MULTICHANNEL NOISE REDUCTION USING DILATED MULTICHANNEL U-NET AND PRE-TRAINED SINGLE-CHANNEL NETWORK

Zhi-Wei Tan[1]*, Anh H. T. Nguyen*[2]*, Yuan Liu*[1]*, and Andy W. H. Khong*[1]

[1]Nanyang Technological University, School of Electrical and Electronics Engineering, Singapore
[2]NextG, FPT Software, Vietnam,
E-mail:zhiwei001@e.ntu.edu.sg, anhnht3@fsoft.com.vn, {yuan.liu, andykhong}@ntu.edu.sg

ABSTRACT

Pre-trained single-channel neural networks have become more prevalent for noise reduction in recent years. However, unlike their multichannel counterparts, these monoaural approaches do not exploit spatial information during the optimization process. Furthermore, while multichannel neural networks exploit spatial information, they are optimized for a specific microphone array configuration; extensive data collection and training are required if a new array configuration is deployed. We propose a transfer learning approach that leverages existing pre-trained single-channel neural networks for the optimization of multichannel neural networks. Simulation results on the CHiME-3 dataset show that the proposed method outperforms the state-of-the-art multichannel neural network and neural beamformer.

Index Terms— Multichannel speech enhancement, transfer learning, fine-tuning, deep learning, data scarcity

1. INTRODUCTION

Deep learning approaches have seen success in speech enhancement, where single-channel neural network (NN) approaches have been shown to achieve good performance under noisy conditions [1, 2]. This development has been driven by deep NNs to discriminate and separate speech from noisy signals. With the push in translating research outcomes toward actual deployment, many of the pre-trained networks on large datasets have been made publicly available by the research community [2, 3]. Fine-tuning these pre-trained models facilitates their deployment to new noisy environments without the need for intensive re-training.

Although existing monoaural approaches perform well in various noisy scenarios [1, 2], they do not exploit spatial information to improve speech enhancement performance. On the other hand, recent works on neural beamformers (NBFs) [4–6] exploit spatial information via the direct application of multiple single-channel networks, each of which performs monoaural denoising before estimating the spatial statistics. These estimated spatial statistics are then utilized for the computation of beamformer weights to achieve noise reduction. Therefore, interchannel statistics are not explicitly modeled by the NBFs during the optimization process of the monoaural model. As opposed to NBFs, multichannel NNs fuses spatial information in a non-linear manner [7, 8]. While results presented in [7] suggest that training for a specific array configuration outperforms existing NBFs such as proposed in [5], methods trained for a fixed microphone array configuration require extensive data collection and re-training when deployed for a different array.

In light of the above, we propose a two-stage transfer learning method where the first stage consists of an existing pre-trained single-channel model. The second stage employs a multichannel NN (that is optimized using a reduced dataset). Optimization of this NN is achieved by our proposed dilated U-Net (DUNet) architecture that fuses spatial information with high-frequency resolution. Long skip connections are also employed to achieve faster convergence in a data scarcity setting. The pre-trained single-channel model (from the first stage) is then used to enhance output of the second stage. Simulation results on the CHiME-3 dataset show that the proposed single-to-multichannel transfer learning (SMTL) architecture achieves similar performance with existing multichannel NN [7] using 10% of the training data while outperforming NBF [5] on the same amount of training data.

2. THE PROPOSED SMTL ARCHITECTURE

2.1. Problem formulation

We first define M as the number of microphones in a microphone array, $h_m(n)$, $s(n)$, and $v_m(n)$ as the source-to-microphone impulse response, source signal, and noise signal, respectively. The noisy signal received by the mth microphone is given by

$$\begin{aligned} y_m(n) =& h_m(n) * s(n) + v_m(n) \\ =& x_m(n) + v_m(n), \end{aligned} \tag{1}$$

Zhi-Wei Tan thanks the Nanyang President's Graduate Scholarship scheme for the support.

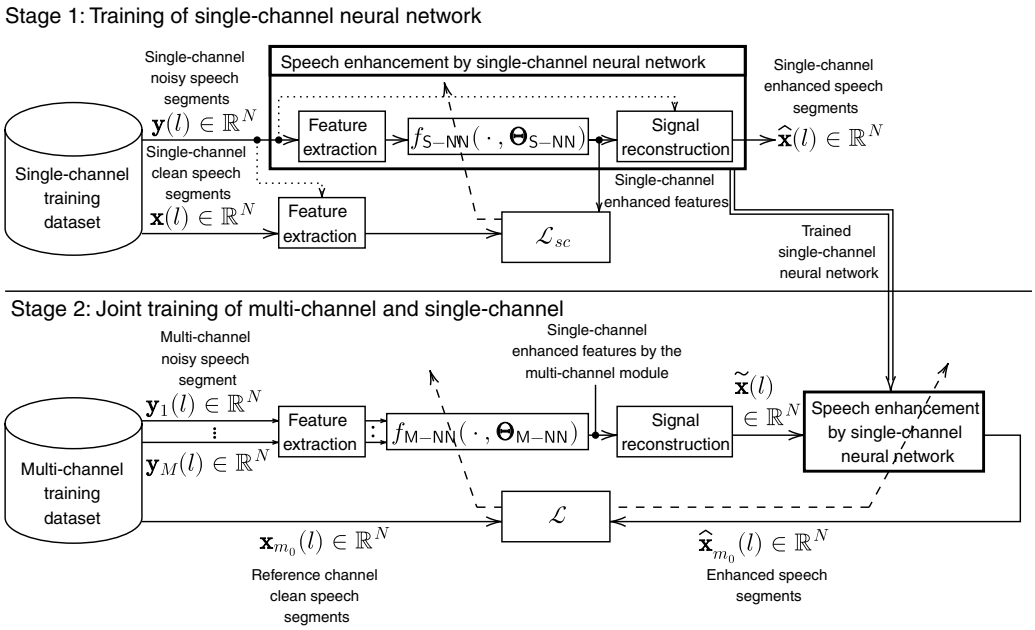

Fig. 1. Transfer learning from single-channel to multichannel approach.

where n is the sample index, $m = 1, \ldots, M$, and $x_m(n) = h_m(n) * s(n)$ with $*$ denoting the convolution operator. Assuming that the impulse response is shorter than the window length of the STFT operation, similar to [5, 9], and that the source is stationary within each utterance, the short-time Fourier transform (STFT) of $y_m(n)$ results in the time-frequency signal

$$\begin{aligned} y_m(t,f) =& h_m(f)s(t,f) + v_m(t,f) \\ =& x_m(t,f) + v_m(t,f), \end{aligned} \tag{2}$$

where t and f are the frame and frequency-bin indices, respectively. Here, we have employed, in our proposed algorithm, a window length of 128 ms (2048 samples at a sampling rate of 16000), which is considerably shorter than a typical acoustic impulse response but sufficiently long to include the direct path component and the early reflections. The objective of speech enhancement is to estimate source signal via

$$\widehat{\mathbf{x}}_{m_0}(l) = [\widehat{x}_{m_0}(l), \ldots, \widehat{x}_{m_0}(l+L)]^T, \tag{3}$$

from the lth segment of the noisy signal

$$\mathbf{y}_{m_0}(l) = [y_{m_0}(l), \ldots, y_{m_0}(l+L)]^T \tag{4}$$

with m_0 denoting the reference microphone index, L as the number of segments, and $(\cdot)^T$ as the transpose operator.

Although pre-trained single-channel NNs can achieve speech enhancement, they do not explicitly model the spatial statistics encapsulated within the (multichannel) noisy signal

$$\mathbf{Y}(l) = [\mathbf{y}_1(l), \ldots, \mathbf{y}_M(l)]^T. \tag{5}$$

2.2. The proposed transfer-learning architecture

We propose a transfer-learning approach comprising two training stages shown in Fig. 1. In the first training stage, either a pre-trained single-channel model [2, 3] is employed, or a new (single-channel) model is being trained, if necessary. We then jointly train a multichannel NN in the second stage with the pre-existing single-channel model. More specifically, we define the following functions $f_{\mathsf{M-FE}}(\cdot)$, $f_{\mathsf{M-NN}}(\cdot)$, and $f_{\mathsf{S-SR}_1}(\cdot)$ as the multichannel feature extraction, its NN, and the single-channel spectral reconstruction, respectively. Defining $\mathbf{\Theta}_{\mathsf{M-NN}}$ as the weights of the multichannel NN, the lth segment of the multichannel input $\mathbf{Y}(l)$ is fed to the multichannel NN via

$$\widetilde{\mathbf{x}}(l) = f_{\mathsf{S-SR}_1}\Big(f_{\mathsf{M-NN}}\big(f_{\mathsf{M-FE}}(\mathbf{Y}(l)), \mathbf{\Theta}_{\mathsf{M-NN}}\big)\Big), \tag{6}$$

where $\widetilde{\mathbf{x}}(l)$ is its single-channel output feature map. Here, we use a single-channel output since a multichannel output is more challenging to optimize [10]. In a similar manner, we define $f_{\mathsf{S-FE}}(\cdot)$, $f_{\mathsf{S-NN}}(\cdot)$, and $f_{\mathsf{S-SR}_2}(\cdot)$ as the pre-trained single-channel feature extraction, its NN, and its single-channel spectral reconstruction. The single-channel denoised signal $\widehat{\mathbf{x}}_{m_0}(l)$ for the lth segment is enhanced via

$$\widehat{\mathbf{x}}_{m_0}(l) = f_{\mathsf{S-SR}_2}\Big(f_{\mathsf{S-NN}}\big(f_{\mathsf{S-FE}}(\widetilde{\mathbf{x}}(l), \mathbf{\Theta}_{\mathsf{S-NN}}\big)\Big), \tag{7}$$

where $\widetilde{\mathbf{x}}(l)$ is the single-channel output of the multichannel NN and $\mathbf{\Theta}_{\mathsf{S-NN}}$ are the weights of the pre-trained single-channel NN.

The learnable parameters $\mathbf{\Theta}_{\mathsf{M-NN}}$ in (6), and the pre-trained parameters $\mathbf{\Theta}_{\mathsf{S-NN}}$ in (7) are jointly optimized by

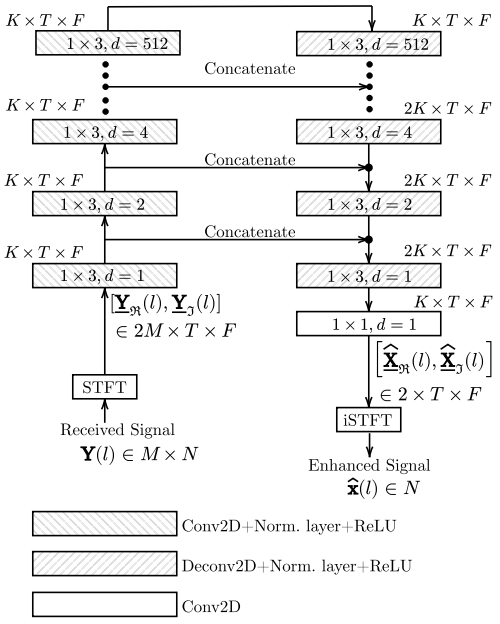

Fig. 2. Architecture of the proposed multichannel dilated U-Net

minimizing, over a multichannel dataset, the negative scale-invariant signal distortion ratio (SI-SDR) loss [11]

$$\mathcal{L} = -\mathsf{SISDR}\big(\widehat{\mathbf{x}}_{\mathsf{m}_0}(l), \mathbf{x}_{m_0}(l)\big). \tag{8}$$

We employ this loss as it accounts for time-shift (or phase differences) in the processed signal [12]. This property results in $\mathcal{L}$ being an effective time-domain loss for multichannel approaches since interchannel phase difference can be optimized. We note that the pre-trained parameters $\boldsymbol{\Theta}_{\mathsf{S-NN}}$ in (7) is jointly optimized with a small learning rate to fine-tune to the reference microphone. Since the pre-trained single-channel can achieve reasonable speech enhancement, minimal learning is required in the multichannel NN and a low amount of training data is therefore required.

2.3. Network architecture

For the multichannel approach in (6), we propose the dilated U-Net (DUNet) inspired from the U-Net encoder-decoder architecture [13]. This new architecture consists of exponentially-dilated convolution [14] and deconvolution layers, as shown in Fig. 2. Each of the ten convolution and ten deconvolution layers has a stride of one. Within each layer, every filter has a size of 1×3 corresponding to the time and frequency dimensions. The proposed DUNet only learns frame-wise mapping since the source is assumed to be stationary within an utterance. Notwithstanding this, the second stage is preferably a network that can model inter-frame dependency.

In DUNet, we employ exponentially-increasing dilation along the frequency dimension [15, 16] for the convolution and deconvolution layers without any downsampling and upsampling layers. This allows subsequent layers to capture more context of the input frequencies, thus preserving high frequency resolution without exponentially increasing the filter size of the convolutions operation [16]. Furthermore, without downsampling, the high frequency resolution that is preserved at each layer is beneficial for beamforming operations. Specifically, the dilation factor d starts with one and doubles after each layer whereas the deconvolution layer starts with 512 but halves after each layer. We utilize long skip connections like in U-Net [13] by concatenating the feature maps of a convolution layer as additional input features to the deconvolution layers. This increases the convergence rate of the network [17] which enables the proposed approach to learn with small amount of training data.

For the multichannel NN, we employ a real-image representation of complex spectrogram as our input to preserve the phase information, i.e., we perform feature extraction via [7, 8]

$$\big[\underline{\mathbf{Y}}_{\Re}(l), \underline{\mathbf{Y}}_{\Im}(l)\big]^T = f_{\mathsf{M-FE}}\big(\mathbf{Y}(l)\big) \in \mathbb{R}^{2M,T,F}, \tag{9}$$

where the real $\underline{\mathbf{Y}}_{\Re}(l)$ and imaginary $\underline{\mathbf{Y}}_{\Im}(l)$ components of the three-dimension noisy signal is concatenated along the microphone dimension.

3. SIMULATION RESULTS

3.1. Dataset, training, and network parameters

We employ the publicly available CHiME-3 challenge dataset [18], which features ($M = 6$) channels speech data that are synthetically mixed with real-world noise from four environments to form the received signal. We employ the fifth channel as the reference channel ($m_0 = 5$) since it has the highest average SNR [19]. During training, eight random segments of each length $L = 10,240$ at a sampling rate of $f_s = 16$ kHz are selected and stacked to form a training mini-batch. For the training dataset, the background noise of each segment is randomly scaled to the SNR of $[-20, 0]$ dB. For testing, the original test set of CHIME-3 is used. In addition, we employ the Adam optimizer [20] for 50 epochs and select the checkpoint with the highest scale-invariant signal distortion ratio (SI-SDR) for testing. Finally, during validation and testing, we set L to be the length of each utterance which, on average, is approximately 3 s long.

For the proposed DU-Net shown in Fig. 2, we set $K = 64$ which results in approximately 350k learnable parameters, and for its STFT parameters, a Hamming window of length 2048 ($F = 1025$) with 75% overlap is used. Since we deploy a mini-batch training scheme, we employ batch normalization layer [21] to stabilize and speed up the training process.

Table 1. Performance on the CHiME-3 test dataset with different percentages of training data (% Trn). Bold-faced values indicate the best performance. Baseline methods with an asterisk (*) are extracted from their corresponding paper, and dash (-) in the result indicates unreported metrics.

Approach	% Trn	PESQ	SDR(dB)	ESTOI
Unprocessed (.CH5)	-	1.15	7.5	0.682
Pre-trained FullSubNet (.CH5) [2]	0	1.62	14.4	0.817
Pre-trained DPTNet (.CH5) [1]	0	2.05	17.1	0.869
NBF* [5]	100	2.29	15.12	-
CA Dense U-Net* [7]	100	**2.44**	18.6	-
DU-Net-FullSubNet (from scratch)	10	1.98	17.4	0.876
DU-Net-FullSubNet (fine-tuning)		2.04	16.7	0.881
DU-Net-DPTNet (from scratch)		2.16	17.1	0.873
DU-Net-DPTNet (fine-tuning)		2.25	18.5	0.898
DU-Net-FullSubNet (from scratch)	100	2.14	18.7	0.900
DU-Net-FullSubNet (fine-tuning)		2.27	18.8	0.907
DU-Net-DPTNet (from scratch)		2.41	18.7	0.907
DU-Net-DPTNet (fine-tuning)		2.41	**19.8**	**0.918**

3.2. Evaluation

We evaluate the proposed framework using the extended short-time objective intelligibility (ESTOI) [22], signal distortion ratio (SDR) [23, 24], and wide-band version of the perceptual evaluation of speech quality (PESQ) as metrics. For the two pre-trained single-channel approaches, we employ the complex-masking FullSubNet[1] [2] trained on the deep noise suppression challenge 2020 dataset [25] and the time-regression dual-path transformer network[2] (DPTNet) [1] trained on the LibriMix dataset [26]. We evaluated these pre-trained models from their checkpoints, and both achieve reasonable speech enhancement, as shown in Table 1. Therefore, we incorporate them in our proposed framework, denoting these two-stage models as DUNet-FullSubnet and DUNet-DPTNet, respectively.

We train the proposed DUNet-DPTNet and DUNet-FullSubNet with two different learning modes: fine-tuning and training from scratch a new model. In the fine-tuning mode, we set the learning rate of the pretrained single-channel models to a small value of 10^{-7} and the multichannel model to 10^{-4}. When training a new model from scratch, parameters in these pre-trained models are re-initialized, and the whole framework is trained with the learning rate of 10^{-4}.

We note from Table 1 that with 10% of training data, the fine-tuned DUNet-DPTNet achieves similar SDR of 18.5 dB to the channel-attention dense U-Net (CA Dense U-Net) [7] of 18.6 dB trained on the full dataset. It also outperforms the SDR of masked-based MVDR beamformer (NBF) [5] by 3.4 dB. This result suggests that the proposed approach can be viable with limited multichannel training data.

[1] https://github.com/haoxiangsnr/FullSubNet/releases/tag/v0.1

[2] https://huggingface.co/JorisCos/DPTNet_Libri1Mix_enhsingle_16k

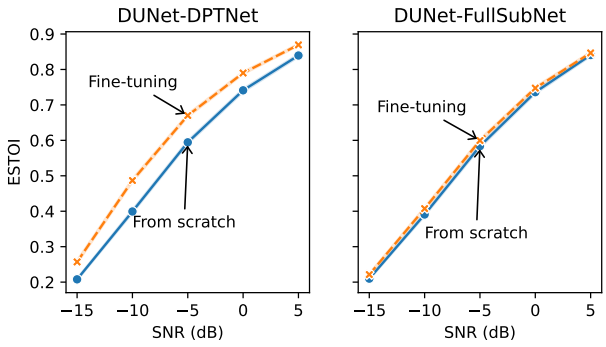

Fig. 3. Test ESTOI performance on the CHiME-3 dataset with 10% training data for different SNRs.

In addition, the fine-tuned DUNet-DPTNet and DUNet-FullSubNet yield better speech enhancement compared to a new model trained from scratch—the fine-tuned DUNet-DPTNet improves SDR by 1.4 dB, PESQ by 0.09, and ESTOI by 2.5% over training a new model from scratch using only 10% training data. Furthermore, as shown in Fig. 3, the proposed SMTL outperforms training a new model from scratch in ESTOI for various SNRs with 10% training data—for $\mathrm{SNR} = -10$ dB, the proposed SMTL improves ESTOI by 8.7% and 1.7% over training a new model from scratch for DUNet-DPTNet and DUNet-FullSubNet, respectively. These results suggest that the proposed fine-tuning approach can leverage the pre-trained single-channel model for a wide range of SNRs to aid its learning with a low amount of training data.

With complete training data as shown in Table 1, the proposed SMTL framework with DU-Net-DPTNet outperforms the SDR of CA Dense U-Net by 1.2 dB. This result suggests that the proposed approach can outperform the state-of-the-art multichannel NN method. We note that performance of the proposed DU-Net may reduce for moving targets as it does not model spatial changes between frames. Features extracted using DOA estimator from acoustic vector sensors [27] or conventional array [28] may aid in such a scenario.

4. CONCLUSION

We propose a transfer learning framework that allows a multichannel NN to optimize jointly with existing single-channel NN. The newly developed DUNet leverages high frequency resolution to fuse spatial information and long skip connections to achieve fast convergence with limited training data. Simulation results on the CHiME-3 dataset show that the proposed SMTL framework only requires a subset of the multichannel data to achieve similar performance with state-of-the-art methods. Furthermore, since the pre-trained single-channel can achieve reasonable speech enhancement, SMTL outperforms a new model trained from scratch.

5. REFERENCES

[1] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," in *Proc. Interspeech*, 2020.

[2] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6633–6637.

[3] M. Pariente *et al.*, "Asteroid: the PyTorch-based audio source separation toolkit for researchers," in *Proc. Interspeech*, 2020.

[4] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *Proc. IEEE Work. Autom. Speech Recognit. Underst.*, 2015, pp. 444–451.

[5] H. Erdogan, J. R. Hershey, S. Watanabe, M. Mandel, and J. L. Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Proc. Interspeech*, 2016, pp. 1981–1985.

[6] Z.-W. Tan, A. H. T. Nguyen, L. T. T. Tran, and A. W. H. Khong, "A joint-loss approach for speech enhancement via single-channel neural network and MVDR beamformer," in *Proc. Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2020, pp. 841–849.

[7] B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnaswamy, "Channel-attention dense u-net for multichannel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 836–840.

[8] Z. Wang and D. Wang, "Multi-microphone complex spectral mapping for speech dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 486–490.

[9] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multi-microphone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, pp. 692–730, 2017.

[10] Z. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1778–1787, 2020.

[11] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?" in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 626–630.

[12] M. Kolbæk, Z.-H. Tan, and J. Jensen, "On loss functions for supervised monaural time-domain speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 825–838, 2020.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015.

[14] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *Proc. Int. Conf. Learn. Represent.*, 2016.

[15] K. Tan, J. Chen, and D. L. Wang, "Gated residual networks with dilated convolutions for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, pp. 189–198, 2019.

[16] Z.-W. Tan, A. H. T. Nguyen, and A. W. H. Khong, "An efficient dilated convolutional neural network for UAV noise reduction at low input SNR," in *Proc. Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2019, pp. 1885–1892.

[17] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Proc. Deep Learn. Data Label. Med. Apps.*, 2016, pp. 179–187.

[18] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Workshop on Autom. Speech Recognit. Underst.*, 2015, pp. 504–511.

[19] ——, "The third 'CHiME' speech separation and recognition challenge: Analysis and outcomes," *Comp. Speech Lang.*, vol. 46, pp. 605 – 626, 2017.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2015, pp. 127–142.

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *Proc. Int. Conf. Mach. Learn.*, pp. 448–456, 2015.

[22] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[23] C. Raffel *et al.*, "mir_eval: A transparent implementation of common MIR metrics," in *Proc. Int. Soc. Music Inf. Retr.*, 2014, pp. 367–372.

[24] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[25] C. K. A. Reddy *et al.*, "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Proc. Interspeech*, 2020.

[26] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "LibriMix: An open-source dataset for generalizable speech separation," in *Proc. Interspeech*, 2020.

[27] K. Wu, V. G. Reju, and A. W. H. Khong, "Multisource doa estimation in a reverberant environment using a single acoustic vector sensor," *IEEE/ACM Trans. Audio, Speech, Lang. Process*, vol. 26, no. 10, pp. 1848–1859, 2018.

[28] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2193–2206, 2013.