Accepted by IEEE for publication at MLSP 2021.

978-1-7281-6338-3/21/$31.00 ©2021 IEEE

DEEP COMPLEX CONVOLUTIONAL RECURRENT NETWORK FOR MULTI-CHANNEL SPEECH ENHANCEMENT AND DEREVERBERATION

Femke B. Gelderblom and Tor Andre Myrvoll

NTNU & SINTEF, Norway

ABSTRACT

This paper proposes a neural network based system for multi-channel speech enhancement and dereverberation. Speech recorded indoors by a far field microphone, is invariably degraded by noise and reflections. Recent single channel enhancement systems have improved denoising performance, but do not reduce reverberation, which also reduces speech quality and intelligibility. To address this, we propose a deep complex convolution recurrent network (DCCRN) based multi-channel system, with integrated minimum power distortionless response (MPDR) beamformer and weighted prediction error (WPE) preprocessing.

PESQ and STOI performance is evaluated on a test set of room impulse responses and noise samples *recorded* by the same setup. The proposed system shows a statistically significant improvement ($p \ll 0.05$) over competitive systems.

Index Terms— speech enhancement, microphone arrays, deep neural networks, dereverberation, beamforming

1. INTRODUCTION

The field of speech enhancement (SE) has undoubtedly been revolutionized by deep learning techniques. Now that the whole world has been forced to adapt to online meetings at an unseen rate, the topic is also more relevant than ever.

Rapid developments in the related field of automatic speech recognition (ASR) have inspired many source separation and denoising systems. However, over the course of only the past year, Microsoft has organized three SE challenges, where the focus was on enhancing single channel signals specifically for human listeners [1, 2, 3]. Additionally, the challenge ConferencingSpeech 2021 targets multi-channel speech enhancement for video conferencing [4].

Most results of these challenges are not yet available. However, top performing systems of the first deep noise suppression (DNS 2020) challenge, demonstrate remarkable performance increases with respect to removing additive noise from speech recordings.

Isik *et al.* proposed PoCoNet; a 2D UNet (with DenseNet blocks and self-attention) with small kernels [5]. They also utilized a semi-supervised method to increase the amount of training data and investigated the effect of different augmentation techniques. Their proposed system with approximately 50M parameters won first place in the non-real-time track.

Hu *et al.* proposed the deep complex convolution recurrent network (DCCRN) [6]. The DCCRN also follows the UNet structure, but uses complex-valued convolutional encoders and decoders, and LSTMs to model the context dependency. With only 3.7M parameters, the DCCRN models ranked first for the real-time-track and second for the non-real-time track. The lower complexity of this network, combined with the fact that it was trained on less data, while obtaining such competitive performance, makes it an ideal candidate for further research.

However, speech quality and intelligibility is also negatively affected by the presence of reverberance [7, 8]. The DCCRN system does not attempt to remove reverberance at all, and PoCoNet only attempted partial dereverberation.

From the field of multi-channel speech enhancement, we know that there lies a huge potential in relying on multi-channel signals as input, and in applying beamforming techniques [9]. Heymann *et al.* proposed a system where a DNN estimates an ideal binary mask (IBM) to deduce the cross-power spectral densities of the target speech and noise. These are then used for beamforming with a generalized eigenvector (GEV) beamformer [10]. Their system did really well on the CHiMe-3 challenge for robust ASR, but as their network estimates the IBM, and not the target signal, performance is inherently capped. We also observe that, despite its definite merits over earlier single-channel systems, the system proposed by Heymann *et al.* struggles to outperform the single channel DCCRN on our test set, even if we rely on oracle IBM masks (see Table 2 in Section 5). Erdogan *et al.* proposed a similar masked based MVDR system with a spectrum magnitude based loss [11]. However, their final system performance was lower.

In this paper, we therefore propose a far-field multi-channel neural network for simultaneous speech dereverberation and enhancement that combines the recent advancements in single channel speech enhancement for human listeners, with mask based neural beamforming from the domain of multi-channel speech enhancement. We integrated the DCCRN with a minimum power distortionless response (MPDR)

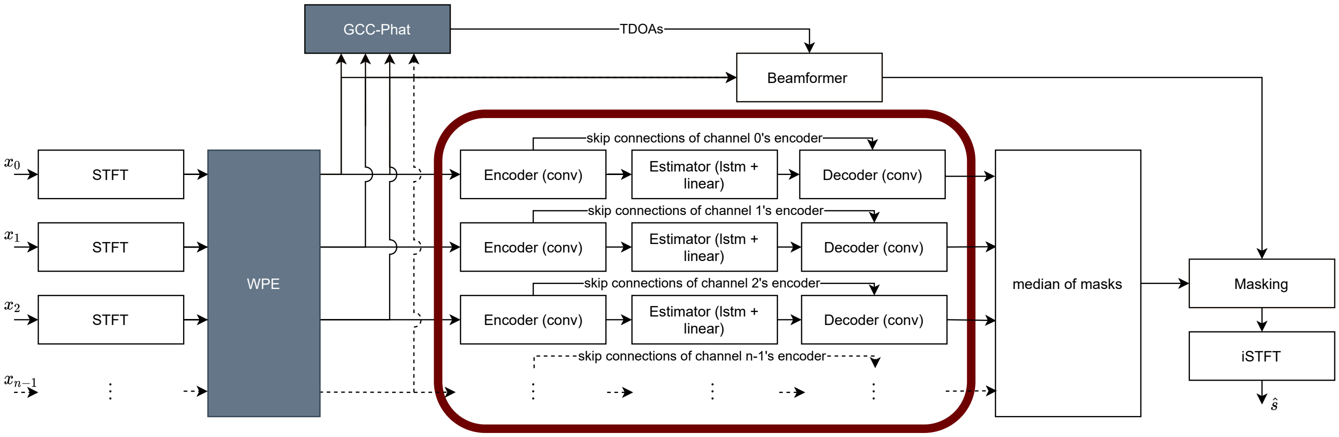
GCC-Phat
TDOAs
Beamformer
x_0
x_1
x_2
x_{n-1}
STFT
STFT
STFT
⋮
WPE
skip connections of channel 0's encoder
Encoder (conv)
Estimator (lstm + linear)
Decoder (conv)
skip connections of channel 1's encoder
Encoder (conv)
Estimator (lstm + linear)
Decoder (conv)
skip connections of channel 2's encoder
Encoder (conv)
Estimator (lstm + linear)
Decoder (conv)
skip connections of channel n-1's encoder
median of masks
Masking
iSTFT
$\hat{s}$

Fig. 1: Overview of the proposed speech enhacement and dereverberation system. The highlighted WPE and GCC-Phat boxes are only employed during inference. The red frame contains all blocks with trainable parameters, where each Encoder-Estimator-Decoder structure represents a single channel DCCRN.

beamformer [12] and added weighted prediction error (WPE) speech dereverberation [13] to the processing pipeline. As such, our system crucially differs from the other mask based beamformering approaches, by relying on a time domain loss, and having complex spectral input features and complex network layers. Furthermore, the usage of the MPDR separates the steering vector estimation from the mask estimation process. The proposed system only requires the corrupted multi-channel speech signal during inference, and as such does not need estimates of noise statistics, or information on microphone layout.

We evaluate the system on two highly realistic test sets. These sets were obtained by combining our own *recorded* multi-channel room impulse responses (RIRs) with clean speech from an open database, and our own multi-channel noise recordings. The latter were recorded by the same array placed at the same location in the same room as where the RIRs were obtained. This setup allows for objective testing with a clean reference signal, while simultaneously avoiding the need for synthetic RIRs that would reduce the realism of the test sets. Furthermore, with this setup we can differentiate between results for speakers looking directly at the array, and the more challenging situation where speakers face the array at a 90° angle.

We compare the system to three state-of-the art baseline systems; i) single channel DCCRN, ii) multi-channel baseline system of the ConferencingSpeech 2021 Challenge, and iii) mask based GEV beamformer with blind analytic normalization postprocessing by Heymann *et al.*

2. THE SYSTEM

2.1. Overview

Figure 1 shows an overview of the proposed system. A multi-channel noisy and reverberant speech signal x is transferred to the frequency domain by a short time Fourier transform (STFT) operation. The resultant signals are fed into the weighted prediction error (WPE) block for deverberation. A DCCRN neural net estimates masks for each channel, where neural network weights are shared across channels (but input is not). All resultant masks are then combined into a single mask using the median operator, because of its resilience to outliers. This mask is then applied to a beamformed result of the dereverberated signal. Lastly, the enhanced signal is taken back to the time domain by an inverse STFT (iSTFT).

The beamformer requires time difference of arrival (TDOA) estimates to obtain an appropriate steering vector. During training, this information is obtained from the known true speaker direction. During the prediction stage, this information is estimated directly from the WPE's output, using generalized cross correlation with phase transform (GCC-PHAT). As such, the final system only requires the corrupted signal as input.

The next subsections provide further processing details.

2.2. Short Time Fourier Transform

Adhering to the original DCCRN paper, we use a Hann window, a FFT length of 512 samples, a window length of 25 ms (400 samples at 16000 Hz) and a hop size of 6.25 ms (100 samples at 16000 Hz) to obtain a complex-valued STFT [6].

2.3. WPE dereverberation

The idea of WPE is to estimate the reverberation tail of the signal and subtract it from the observation with a maximum likelihood approach [13]. We have tested our system with one iteration (using the Nara-WPE implementation [14]), as this has been shown to already provide significant benefit, while multiple iterations quickly become highly time consuming.

2.4. Beamforming

Beamforming is a signal processing technique, where the channels of a multi-channel signal are delayed, weighted, and then combined into a single signal that is steered towards a specific source/direction. Depending on the chosen algorithm, a beamformer can both denoise and dereverberate a multi-channel signal.

One popular beamformer, is the minimum variance distortionless response (MVDR) beamformer. It requires statistical noise characteristics, which are particularly difficult to obtain when the noise is non-stationary as well as mixed with the signal of interest.

One implementation of the MVDR-related algorithm avoids this problem, by deriving the distortionless filter for a specified steering direction that minimizes the mean square output power, and as such only requires the corrupted input signal. Although this implementation is often referred to as an MVDR in the literature, we comply with Van Trees' practice of referring to it as the minimum power distortionless response (MPDR) beamformer for unambiguity [12].

The weights of the MPDR beamformer are obtained as follows:

$$\mathbf{w}_{\mathrm{mpdr}}^{H} = \frac{\mathbf{v}^{H}\mathbf{X}^{-1}}{\mathbf{v}^{H}\mathbf{X}^{-1}\mathbf{v}} \tag{1}$$

where $\mathbf{X}$ is the spectral matrix of the entire input, and $\mathbf{v}$ the steering vector.

When the steering direction is equal to the desired signal direction, the MPDR beamformer reduces to the standard MVDR beamformer [12]. As the target direction is known during training, we effectively train the algorithm with an MVDR beamformer. During inference, the target direction has to be estimated as discussed in Section 2.5.

2.5. GCC-Phat

During inference, one cannot expect the true azimuth of speakers to be available and once the steering vector starts to deviate from the signal vector, the performance between an MPDR and MVDR may differ significantly.

There are many DOA estimation techniques available, both traditional [15, 16], and neural network based [17]. We leave the problem of estimating the azimuth largely outside the scope of this study, but present the results for the final system, both for the ideal situation where the speaker azimuth is known, and for an estimated azimuth using generalized cross correlation with phase transform (GCC-PHAT) [15]. This method allows us to estimate the steering vector without needing to provide the microphone layout.

2.6. DCCRN single channel speech enhancement

The DCCRN single channel SE system was first proposed in [6]. Its goal is to estimate a complex ratio mask (CRM) for the complex-valued STFT. The DCCRN therefore receives both real and imaginary information. This in contrast with SE systems that try to enhance the magnitude of a signal, but rely on the noisy phase.

The DCCRN network can be structured into three parts: the encoder, the estimator and the decoder.

The encoder and decoder contain 6 encoder/decoder blocks each. Each of these blocks consist of a 2D complex convolutional (or deconvolutional) layer, followed by real-valued 2D batch normalization (BN) and leaky ReLU activation. Encoder and decoder blocks (with output channels [32, 64, 128, 128, 256, 256]) are furthermore connected through skip connections.

The encoder extracts high-level features from the input, while the symmetric encoder-decoder architecture ensures that the decoder takes these features (after the estimator stage) back to the same shape as the input. Skip connections between encoder and decoder blocks, make that the noisy input (translated into the corresponding feature spaces), are available during decoding.

At the estimator stage, the network needs to identify the desired signal from the noise, to construct a mask like structure in the encoded feature space. For this, it is important to leverage long-term contexts, which the DCCRN does with LSTM layers. The estimator therefore consist of two real valued LSTM layers (not bidirectional, and each with 256 nodes) followed by a linear layer (1024 nodes). We relied on the polar coordinate masking approach (DCCRN-E).

3. TRAINING

3.1. Setup

We first trained a single channel DCCRN SE model as a pretraining step. This model also functions as one of the reference systems. We then initialize the multi-channel system with the obtained weights.

Both single channel and multi-channel systems were trained with the SI-SNR loss function [6] and the Adam optimizer. While the DCCRN model itself has been kept equal to the original, we made changes to the data synthesis process, updated to the newer 2021 dataset for training, and changed the learning rate; all for improved performance. We used a learning rate of .002, and .0005 during single channel pretraining and multi-channel fine-tuning, respectively.

3.2. Training Data

3.2.1. Single channel pretraining dataset

The DNS Challenge 2021 speech and noise data was used during the pretraining stage, but we relied on the ISM-dir dataset described in [17] for the RIRs. RIRs in this set are simulated with the image source method (ISM) where speaker sources are modelled as directive sources with an average speaker pattern directivity.

For 80% of the time, reverberant speech was obtained from combining clean speech with a random single-channel RIR. For the remaining 20%, speech was left non-reverberant. Noise (always non-reverberant) was then added to obtain the noisy input of SNR within the -5 to 20 dB range. We trained the single channel model using reverberant speech as the target, as training to a clean reference did not improve performance.

3.2.2. *Multi-channel fine-tuning dataset*

For the multi-channel system, also the noise was made multi-channel and reverberant using synthetic RIRs, but here sources were modelled as omnidirectional during simulation. Speech and noise sources were simulated as if from the same room, but at different random locations. The multi-channel system was trained to a clean (non-reverberant) target, by combining above RIRs with the DNS Challenge 2021 speech and noise.

4. EVALUATION

4.1. Testing setup

We test the performance of our system with PESQ, an objective measure of speech quality, and STOI, an objective measure of speech intelligibility. When calulating these objective measures, it is important to compare to the right reference signal. A dereverberating system will *appear* to have worse performance when a reverberant reference is used, as it is 'punished' for dereverberating the input, bringing the enhanced output away from the reference it is tested against. However, the single channel systems from the literature were tested against the reverberant speech signal. Therefore we switch from using a reverberant reference signal for the single-channel system (allowing for fair comparison), to the clean non-reverberant target for the multi-channel system (to take the dereverberation into account).

4.2. Testing Data

4.2.1. *Single channel test set*

To anchor the performance of the single channel SE system to a known test set, we test it with the DNS Challenge 2020 test set.

4.2.2. *Multi-channel test sets*

To create realistic multi-channel test data, RIRs were measured manually with a 9-channel circular array (planar) with 4 cm radius, positioned on a table approximately in the middle of a typical meeting room. See [17] for further details. Two types of RIRs were measured: i) speaker facing towards the array (the 'Easy' set), and ii) the speaker rotated at 90° away from the microphone (the 'Challenging' set).

Obtained RIRs were combined with random speech samples from 'NB Tale', an open Norwegian speech database. None of the training sets contained Norwegian speech.

Additionally, we recorded typical meeting room like noises (see Figure 2) in the same room, using the same array at the same location, as where the RIRs were measured. This means that all recordings also contained more general background noise, like the room's ventilation system.

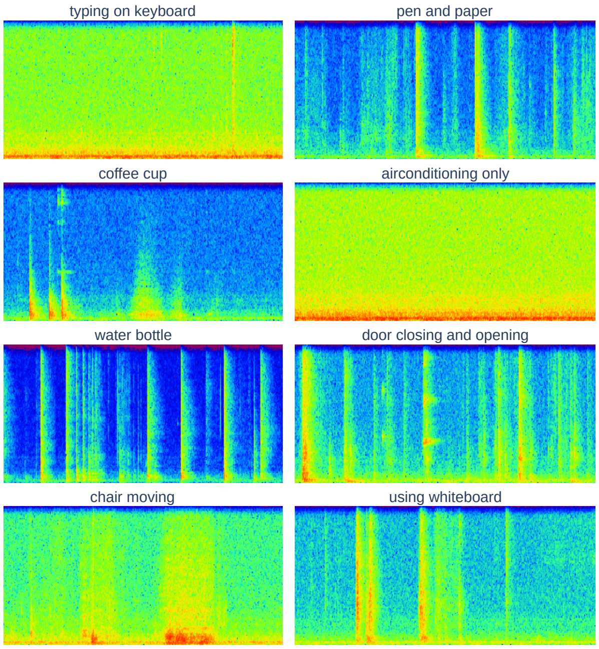

Fig. 2: Sample spectograms of recorded test noises (only the first channel is shown)

The true DOAs were measured with an uncertainty of ±1° at random angles uniformly distributed around the array. As such, it was also possible to test using the oracle steering direction for the beamformer, which normally isn't available during inference.

4.3. Reference systems

We compare results to the performance of three reference systems from the literature, and an alternative to our proposed system:

1. **ConferencingSpeech 2021 baseline**: The multi-channel SE system described in [4], trained with our own multi-channel training set.
2. **Single channel DCCRN**: The pretrained single channel DCCRN model, where we ignore all but the first channel of our test data.
3. **GEV (oracle IBM mask) with BAN**: Mask based GEV beamformer, where the IBM mask is not estimated by a DNN, but obtained directly from the known target/noise signals.

4. **MPDR (oracle TDOAs) + Single channel DCCRN**: Here the MPDR beamformer (suplied with oracle TDOAs) is added as a standalone preprocessing step for the single channel DCCRN.

All of these systems are applied to the noisy signal directly, or to a signal that has first been preprocessed by a standalone WPE block.

5. RESULTS AND DISCUSSION

Table 1 shows the PESQ results for the pretrained DCCRN system. Our single channel system performs on par with the two winning systems, when looking at PESQ scores for the non-reverberant test set. Furthermore, the changes to the training setup give it superior performance on the reverb set, when compared to the original DCCRN-E, and also possibly when compared to PoCoNet, depending on the standard deviation of their test scores (not published). From these results we are confident that our DCCRN acts as a competitive baseline system for our multi-channel results.

Table 1: Narrowband and wideband PESQ results for the DNS Challenge 2020 channel dataset. Reverberant signal used as reference.

	PESQ nb		**PESQ wb**	
	No reverb	**Reverb**	**No reverb**	**Reverb**
Noisy	2.16	2.52	1.58	1.82
PoCoNet [5]	-	-	2.75	2.83[a]
DCCRN-E [6]	3.27	3.08	-	-
Our DCCRN	3.28	3.44	2.76	2.94

[a] Result without partial dereverberation, for unbiased comparison

Table 2 shows the PESQ and STOI results for the multi-channel testsets. Generally speaking, we obtain much lower PESQ scores than those observed in Table 1, despite similar SNRs in both test sets. This is because we are now calculating PESQ with respect to the clean (instead of the reverberant) speech signal.

Independent of the test set used, we see that all enhancement systems benefit from the WPE preprocessing step, even if for STOI scores the difference isn't always significant. This shows that although all systems are trained with reverberant data, they do not learn to deal with it as effectively as WPE.

The independent two-sample t-test shows that all our three systems have statistically significant higher performance than the three reference systems ($p \ll 0.05$). This is true, both when measuring performance in PESQ, or in STOI.

Table 2 furthermore shows that when the speaker is looking at the array ('Easy' set), there is no statistically significant difference in performance, between integrating the MPDR in the training loop, or simply adding it as a preprocessing step to the single channel DCCRN. The same comparison does however find a significant performance difference for the challenging dataset for the SNRs of 5 and 10 dB. Here the alternative to the proposed system (where the MPDR is added as a standalone preprocessing step before the pretrained DCCRN) performs statistically significant worse ($p < 0.05$). This suggest that integrating the MPDR into the training loop, actually allows the enhancement system to learn information that makes it better equipped to deal with a speaker looking in the wrong direction, than the MPDR is capable of on its own, unless there is too much noise.

The performance decrease from moving from *oracle* TDOAs to *estimated* TDOAs is statistically significant for lower SNRs, as expected. At low SNRs, the estimated

Table 2: Wideband PESQ and STOI results for the different multi-channel datasets. Clean signal used as reference. Best scores per SNR are shown in bold, where multiple highlighted values indicate that the difference was not statistically significant.

	WPE	**Easy (looking towards array)**						**Challenging (looking away at a 90° angle)**					
		PESQ wb			**STOI**			**PESQ wb**			**STOI**		
SNR [dB]		**0**	**5**	**10**	**0**	**5**	**10**	**0**	**5**	**10**	**0**	**5**	**10**
No enhancement	No	1.25	1.33	1.39	0.69	0.72	0.74	1.22	1.29	1.35	0.60	0.62	0.63
	Yes	1.33	1.44	1.56	0.72	0.76	0.78	1.27	1.36	1.46	0.18	0.66	0.68
ConferencingSpeech 2021 baseline [4]	No	1.33	1.36	1.48	0.68	0.72	0.73	1.27	1.31	1.41	0.59	0.61	0.62
	Yes	1.40	1.46	1.63	0.71	0.75	0.77	1.33	1.39	1.52	0.63	0.66	0.67
Single channel DCCRN, by Hu *et al.* [6]	No	1.46	1.49	1.51	0.73	0.75	0.75	1.41	1.44	1.46	0.64	0.64	0.65
	Yes	1.64	1.71	1.76	0.77	0.78	0.79	1.55	1.61	1.66	0.68	0.69	0.70
GEV (oracle IBM mask) with BAN, by Heymann *et al.* [10]	No	1.48	1.59	1.60	0.77	0.78	0.79	1.41	1.46	1.52	0.61	0.66	0.67
	Yes	1.58	1.75	1.80	0.78	0.80	0.81	1.49	1.58	1.67	0.68	0.69	0.71
MPDR (oracle TDOAs) + Single channel DCCRN	No	1.68	1.73	1.76	**0.80**	**0.81**	**0.81**	1.54	1.59	1.62	0.71	0.72	0.73
	Yes	**1.89**	**1.98**	**2.04**	**0.81**	**0.82**	**0.83**	**1.71**	1.79	1.85	**0.74**	0.74	0.75
Proposed system (oracle TDOA)	No	1.68	1.86	1.88	**0.80**	**0.82**	**0.83**	1.61	1.73	1.78	**0.75**	**0.76**	**0.77**
	Yes	**1.80**	**2.02**	**2.06**	**0.80**	**0.83**	**0.83**	**1.74**	**1.89**	**1.94**	**0.76**	**0.77**	**0.78**
Proposed system (estimated TDOAs)	No	1.60	1.80	1.85	0.78	**0.81**	**0.82**	1.50	1.62	1.69	0.72	0.73	0.75
	Yes	1.74	**1.95**	**2.04**	**0.79**	**0.82**	**0.83**	1.63	1.79	**1.88**	0.73	**0.75**	**0.77**

TDOAs are more likely to cause the MPDR to point towards the noise, and this even more likely to happen when the speaker is not looking at the array, weakening the direct signal. However, the TDOA estimation method used, leaves a lot of room for improvement to bring the performance closer. As such, it is very promising that the effect size of the performance degradation is this limited. It establishes the MPDR beamformer as a valid candidate for speech enhancement, especially for challenging noise types where the noise statistics are difficult to estimate.

6. CONCLUSION

We proposed a neural network-based system for multichannel speech enhancement and dereverberation, based on WPE dereverberation, the MPDR beamformer and the DCCRN denoiser. The proposed model outperforms state of the art reference systems with respect to speech quality as measured with PESQ and speech intelligibility measured with STOI.

Future work will include improving the estimation of TDOAs by investigating other methods, and exploring opportunities within the system. Furthermore, we plan to evaluate the systems subjectively.

7. REFERENCES

[1] Chandan K A Reddy, Ebrahim Beyrami, Harishchandra Dubey, Vishak Gopal, Roger Cheng, Ross Cutler, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, Sriram Srinivasan, and Johannes Gehrke, "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Speech Quality and Testing Framework," in *INTERSPEECH*, Shanghai, China, 2020, p. 5.

[2] Chandan K. A. Reddy, Harishchandra Dubey, Kazuhito Koishida, Arun Nair, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan, "Interspeech 2021 Deep Noise Suppression Challenge," in *INTERSPEECH*, Brno, Czechia, 2021.

[3] Chandan K. A. Reddy, Harishchandra Dubey, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan, "ICASSP 2021 Deep Noise Suppression Challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, Canada, June 2021, pp. 6623–6627.

[4] Wei Rao, Yihui Fu, Yanxin Hu, Xin Xu, Yvkai Jv, Jiangyu Han, Zhongjie Jiang, Lei Xie, Yannan Wang, Shinji Watanabe, Zheng-Hua Tan, Hui Bu, Tao Yu, and Shidong Shang, "INTERSPEECH 2021 ConferencingSpeech Challenge: Towards Far-field Multi-Channel Speech Enhancement for Video Conferencing," in *INTERSPEECH*, Brno, Czechia, Apr. 2021.

[5] Umut Isik, Ritwik Giri, Neerad Phansalkar, Jean-Marc Valin, Karim Helwani, and Arvindh Krishnaswamy, "PoCoNet: Better Speech Enhancement with Frequency-Positional Embeddings, Semi-Supervised Conversational Data, and Biased Loss," in *Interspeech 2020*, Shanghai, China, Oct. 2020, pp. 2487–2491.

[6] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *INTERSPEECH*, Shanghai, China, 2020, pp. 2472–2476.

[7] Amaro A. de Lima, Sergio L. Netto, Luiz W. P. Biscainho, Fabio P. Freeland, Bruno C. Bispo, Rafael A. de Jesus, Ronald Schafer, Amir Said, Bowon Lee, and Ton Kalker, "Quality Evaluation of Reverberation in Audioband Speech Signals," in *E-Business and Telecommunications*, Joaquim Filipe and Mohammad S. Obaidat, Eds., vol. 48, pp. 384–396. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[8] Karen S. Helfer and Laura A. Wilber, "Hearing Loss, Aging, and Speech Perception in Reverberation and Noise," *Journal of Speech, Language, and Hearing Research*, vol. 33, no. 1, pp. 149–155, Mar. 1990.

[9] DeLiang Wang and Jitong Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[10] Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 196–200.

[11] Hakan Erdogan, John R. Hershey, Shinji Watanabe, Michael I. Mandel, and Jonathan Le Roux, "Improved MVDR Beamforming Using Single-Channel Mask Prediction Networks," in *INTERSPEECH*. 2016, pp. 1981–1985, ISCA.

[12] Harry L Van Trees, *Optimim Array Processing*, Wiley, 2002.

[13] Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang, "Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, Sept. 2010.

[14] Lukas Drude, Jahn Heymann, Christoph Boeddeker, and Reinhold Haeb-Umbach, "NARA-WPE: A Python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing," in *Speech Communication; 13th ITG-Symposium*, 2018, pp. 1–5.

[15] M.S. Brandstein and H.F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, 1997, vol. 1, pp. 375–378.

[16] Joseph Hector DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*, Ph.D. thesis, Brown University, Providence, Rhode Island, USA, 2000.

[17] Femke B. Gelderblom, Yi Liu, Johannes Kvam, and Tor Andre Myrvoll, "Synthetic Data For Dnn-Based Doa Estimation of Indoor Speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, Canada, June 2021, pp. 4390–4394.