1)$\times$161$\times$2, 64, 64, 64 and 161 units, respectively.

We train both the CRN and the DNN with the AMSGrad optimizer [21]. The learning rate is set to 0.001. We use the MSE as the objective function. The minibatch size is set to 16 at the utterance level. Within a minibatch, all training samples are zero-padded to have the same number of time steps as the longest sample.

3.2. Experimental results

3.2.1. Comparisons of different approaches

In our experiments, we use short-time objective intelligibility (STOI) [22] and perceptual evaluation of speech quality (PESQ) [23] as evaluation metrics. Tables 2 and 3 present comprehensive evaluations for different approaches on quasi-diffuse noise and point noise, respectively. The best results in each case are highlighted by boldface. On the two highly nonstationary noises for both noise fields, the two traditional methods yield no improvements in STOI, and relatively small improvements in PESQ over the unprocessed mixtures. By contrast, the deep learning based methods, *i.e.*, the DNN and the CRN, significantly improve both STOI and PESQ metrics. Take, for example the -5 dB SNR case for quasi-diffuse noise. The DNN improves STOI by 23.22% and PESQ by 0.69, while the CRN improves STOI by 34.94% and PESQ by 1.40, over the unprocessed mixtures.

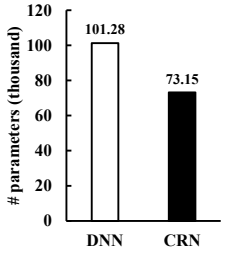

Fig. 4. The number of trainable parameters (unit: thousand).

Table 4. Evaluation of the inter-channels features and the phase of noisy signal difference between channels in terms of STOI and PESQ. The numbers represent the averages over diffuse noise and point noise. See the text for the definitions of (i), (ii), (iii) and (iv).

metrics	STOI (in %)				PESQ			
SNR	-5 dB	0 dB	5 dB	10 dB	-5 dB	0 dB	5 dB	10 dB
noisy	57.62	69.74	80.79	89.23	1.50	1.77	2.09	2.43
(i)	83.67	89.00	93.04	95.79	2.38	2.71	3.02	3.32
(ii)	86.75	91.36	94.65	96.84	2.56	2.88	3.21	2.50
(iii)	88.96	92.44	95.02	96.85	2.65	2.97	3.25	3.50
(iv)	**92.17**	**94.82**	**96.60**	**97.86**	**2.87**	**3.19**	**3.47**	**3.69**

In addition, our proposed CRN-based method consistently outperforms the DNN baseline in both metrics. In the 0 dB SNR case for quasi-diffuse noise, for example, going from the DNN to the CRN yields a 7.88% STOI improvement and a 0.71 PESQ improvement. In addition, the CRN has fewer trainable parameters than the DNN, as shown in Fig. 4.

3.2.2. The effectiveness of the inter-channel features and the phase of noisy signal difference between channels

We evaluate the effectiveness of the inter-channel features and the phase of noisy signal difference between channels for waveform resynthesis. Four cases are considered: (i) intra-channel features + the phase of y_1; (ii) both intra-channel and inter-channel features + the phase of y_1; (iii) intra-channel features + the phase of $y_1 - y_2$; (iv) both intra-channel and inter-channel features + the phase of $y_1 - y_2$. As shown in Table 4, the inclusion of the inter-channel features consistently improves both metrics. Moreover, the phase of $y_1 - y_2$ leads to higher STOI and PESQ scores over the phase of y_1. It can be observed that going from (i) to (iv) improves STOI by 8.5% and PESQ by 0.49 at -5 dB SNR, which reveals that the use of the inter-channel features and the phase of noisy signal difference between channels is advantageous.

4. CONCLUSION

In this study, we have proposed a new deep learning based framework for real-time speech enhancement on dual-microphone mobile phones in a close-talk scenario. The proposed framework incorporates a computationally efficient CRN, which is trained from both intra-channel and inter-channel features. In addition, we propose to use the phase of noisy signal difference between channels to resynthesize the waveform. The experimental results show that the proposed approach consistently outperforms a DNN-based method, as well as two traditional speech enhancement methods.

5. REFERENCES

[1] N. Yousefian, A. Akbari, and M. Rahmani, "Using power level difference for near field dual-microphone speech enhancement," *Applied Acoustics*, vol. 70, no. 11-12, pp. 1412–1421, 2009.

[2] M. Jeub, C. Herglotz, C. Nelke, C. Beaugeant, and P. Vary, "Noise reduction for dual-microphone mobile phones exploiting power level differences," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 1693–1696.

[3] J. Zhang, R. Xia, Z. Fu, J. Li, and Y. Yan, "A fast two-microphone noise reduction algorithm based on power level ratio for mobile phone," in *8th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2012, pp. 206–209.

[4] Z.-H. Fu, F. Fan, and J.-D. Huang, "Dual-microphone noise reduction for mobile phone application," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 7239–7243.

[5] W. Nabi, N. Aloui, and A. Cherif, "Speech enhancement in dual-microphone mobile phones using kalman filter," *Applied Acoustics*, vol. 109, pp. 1–4, 2016.

[6] 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 *41st International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, 2018, pp. 1–5.

[7] Y.-Y. Chen, "Speech enhancement of mobile devices based on the integration of a dual microphone array and a background noise elimination algorithm," *Sensors*, vol. 18, no. 5, pp. 1467, 2018.

[8] D. L. Wang and G. J. Brown, Eds., *Computational auditory scene analysis: Principles, algorithms, and applications*, Wiley-IEEE press, 2006.

[9] Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[10] I. López-Espejo, J. 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*, pp. 119–128. Springer, 2014.

[11] 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 *International Conference on Advances in Speech and Language Technologies for Iberian Languages*. Springer, 2016, pp. 117–127.

[12] K. Tan and D. L. Wang, "A convolutional recurrent neural network for real-time speech enhancement," *Proc. Interspeech*, pp. 3229–3233, 2018.

[13] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 708–712.

[14] Y. Wang and D. L. Wang, "A deep neural network for time-domain signal reconstruction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4390–4394.

[15] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," in *International Conference on Learning Representations*, 2016.

[16] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[17] X. Zhang and D. L. Wang, "Deep learning based binaural speech separation in reverberant environments," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 25, no. 5, pp. 1075–1084, 2017.

[18] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[19] R. C. Hendriks, R. Heusdens, and J. Jensen, "Mmse based noise psd tracking with low complexity," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*. IEEE, 2010, pp. 4266–4269.

[20] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on speech and audio processing*, vol. 9, no. 5, pp. 504–512, 2001.

[21] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," in *International Conference on Learning Representations*, 2018.

[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.

[23] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2001, vol. 2, pp. 749–752.