END-TO-END COMPLEX-VALUED MULTIDILATED CONVOLUTIONAL NEURAL NETWORK FOR JOINT ACOUSTIC ECHO CANCELLATION AND NOISE SUPPRESSION

Karn N. Watcharasupat⋆, Thi Ngoc Tho Nguyen⋆, Woon-Seng Gan⋆, Shengkui Zhao†, and Bin Ma†

⋆School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore
†Alibaba Group
{karn001, nguyenth003, ewsgan}@ntu.edu.sg, {shengkui.zhao, b.ma}@alibaba-inc.com.

ABSTRACT

Echo and noise suppression is an integral part of a full-duplex communication system. Many recent acoustic echo cancellation (AEC) systems rely on a separate adaptive filtering module for linear echo suppression and a neural module for residual echo suppression. However, in practice, adaptive filtering modules require time to converge and remain susceptible to changes in the acoustic environment. This introduces unnecessary delays to AEC systems using this two-stage framework, despite neural modules already having the capability to suppress both linear and nonlinear echo components. In this paper, we exploit the offset-compensating property of complex time-frequency masks and propose an end-to-end complex-valued neural network architecture. The building block of the proposed model is a pseudocomplex extension of the densely-connected multidilated DenseNet (D3Net), resulting in a very small network of only 354K parameters. The architecture utilized the multi-resolution nature of the D3Net to eliminate the need for pooling, allowing feature extraction using large receptive fields without any loss of output resolution. We also propose a dual-mask technique for joint echo and noise suppression with simultaneous speech enhancement. Evaluation on both synthetic and real test sets demonstrated promising results across multiple energy-based metrics and perceptual proxies.

Index Terms— Acoustic echo cancellation, deep noise suppression, speech enhancement, complex neural network

1. INTRODUCTION

Noises and echoes are well-known sources of degradation for speech quality in a full-duplex communication (FDC) system. In a FDC system, several speech and noise sources can be at play at each end of the communication line, in addition to the echo signal incident on the nearend microphone due to the playback of the farend loopback signal. Consider a realistic scenario with two desired speech sources, one nearend and one farend, with noise sources on each end. The input signal to the nearend microphone is given by

$$p[t] = s[t] + n[t] + f(q)[t], \qquad (1)$$

where t is the time index, $q[t] = r[t] + m[t]$ is the farend microphone input; $f(\cdot)$ is a distortion function responsible for both linear and nonlinear transformations in the communication system; $s[t]$ and $r[t]$ are speech sources from the nearend and farend speakers,

respectively; and, $n[t]$ and $m[t]$ are noise sources from the nearend and farend environments, respectively. For simplicity, we absorb any linear time-invariant transformation of the sound sources due to the environment and the sensors into the signal terms. We denote the complex-valued time-frequency (TF) domain counterpart for each signal by its corresponding boldface uppercase, e.g., $\mathbf{P}[\omega, k]$ for $p[t]$, with ω and k being the frequency and frame indices, respectively.

In traditional acoustic echo cancellation (AEC) that only utilized linear filtering techniques, noise terms are often ignored or absorbed into the speech terms. Hence, the signal models are often viewed as some variants of

$$\mathbf{P}[\omega, k] = \mathbf{S}[\omega, k] + \mathbf{W}[\omega, k]\mathbf{Q}[\omega, k] + g(\mathbf{Q}[\omega, k]), \qquad (2)$$

with the goal of estimating $\mathbf{W}^{-1}[\omega, k]$ in either time or TF domain to remove the linearly transformed component of the farend signal. The term $g(\mathbf{Q}[\omega, k])$ is often called residual echo term, which captures any nonlinear components that are not removable by the linear AEC system. Many recent AEC systems have adopted a two-stage approach with a traditional digital signal processing (DSP) module for linear AEC followed by a neural network module for residual echo suppression (RES) [1–8].

Neural RES is highly related to the task of deep noise suppression (DNS), the latter focusing on noise removal rather than echo cancellation. In a broad sense, the residual echoes can be seen as a noise source, thus a few works have adapted DNS models as the RES module for AEC systems [9, 10]. Given that AEC, RES, and DNS share more similarities than differences, a more unifying formulation naturally follows. Both speech enhancement tasks share a common goal of retrieving the clean speech term s, with the differences mainly lying in the sources of degradation. Moreover, neural networks are often capable of performing both linear and nonlinear filtering, thus diminishing the overall benefits of having a linear DSP preprocessor. In particular, modules such as adaptive filters and delay compensators require a 'warm-up' period for convergence and can be susceptible to changes in the acoustic environments [10], especially in a doubletalk scenario. On the other hand, neural modules, once trained, operate on a fixed set of weights and provide consistent enhancement without the need for an initial convergence period.

With these considerations in mind, we propose an end-to-end joint noise and echo suppression model based on the D3Net building blocks [11]. Inspired by [12], we introduce a pseudocomplex extension of the D3Net, allowing fast computation even on hardware without native complex number supports. Unlike most systems based on convolutional neural networks (CNNs), the proposed architecture exploits the multi-resolution nature of the D3Net building blocks, and operates without any pooling, preserving full feature resolutions throughout the network.

We further proposed a dual-masking technique using two complex-valued output masks to allow the model to simultaneously perform echo/noise suppression and speech enhancement. The complex-valued nature of the masks allows the model to natively compensate delays between the loopback and the microphone signals, eliminating the need for an additional delay compensation module. Due to the purely convolutional nature, the model, with only about 354K parameters, can be easily adapted for causal real-time implementation in future work.

2. PROPOSED SYSTEM

2.1. Complex Densely-Connected Multidilated DenseNet

The densely-connected multidilated DenseNet (D3Net) was formally introduced in [11] as a building block of a U-Net architecture for semantic segmentation and music source separation, both achieving state-of-the-art performances. Variations of the D3Net have also been used successfully with near state-of-the-art performances on sound event localization and detection [13, 14].

Although the U-Net structure has been shown to be very effective in source separation, noise suppression, and speech enhancement, it is notorious for being extremely computationally expensive and can rarely be adapted for real-time implementation. Moreover, while the pooling operations have been useful in non-dilated CNNs to capture features at different scales, the multidilated nature of the D3Net is already capable of capturing these scale-dependent features without the need for pooling, which is prone to poor output resolutions. As a result, the proposed model was designed in a sequential manner to ensure adaptability for real-time implementation, and without pooling to preserve feature resolutions.

In this work, we adapted the D3Net building blocks described in [11] by extending them to the complex-valued domain using the pseudocomplex technique in [12]. For a layer $\mathcal{H}$ with trainable parameters, such as the convolutional layer, a pseudocomplex extension is given by

$$\tilde{\mathcal{H}}(\mathbf{Z}) = \left[\mathcal{H}_{\mathrm{R}}\left(\Re\mathbf{Z}\right) - \mathcal{H}_{\mathrm{R}}\left(\Im\mathbf{Z}\right)\right] + \jmath\left[\mathcal{H}_{\mathrm{I}}\left(\Im\mathbf{Z}\right) + \mathcal{H}_{\mathrm{I}}\left(\Im\mathbf{Z}\right)\right], \quad (3)$$

where $\mathcal{H}_{\mathrm{R}}, \mathcal{H}_{\mathrm{I}}$ are independent layers of identical functions. For a parameterless operation h, such as activations, the pseudocomplex extension $\tilde{h}$ is simply h applied on the real and imaginary parts separately, that is,

$$\tilde{h}(z) = h\left(\Re z\right) + \jmath h\left(\Im z\right). \quad (4)$$

The overall architecture, the complex densely-connected multidilated DenseNet (cD3Net), is shown in Figure 1. We replaced the ReLU activations in the original model with leaky ReLU to ensure the output masks can take any value from the entire complex space. The pseudocomplex extension was applied at the layer level, i.e., all convolutional layers are individually pseudocomplexified. Since the pseudocomplex extensions are real-valued modules under the hood, training and inference of the model can be done efficiently even on hardware without native complex number support.

2.2. Input Representation

The complex-valued TF-domain nearend microphone signal $\mathbf{P}$ and the farend loopback signal $\mathbf{Q}$ are used as the input signals. The input tensor was formed via channel-wise stacking of the TF-domain input signals such that

$$\tilde{\mathbf{X}} = \mathrm{Concat}\left[\mathbf{P}, \mathbf{Q}, \mathbf{P} + \mathbf{Q}, \mathbf{P} - \mathbf{Q}\right] \in \mathbb{C}^{F \times K \times 4}. \quad (5)$$
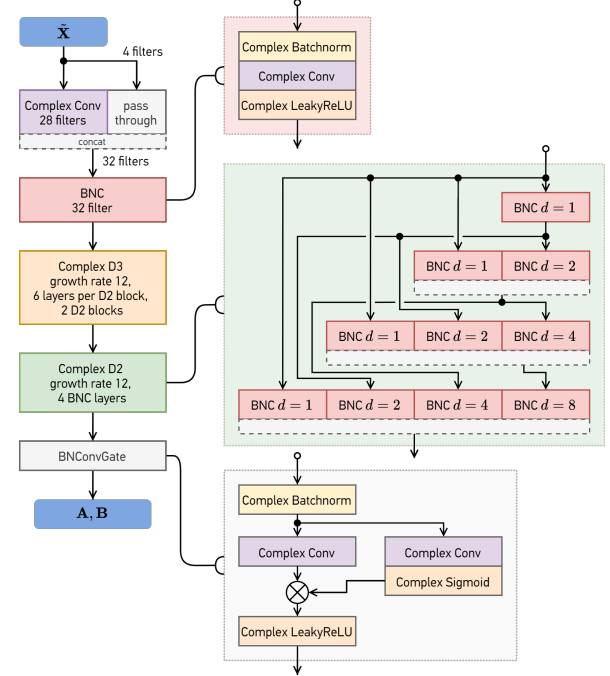

Fig. 1. cD3Net Architecture. BNC: Batch normalization-convolutional block. D2: multidilated DenseNet block. D3: densely-connected D2 block.

The latter two channels were explicitly provided to the model to facilitate the implicit learning of the distortion function f required to generate the echo cancellation masks. The latter two channels also provide temporal cues to the model, especially when there is a non-negligible time offset between the echo and the loopback signals.

2.3. Time-Frequency Masking

Conventionally, the most popular TF masking method used in AEC, DNS, and several other speech enhancement tasks, has been to apply a mask $\mathbf{A} \in \mathbb{C}^{T \times F}$ (or $\mathbb{R}^{T \times F}$) on the nearend microphone signal $\mathbf{P} \in \mathbb{C}^{T \times F}$ to obtain an enhanced speech signal $\widehat{\mathbf{S}} = \mathbf{A} \circ \mathbf{P}$, where $\circ$ is the elementwise multiplication operator. Although this method works very well in a single-track enhancement, this masking technique would discard the rich acoustic information already available in the farend loopback signal for echo suppression.

Since the farend signal is available, we consider the use of two separate masks, $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{T \times F}$, where $\mathbf{B}$ is mainly responsible for echo suppression, and $\mathbf{A}$ for noise suppression and speech enhancement. The dual masking operation is given by

$$\widehat{\mathbf{S}} = \mathbf{A} \circ \left(\mathbf{P} - \mathbf{B} \circ \mathbf{Q}\right). \quad (6)$$

In particular, this dual masking scheme is inspired by the traditional cascade of echo and noise suppressors, but adapted for a unified joint enhancement network. Since $\mathbf{B}$ is complex-valued, when the misalignment between the farend loopback and the nearend microphone signal is not significantly greater than the short-time Fourier transform (STFT) window size, there is a reduced need for time-domain compensation as $\mathbf{B}$ itself can partially account for the delay operation.

3. EXPERIMENTS

3.1. Training

All models in this paper were trained using only the synthetic dataset of the Microsoft AEC Challenges [9, 10]. The synthetic dataset provides four signals for each acoustic scene: clean nearend speech signal, potentially noisy farend loopback signal, potentially noisy echo signal, and potentially noisy nearend microphone signal (a mixture of nearend speech, nearend noise, and echo signal). We used the nearend microphone and the farend loopback signals as inputs. The clean nearend speech signals used as the training targets were first scaled to the same amplitude as the speech content in the nearend microphone signal, using the scaling factors provided in the ground truths. The provided echo signals were not used in this paper. The official train and test set consists of 9.5K and 500 acoustic scenes, respectively. We take another 500 samples out of the train set to form a validation set, leaving 9K acoustic scenes for training.

All models in this paper were trained for 75 epochs each. Each epoch drew a random subset of 2048 scenes from the train set. All models were trained with an effective batch size of 32 using an Adam optimizer [15] with a learning rate of 1×10^{-3}, weight decay of 1×10^{-6}, and learning rate reduction by a factor of 0.9 for every three epochs without validation loss improvement. Since the audio signals were provided at 16 kHz sampling rate, we used a 512-sample (32 ms) square-root Hann window with 50 % overlaps for STFT.

3.2. Loss Function

In order to balance between echo and noise cancellation in an energy-based sense and speech enhancement in a perceptual sense, we consider the use of a weighted loss function between the negative signal-to-distortion ratio loss and the perceptual PMSQE loss [16], given by

$$\mathcal{L}(\hat{s}; s) = \alpha \cdot \text{NegSDR}(\hat{s}; s) + \beta \cdot \text{PMSQE}(\hat{s}; s), \qquad (7)$$

where α, β are weight terms which sum to one.

Unlike most systems utilizing SDR-based losses [17], the negative SDR loss is implemented using the scale-dependent SDR (SD-SDR), instead of the scale-invariant counterpart (SI-SDR)[1], in order to encourage the model to preserve amplitude scaling of the clean speech estimates [18], especially since the amplitude scaling information between the clean speech and the nearend microphone signals is available. For zero-mean estimates and targets, the negative SD-SDR loss is given by

$$\text{NegSDR}(\hat{s}; s) = -10 \log_{10} \left[\frac{\sum_t (\gamma s[t])^2}{\sum_t (\hat{s}[t] - s[t])^2} \right], \qquad (8)$$

where $\gamma = \sum_\tau (\hat{s}[\tau] s[\tau]) / \sum_\tau (s^2[\tau])$.

3.3. Data Augmentation

We experimented with a few data augmentation techniques[2] in the raw signal domain to allow the model to be more robust to realistic signal degradation in a FDC system. For this work, the augmentation

was always applied to the farend loopback signal since not all components of the signals that constitute the nearend microphone signal were available in the dataset. Whenever activated, each augmentation technique was applied with a 50 % probability independently of one another.

Specifically, we experimented with time shifting between the farend loopback signal and the nearend microphone signal to simulate asynchronous signal arrivals, and amplitude scaling to allow the model to be more robust to discrepancy between the amplitudes of the audio content in the loopback signal and the echo signal. A time shift of up to ±512 samples, which is the STFT block size used in this paper, was applied on the farend loopback signal. Amplitude scaling of a factor between 0.5 to 1.5 was applied to the farend loopback signal.

4. RESULTS AND DISCUSSION

4.1. Baseline Models

For comparison, we adapted NSNet [19], Conv-TasNet [20], and DCCRN [12] for AEC. All three models have their first layers modified to support a multichannel input consisting of the nearend microphone and farend loopback signals. NSNet was additionally modified to use the same STFT parameters as the cD3Net. All other settings of DCCRN and Conv-Tasnet followed the defaults provided by Asteroid [21]. All baseline models used the traditional single-mask method and were trained using the same amount of data and epochs as the proposed models.

4.2. Evaluation

We report two sets of results for each model: one on the test split of the synthetic dataset [9] used for the development, and one on the Interspeech 2021 AEC Challenge blind test set [10]. The blind test set consists of 800 real-world recordings with challenging practical conditions, e.g., echo path changes, clock drifts, glitches, noises, gain variations, and onboard DSP.

For the synthetic test set, we evaluate the performance of the models based on four objective metrics commonly used in AEC: scale-invariant source-to-distortion ratio (SI-SDR) [18], standard source-to-distortion ratio (SDR) [22], short-time objective intelligibility (STOI) [23], and perceptual evaluation of speech quality (PESQ) [24].

For the blind real test set, we evaluate the model using objective proxies of mean opinion score (MOS) via the DNSMOS system [25] and degradation MOS via the AECMOS system [10, 26]. The AECMOS provides two degradation MOS (DMOS) scores. The Echo DMOS score captures degradation due to farend echo while the Other DMOS captures degradation due to any other sources. Both DMOS are rated on a 1-to-5 scale, with a higher score reflecting a better audio quality. As per the convention in the ICASSP and Interspeech AEC Challenges [9, 10], we report MOS for the nearend singletalk scenario, Echo DMOS for the farend singletalk scenario, and both Echo and Other DMOS for the doubletalk scenario.

4.3. Discussion

The results for all models evaluated are shown in Table 1. Note that we **do not** perform any preprocessing nor postprocessing on the test set, i.e., no delay compensation or gain adjustment. The networks are fully responsible for all processing required to perform end-to-end echo/noise suppression and speech enhancement with the test set tracks as-is.

---

[1] We have repeated the reported experiments with the SDR and SI-SDR losses replacing the SD-SDR loss; SD-SDR loss performed best across all metrics. Full experimental details were omitted due to space constraints.

[2] We also experimented with additional noise and codec companding, but these techniques did not result in appreciable performance gains.

Model	Params.	α	β	Data Augmentation	Synthetic Test Set				Real Test Set MOS			
					SI-SDR	SDR	STOI	PESQ	NE	FE	DT/E	DT/O
NSNet [10]	3.5M	1	0	–	9.95	10.46	.84	1.45	3.77	3.73	3.99	2.23
Conv-TasNet [20]	5.1M	1	0	–	7.97	9.38	.83	1.30	3.03	3.21	3.93	2.25
DCCRN [12]	3.7M	1	0	–	5.54	5.65	.84	1.33	3.66	2.83	3.60	2.71
cD3Net single mask	352K	1	0	–	13.10	13.60	**.90**	1.77	3.94	4.11	4.16	2.60
cD3Net dual mask	354K	1	0	–	13.22	**13.74**	**.90**	1.80	3.94	4.19	4.12	3.13
		0.75	0.25	–	13.00	13.51	**.90**	1.84	3.94	4.06	4.10	2.40
		0.50	0.50	–	13.21	13.70	**.91**	1.99	3.96	4.28	4.12	3.29
		0.25	0.75	–	**13.26**	**13.74**	**.91**	2.05	3.94	4.26	4.13	3.16
		0	1	–	9.49	10.74	**.90**	1.97	**4.00**	4.37	4.22	3.68
		0.25	0.75	farend shift	13.23	**13.73**	**.91**	2.04	**3.99**	4.42	**4.27**	3.89
				farend scale	13.15	13.67	**.91**	1.98	3.97	4.28	4.14	3.20
				farend shift and scale	13.21	13.70	**.91**	2.06	**3.99**	**4.44**	4.26	**4.00**

Table 1. Evaluation results on the synthetic test set [2] and real test set [10]. *Legends for MOS scores*: NE – singletalk nearend MOS; FE – singletalk farend Echo DMOS; DT/E – doubletalk Echo DMOS; DT/O – doubletalk Other DMOS. Metric values within 0.01 of the respective top scores are considered practically equivalent.

4.3.1. Architectures

In both the synthetic and real test sets, cD3Net models of any config-uration consistently outperformed the three baseline models across most metrics, despite only having about a tenth of the parameters of the smallest baseline model. This demonstrates the ability of cD3Net architecture to learn useful intermediate features in a very parameter-efficient manner. Interestingly, NSNet, which performed the best among the baselines on the synthetic test set for most met-rics, has the simplest architecture amongst the baselines and also does not use pooling. The unexpected poor performance by DC-CRN and Conv-TasNet could be due to the limited number of epochs all models are trained on, as well as inherent shortcomings in their architectures. DCCRN perform frequency downsampling in each layer, while Conv-TasNet uses a basis kernel in lieu of STFT [20]. The small 16-sample time-domain kernel is likely insufficient for learning effective representations for echo cancellation due to the inherent time offsets between the nearend and loopback signals.

4.3.2. Masking techniques

The dual-mask variant of cD3Net shows a slight improvement on the synthetic test set results compared to the single-mask variant. Although the real test set results are comparable for nearend MOS, farend singletalk Echo DMOS, and doubletalk Echo MOS, the dou-bletalk Other DMOS shows a significant improvement with the dual-mask technique. This is somewhat expected by design, as the single-mask technique is unlikely to be able to cope with both echo/noise suppression and speech enhancement in a doubletalk scenario.

4.3.3. Loss function

The models discussed so far were all trained purely using the nega-tive SDR loss. Using a pure PMSQE loss on the dual-mask cD3Net, we found that the SDR and SI-SDR metrics on the synthetic test set degraded considerably, despite the real test set MOS indicating otherwise. Due to the lack of ground truth data on the real test set, it is not possible to exactly analyze the cause of this discrepancy, although we suspect that PMSQE may simply teach the model to optimize for perceived speech quality without regard to the actual amount of echo or noise suppressed.

Regardless, we found that using a weighted composite loss of negative SDR and PMSQE results in a better trade-off between the SDR-based metrics and the perceptual proxies, and even improves STOI and PESQ with the 1:3 and 1:1 SDR-to-PMSQE weightings.

4.3.4. Data Augmentation

Using the 1:3 weighting, which performed best on the synthetic test set, we experimented with data augmentation in the form of time shifting and amplitude scaling on the loopback signal, as described in Section 3.3. It is important to note that the echo and the loopback signals in the synthetic test set are nearly exactly time-aligned. On the other hand, the time offsets between the loopback and the echo signals in the real test set can be much more significant. Since we are using end-to-end models without offset compensation, it is not sur-prising that adding random time shifts to the farend loopback signal considerably improved the model performance on the real test set. The slight drop in the performance on the synthetic test set is likely due to the domain shift introduced by the augmentation.

Although the introduction of random farend amplitude scaling alone worsens the performance across both test sets, when used with random shifting, amplitude scaling further improved the DMOS per-formance in the real test set, achieving the best performances across the MOS metrics.

5. CONCLUSION

In this work, we proposed cD3Net, an end-to-end pooling-free net-work for joint echo cancellation, noise suppression, and speech en-hancement, using a complex-valued extension of the D3Net build-ing block, with a very small parameter count of only 354K. The complex-valued extension eliminates the need for additional linear filtering or preprocessing modules while the multidilated building blocks allow feature extraction using large receptive fields without pooling, preserving full feature resolution throughout the network. To fully utilize the farend loopback signal information, we also intro-duced a time-frequency domain enhancement technique using dual masks for joint AEC, DNS, and speech enhancement. Using a mixed loss training with scale-dependent SDR and PMSQE to account for enhancement in both energy-based and perceptual senses, evaluation on both synthetic and real test sets showed promising results. The proposed model achieved upwards of 13.2 dB for SI-SDR, 13.7 dB for SDR, and upwards of 0.91 for STOI, as well as scored 4 or above for all deep proxies of degradation MOS.

6. REFERENCES

[1] J.-M. Valin, S. Tenneti, K. Helwani, U. Isik, and A. Krishnaswamy, "Low-Complexity, Real-Time Joint Neural Echo Control and Speech Enhancement Based On Percepnet," in *Proc. ICASSP*, 2021, pp. 7133–7137.

[2] R. Peng, L. Cheng, C. Zheng, and X. Li, "ICASSP 2021 Acoustic Echo Cancellation Challenge: Integrated Adaptive Echo Cancellation with Time Alignment and Deep Learning-Based Residual Echo Plus Noise Suppression," in *Proc. ICASSP*, 2021, pp. 146–150.

[3] M. M. Halimeh, T. Haubner, A. Briegleb, A. Schmidt, and W. Kellermann, "Combining Adaptive Filtering And Complex-Valued Deep Postfiltering For Acoustic Echo Cancellation," in *Proc. ICASSP*, 2021, pp. 121–125.

[4] A. Ivry, I. Cohen, and B. Berdugo, "Deep Residual Echo Suppression With A Tunable Tradeoff Between Signal Distortion And Echo Suppression," in *Proc. ICASSP*, 2021, pp. 126–130.

[5] Z. Wang, Y. Na, Z. Liu, B. Tian, and Q. Fu, "Weighted Recursive Least Square Filter and Neural Network Based Residual ECHO Suppression for the AEC-Challenge," in *Proc. ICASSP*, 2021, pp. 141–145.

[6] L. Pfeifenberger, M. Zoehrer, and F. Pernkopf, "Acoustic Echo Cancellation with Cross-Domain Learning," in *Proc. Interspeech*, 2021, pp. 4753–4757.

[7] E. Seidel, J. Franzen, M. Strake, and T. Fingscheidt, "Y²-Net FCRN for Acoustic Echo and Noise Suppression," in *Proc. Interspeech*, 2021, pp. 4763–4767.

[8] R. Peng, L. Cheng, C. Zheng, and X. Li, "Acoustic Echo Cancellation using Deep Complex Neural Network with Nonlinear Magnitude Compression and Phase Information," in *Proc. Interspeech*, 2021, pp. 4768–4772.

[9] K. Sridhar, R. Cutler, A. Saabas, T. Parnamaa, M. Loide, H. Gamper, S. Braun, R. Aichner, and S. Srinivasan, "ICASSP 2021 Acoustic Echo Cancellation Challenge: Datasets, Testing Framework, and Results," in *Proc. ICASSP*, 2021, pp. 151–155.

[10] S. Braun, R. Cutler, A. Saabas, T. Parnamaa, M. Loide, S. Sootla, M. Purin, H. Gamper, K. Sorensen, R. Aichner, and S. Srinivasan, "Interspeech 2021 Acoustic Echo Cancellation Challenge," in *Proc. Interspeech*, 2021, pp. 4748–4752.

[11] N. Takahashi and Y. Mitsufuji, "Densely connected multidilated convolutional networks for dense prediction tasks," in *Proc. CVPR*, 2021.

[12] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Interspeech*, 2020, pp. 2472–2476.

[13] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "ACCDOA: Activity-Coupled Cartesian Direction of Arrival Representation for Sound Event Localization And Detection," in *Proc. ICASSP*, 2021, pp. 915–919.

[14] K. Shimada, N. Takahashi, Y. Koyama, S. Takahashi, E. Tsunoo, M. Takahashi, and Y. Mitsufuji, "Ensemble of ACCDOA- and EINV2-based Systems with D3Nets and Impulse Response Simulation for Sound Event Localization and Detection," Tech. Rep., 2021.

[15] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. ICLR*, 2014.

[16] J. M. Martin-Donas, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, "A deep learning loss function based on the perceptual evaluation of the speech quality," *IEEE Signal Process. Lett.*, vol. 25, no. 11, pp. 1680–1684, 2018.

[17] M. Kolbaek, Z. H. Tan, S. H. Jensen, and J. Jensen, "On Loss Functions for Supervised Monaural Time-Domain Speech Enhancement," *IEEE/ACM Trans. ASLP*, vol. 28, pp. 825–838, 2020.

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

[19] Y. Xia, S. Braun, C. K. Reddy, H. Dubey, R. Cutler, and I. Tashev, "Weighted Speech Distortion Losses for Neural-Network-Based Real-Time Speech Enhancement," in *Proc. ICASSP*, 2020, pp. 871–875.

[20] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Trans. ASLP*, vol. 27, no. 8, pp. 1256–1266, 2019.

[21] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F. R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, "Asteroid: The PyTorch-based audio source separation toolkit for researchers," in *Proc. Interspeech*, 2020, pp. 2637–2641.

[22] E. Vincent, R. Gribonval, and C. Févotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[23] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, 2010, pp. 4214–4217.

[24] 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 *Proc. ICASSP*, 2001, pp. 749–752.

[25] C. K. A. Reddy, E. Beyrami, H. Dubey, V. Gopal, R. Cheng, R. Cutler, S. Matusevych, R. Aichner, A. Aazami, S. Braun, S. Srinivasan, and J. Gehrke, "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Speech Quality and Testing Framework," in *Proc. Interspeech*, 2020.

[26] R. Cutler, B. Nadari, M. Loide, S. Sootla, and A. Saabas, "Crowdsourcing Approach for Subjective Evaluation of Echo Impairment," in *Proc. ICASSP*, 2021, pp. 406–410.