

Complex-valued Spatial Autoencoders for Multichannel Speech Enhancement

Mhd Modar Halimeh, *Student member, IEEE*, and Walter Kellermann, *Fellow, IEEE*

Abstract—In this contribution, we present a novel online approach to multichannel speech enhancement. The proposed method estimates the enhanced signal through a filter-and-sum framework. More specifically, complex-valued masks are estimated by a deep complex-valued neural network, termed the complex-valued spatial autoencoder. The proposed network is capable of exploiting as well as manipulating both the phase and the amplitude of the microphone signals. As shown by the experimental results, the proposed approach is able to exploit both spatial and spectral characteristics of the desired source signal resulting in a physically plausible spatial selectivity and superior speech quality compared to other baseline methods.

Index Terms—Multichannel signal processing, speech enhancement, deep learning, complex-valued networks.

I. INTRODUCTION

The widespread availability of devices with multiple microphones have boosted the interest in multichannel speech enhancement techniques for, e.g., source separation, source extraction, or noise suppression [1], [2].

The most commonly used multichannel speech enhancement technique is beamforming, where the spatial diversity of the different sound sources is exploited to emphasize sounds coming from the desired source's direction while suppressing sounds that arrive from other directions [3]–[5]. Many beamformers can be found in the literature derived under different constraints such as the popular Minimum Variance Distortionless Response (MVDR) beamformer [6], the Generalized MVDR (GMVDR) beamformer [7], the Generalized Eigenvalue (GEV) beamformer [8], [9], the Multichannel Wiener Filter (MWF) [10], and modulation-domain multichannel Kalman filter [11].

In general, conventional beamformers share the need for spatial information, whether in the form of steering vectors or spatial covariance matrices, in order to function properly. Recently, several data-driven methods have been proposed to estimate this information, e.g., in [12] a combination of a Deep Neural Network (DNN) and a maximum likelihood estimator is used to estimate the clean speech statistics and speech presence probability which are then used to compute the beamformer's weights. The authors in [13] proposed to use the Multichannel Non-negative Matrix Factorization (MNMF) to decompose time-frequency bins into speech and noise components to be used in obtaining the necessary statistics for an MVDR beamformer. MNMF is replaced by a DNN-based speech prior in [14] to estimate clean speech statistics.

Alternatively, departing from statistically optimum beamformers, a beamformer's weights can be directly estimated using DNNs. The authors in [15] proposed to train a DNN to estimate a beamformer's weights for maximizing the performance of a subsequent Automatic Speech Recognition (ASR) system without guaranteeing better speech quality. Similarly, time-domain beamformer weights are estimated using Long-Short Term Memory (LSTM) layers in [16] for better speech recognition performance. Robust speech recognition was also the aim in [17] where 'deep LSTM adaptive beamforming' is introduced. Another variant is to infer a time-frequency mask that is applied to the reference microphone to estimate the desired signal. This was done in [18] by employing a shared LSTM network across subbands and in [19] using a convolutional recurrent network. Sinc and dilated convolutional layers were used in [20] to perform waveform mapping.

In this paper, we present a novel approach to data-driven online multichannel spatio-spectral filtering using complex-valued DNNs. The proposed approach adopts the filter-and-sum technique from conventional beamforming as each channel is filtered by a complex-valued mask and the filtered channels are then added to produce the enhanced signal. This allows the network to produce effects such as phase-aligned superposition of the desired signal, in contrast to, e.g., [18]. Moreover, unlike, e.g., [15], the network is trained for speech quality enhancement and is not a preprocessing block for an ASR system, nor is it a supporting block for a conventional beamformer as, e.g., in [12]. Finally, we verify the validity of the proposed approach under various acoustic conditions, where the proposed network is shown to be capable of localizing and extracting the desired speech signal. In the following, signals in the Short Time Fourier Transform (STFT) domain are denoted by uppercase letters while signals in the time domain are denoted by lowercase letters. Furthermore, transposition is denoted by $(\cdot)^T$, $'*$ ' denotes conjugate complex, while vectors are denoted by boldface letters.

II. COMPLEX-VALUED SPATIAL AUTOENCODERS

We consider a scenario with M microphones, where at time-frequency bin (τ, f) the m -th microphone signal is given by

$$X_m(\tau, f) = D_m(\tau, f) + N_m(\tau, f), \quad (1)$$

where $D_m(\tau, f) = H_m^*(\tau, f)S(\tau, f)$ denotes the reverberant source signal, H_m denotes the Acoustic Transfer Function (ATF) from the desired source's position to the m -th microphone, while N_m denotes the background noise components as picked up by the m -th microphone. It must be pointed out that undesired components captured by N_m are restricted to non-speech components, i.e., interfering speakers are not

M. M. Halimeh and W. Kellermann are with the Chair of Multimedia Communications and Signal Processing (LMS), University of Erlangen-Nuremberg, 91058 Erlangen, Germany (e-mail: mhd.m.halimeh@fau.de; walter.kellermann@fau.de).

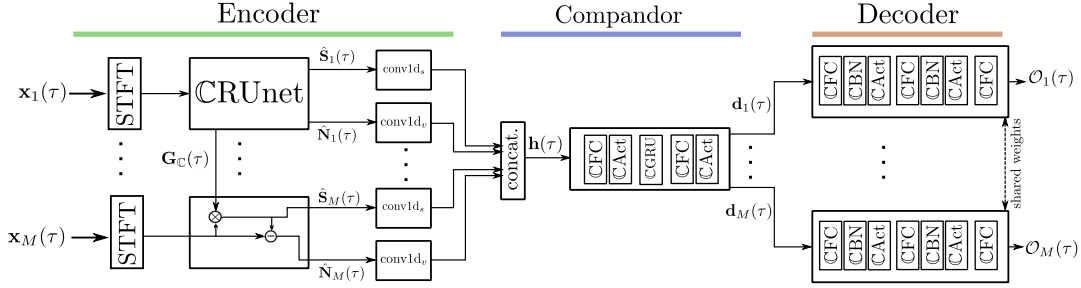


Fig. 1: The proposed complex-valued spatial autoencoder structure.

considered in this work. Nevertheless, N_m is not restricted to stationary nor diffuse noises, but it can represent arbitrary noises. Our goal in this paper is to extract the source signal S , or a reverberant version of it, with minimal distortions while suppressing the noise components N .

Complex-valued DNNs [21] have shown convincing results in single-channel speech enhancement [22], [23] as well as in echo suppression [24]. Their ability to manipulate and exploit phase information makes them a natural candidate for our multichannel signal processing task.

The proposed network architecture is shown in Fig. 1. As input, the network takes one frame of the time-domain signal $\mathbf{x}_m(\tau)$ comprising L samples per channel m , i.e., $\mathbf{x}_m(\tau) = [x_m(\tau), \dots, x_m(\tau - L + 1)]^T$ and outputs one complex-valued mask $\mathcal{M}_m(\tau)$ per channel. For each time-frequency bin (τ, f) , an estimate of the desired source signal is obtained as follows

$$\hat{S}(\tau, f) = \sum_{m=1}^M \mathcal{M}_m(\tau, f) \cdot X_m(\tau, f). \quad (2)$$

As seen from the figure, the networks starts by processing each channel's signal separately. Afterwards, information from all channels is processed jointly at the compandor unit in the middle. Finally, each channel's mask is constructed separately. This structure resembles the commonly used autoencoder structures and therefore, we denote it a *spatial autoencoder*.

A. Spatial Encoders

As seen in Fig. 1, M frames of length L of the M microphone signals $\{\mathbf{x}_m(\tau)\}_{m=1}^M$ are processed as follows: First, an STFT is performed to obtain $\{\mathbf{X}_m(\tau)\}_{m=1}^M$. Afterwards, the complex-valued signal $\mathbf{X}_1(\tau)$ is fed into a complex-valued subnetwork, denoted by CRUnet, that is a smaller variant of the network in [24] consisting of eight complex-valued convolutional modules with a complex-valued Gated Recurrent Unit (GRU) and a complex-valued Fully Connected (FC) layer in between. The CRUnet produces a complex-valued mask $\mathbf{G}_C(\tau)$ that is used across all M channels to obtain initial estimates of the desired speech components $\mathbf{S}_m(\tau)$ and undesired noise components $\mathbf{N}_m(\tau)$ as

$$\hat{\mathbf{S}}_m(\tau) = \mathbf{G}_C(\tau) \odot \mathbf{X}_m(\tau), \quad (3)$$

$$\hat{\mathbf{N}}_m(\tau) = (1 - \mathbf{G}_C(\tau)) \odot \mathbf{X}_m(\tau), \quad (4)$$

where \odot denotes the Hadamard product operator.

The use of the same mask $\mathbf{G}_C(\tau)$ across all microphone signals ensures the preservation of relative phase differences and therefore, the preservation of spatial information as encoded in the original microphone signals. On the other hand,

one should acknowledge that using a single complex-valued mask across the different channels cannot effectuate spatially selective filtering.

The initial signal components estimates are then downsampled using two single-dimensional convolutional layers denoted by conv1d_s and conv1d_n to reduce their dimensionality to $L_1 < L$. More specifically, the initial source estimates $\{\hat{\mathbf{S}}_m(\tau)\}_{m=1}^M$ are downsampled using conv1d_s that is shared across all M channels, while the noise estimates $\{\hat{\mathbf{N}}_m(\tau)\}_{m=1}^M$ are similarly downsampled using conv1d_n . This downsampling is done for purely computational purposes as a certain degree of redundancy is to be expected in the signals $\hat{\mathbf{S}}_m(\tau)$ and $\hat{\mathbf{N}}_m(\tau)$.

B. Spatial Compandor

The encoders lead to a compandor unit. The goal of the compandor unit is to estimate the necessary complex equalization, or an abstract representation thereof, that adjusts both the amplitude and phase of the different channels in order to extract the desired source exploiting both the spatial and the spectrotemporal domain. As the compandor is the only part of the network that has access to all channels simultaneously and where different channels are processed differently to lead to the desired spatial selectivity, it is also the part where spatial filtering is accomplished. Inspired by the coding literature, the term compandor here refers to the compression at the input side, where information from all channels is fused into a single channel stream to be processed jointly, while on the output side the single stream of information is expanded to the original number of channels. More specifically, at the input of the compandor, the signals resulting from the encoding stage are collected in the vector $\mathbf{h}(\tau) \in \mathbb{C}^{2ML_1}$. Therefore, $\mathbf{h}(\tau)$ encapsulates both spatial and spectral information regarding the desired source and any active noise sources.

The vector $\mathbf{h}(\tau)$ is then processed by a cascade of a complex-valued FC layer denoted by (CFC), a complex-valued leaky Rectified Linear Unit (ReLU) activation function [22] denoted by (CAct), a complex-valued GRU (CGRU), and finally a CFC and a CAct. These different layers will be characterized by their output sizes which are denoted by $\{L_2, L_3, ML_4\}$, respectively. The inclusion of the the CGRU enables the compandor to not only recognize and exploit instantaneous spatial and spectral patterns, but to also exploit the temporal evolution of these patterns.

Finally, the compandor outputs the vector $\mathbf{d}(\tau) \in \mathbb{C}^{ML_4}$, which is decomposed into M excitation vectors $\{\mathbf{d}_m(\tau)\}_{m=1}^M$ of length L_4 such that each vector is used to construct a complex-valued mask at the decoding stage.

C. Spatial Decoders

Following the compander unit is the decoding stage, where each excitation vector $\mathbf{d}_m(\tau)$ is fed into a decoder network consisting of a CFC, a complex-valued Batch Normalization (BN), and a CAct. This cascade is repeated once more and then followed by the final CFC layer. These layers will be characterized by their outputs' dimensions $\{L_5, L_6, L\}$, respectively. The final decoder layer outputs an unprocessed mask $\mathcal{O}_m(\tau, f)$ for each time-frequency bin (τ, f) that is used to obtain the complex-valued mask $\mathcal{M}_m(\tau, f)$ as follows [22]

$$|\mathcal{M}_m(\tau, f)| = \tanh(|\mathcal{O}_m(\tau, f)|), \quad (5)$$

and

$$e^{i\theta_{\mathcal{M}_m}(\tau, f)} = \frac{\mathcal{O}_m(\tau, f)}{|\mathcal{O}_m(\tau, f)|}. \quad (6)$$

It is worth noting that the aforementioned M decoder networks are identical, i.e., weights are shared across the M decoding channels, and as a consequence, any differences between the M masks $\{\mathcal{M}_m(\tau, f)\}_{m=1}^M$ can stem only from differences in the excitation vectors $\{\mathbf{d}_m(\tau)\}_{m=1}^M$ rather than channel-specific decoder networks.

Using Eq. (2) we obtain the STFT-domain estimate of the source signal $\hat{\mathbf{S}}(\tau)$ which can be transformed back to time domain to obtain the estimated signal frame $\hat{\mathbf{s}}(\tau)$.

D. Training and Optimization

As a training target, we employ the clean reverberant source signals filtered by an MVDR beamformer steered towards the source position

$$S_{\text{target}}(\tau, f) = \sum_{m=1}^M W_m^*(\tau, f) D_m(\tau, f), \quad (7)$$

where $W_m(\tau, f)$ denotes an MVDR beamformer weight at the (τ, f) time-frequency bin. The beamformer weights $W_m(\tau, f)$ are calculated using a recursively estimated noise spatial covariance matrix $\mathbf{R}_{NN}(\tau, f)$ using the ground truth noise signals $\{N_m(\tau, f)\}_{m=1}^M$ and a free-field steering vector towards to the source's ground truth Direction Of Arrival (DOA). A simple rearrangement of Eq. (7) as a function of the microphone signals yields

$$S_{\text{target}}(\tau, f) = \sum_{m=1}^M W_m^*(\tau, f) (cR_m(\tau, f) X_m(\tau, f)), \quad (8)$$

where $cR_m(\tau, f)$ denotes the ideal complex-valued ratio mask at microphone m and time-frequency bin (τ, f) . Clearly, this target is not attainable using only a spatial filter, i.e., $W_m(\tau, f)$, but instead, spectral filtering as represented by $cR_m(\tau, f)$ is needed, highlighting the difference to learning a conventional beamformer. Furthermore, compared to utilizing the 'dry' non-reverberant source signal, the proposed target is a reverberant image of the source signal and therefore, dereverberation is not targeted by the network.

To optimize the network's weights, the Signal-to-Noise-Ratio (SNR) loss function is used [25]

$$\mathcal{J}_{\text{SNR}}(\mathbf{s}_{\text{target}}(\tau), \hat{\mathbf{s}}(\tau)) = -10 \log_{10} \left(\frac{\|\mathbf{s}_{\text{target}}(\tau)\|^2}{\|\mathbf{s}_{\text{target}}(\tau) - \hat{\mathbf{s}}(\tau)\|^2} \right),$$

where $\|\cdot\|$ denotes the Euclidean norm, while $\mathbf{s}_{\text{target}}(\tau)$ and $\hat{\mathbf{s}}(\tau)$ denote the time-domain target signal and estimated desired signal, respectively.

TABLE I: Average performance of the various approaches.

	ΔSINR [dB]	SDR [dB]	ΔPESQ	ΔSTOI
CRUnet	7.7	4.2	0.16	0.03
DNN-MVDR	5.3	-	0.08	0.07
OMVDR	5.0	-	0.1	0.09
OGMVDR	14.3	-	0.24	0.12
COSPA	7.5	5.3	0.23	0.09

III. EXPERIMENTAL RESULTS

For evaluation we compare the proposed approach, denoted by Complex-valued Spatial Autoencoder (COSPA) to four different baseline methods:

- The use of the CRUnet as a DNN-based single-channel speech enhancement method. This network is trained to estimate a complex-valued mask that extracts $\mathbf{s}_m(\tau)$ from $\mathbf{x}_m(\tau)$ and is optimized using the SNR loss function [25]. This network had approximately 0.5 M parameters. When applied across M microphones, this approach provides one source signal estimate per microphone signal and therefore, its results were averaged over the M channels.
- A DNN-driven MVDR beamformer (denoted DNN-MVDR) which uses free-field steering vectors steered towards the true source DOA. The noise spatial covariance matrices are recursively estimated using the estimated noise microphone signals. The noise signals are estimated using complex-valued masks estimated by a pre-trained CRUnet. This approach is used as a representative of DNN-supported beamforming methods.
- An oracle knowledge MVDR beamformer (denoted OMVDR) which, similarly to the DNN-MVDR, uses free-field steering vectors steered towards the true source DOA. The noise spatial covariance matrices are recursively estimated using the ground truth noise microphone signals. This beamformer represents an upper bound for similar methods which rely on estimating the noise components in calculating the spatial covariance matrices.
- An oracle knowledge Generalized MVDR (GMVDR) beamformer (denoted OGMVDR) which uses the true Relative Transfer Function (RTF)s calculated w.r.t. the source position in addition to the true noise microphone signals for recursively estimating the spatial covariance matrices. This beamformer represents an upper bound for achievable performance using MVDR beamformers as it uses oracle spectral and spatial knowledge.

For all considered algorithms, online processing was carried out for a linear array with $M = 5$ omnidirectional microphones with uniform spacing of 4 cm, using signal frames of length 1024 samples and with frame shifts of 512 samples for a sampling frequency of $f_s = 16$ kHz. The COSPA was configured with $\{L_1 = 260, L_2 = L_3 = 128, L_4 = 513, L_5 = L_6 = 256\}$ resulting in approximately 2.7 M free parameters.

For this evaluation, two datasets were generated. In all datasets, each scenario included one desired speech source and two interferers, a noise source and a music source. The speech utterances were taken from the TIMIT dataset [26] with disjoint speakers for training and testing. The noise and music sequences were obtained from the MUSAN dataset [27], which includes singing voices among other types of noise, and

for which training and testing sequences were also disjoint. To generate the training dataset, 6000 scenarios, each 7 s long, (11hrs 40min) were created. Each scenario consisted of a room of random dimensions between $[3, 3, 1]$ m and $[8, 8, 4]$ m and a reverberation time sampled randomly from the range $[0.3 - 0.7]$ s. The positions of the microphone array, desired speech source, noise source and music source were also sampled randomly within the simulated room. The Room Impulse Responses (RIRs) of the simulated sources were generated using the image-source method [28].

As for the test dataset, 300 scenarios were generated using randomly sampled room dimensions, reverberation times, array, speech source, noise source and music source positions similar to the training dataset. The inter-microphone distance was identical across all scenarios in both datasets.

For both datasets, the SNR and signal-to-music ratio was sampled randomly per scenario from the range $[-7, 0]$ dB, individually. In addition, to simulate microphone noise, white additive noise for an SNR of 30 dB was added to each microphone signal.

To compare the different approaches, four different measures are used¹, averaged over time and scenarios:

- Δ SINR: describes the gain in terms of Signal-to-Interferer and Noise Ratio (SINR) when comparing the SINR at the first microphone to that of the enhanced signal. The SINR is calculated as the ratio between the energy of the (filtered) source signal to the energy of the (filtered) music and noise signals.
- SDR: describes the Signal-to-Distortion Ratio (SDR) as calculated for the (filtered) source signal to quantify the distortions introduced by the filtering [29].
- Δ PESQ: describes the PESQ (Perceptual Evaluation of Speech Quality [30]) difference between the unprocessed first microphone signal and the enhanced signal.
- Δ STOI: describes the STOI (Short-Time Objective Intelligibility [31]) difference between the unprocessed first microphone signal and the enhanced signal.

As a reference signal for the SDR, PESQ and STOI calculations, the *dry* non-reverberant source signal was used.

The averaged performance results are provided in TABLE I, where the OGMVDR beamformer performs best as it utilizes perfect spatial and spectral knowledge. When comparing the COSPA to single-channel CRUnet, clear gains are observed due to the utilization of spatio-spectral filtering in comparison to spectral filtering only. We must point out that due to the random nature of the testing dataset, it included scenarios of limited spatial diversity, in which the advantages of spatial filtering are less pronounced, driving the average results of the single-channel approach closer to other multichannel ones. A comparison between OMVDR, OGMVDR and COSPA places COSPA in terms of performance in-between the two oracle knowledge methods which is very encouraging given that COSPA is not provided any side information such as source DOA. It is worth mentioning that no SDR values are provided

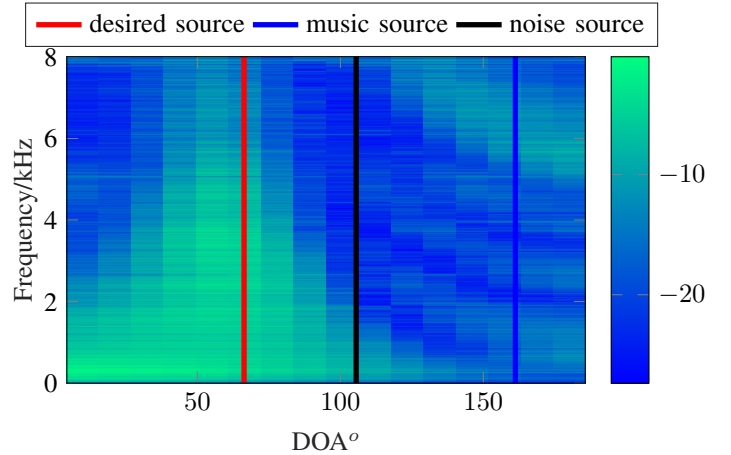


Fig. 2: An exemplary COSPA beampattern.

for the different MVDR beamformer variants, as distortionless response is guaranteed in the source’s direction.

To better examine the spatial selectivity of the proposed approach, the average results in Table I are complemented by the beampattern depicted in Fig. 2. This beampattern is generated by simulating 36 equidistant white noise sources placed at different DOAs with angular distance increments of 5° , under the free-field propagation assumption. Then, a set of complex-valued masks $\{\mathcal{M}_m(\tau); \tau = 1, 2, \dots\}_{m=1}^M$ is generated for one sample in the test set, i.e., to extract one desired speech signal from a noisy mixture, as described earlier. Using the masks $\{\mathcal{M}_m(\tau); \tau = 1, 2, \dots\}_{m=1}^M$, the white noise sources’ microphone signals are filtered, and the power of the filtered signals, averaged over the signals’ duration, is depicted in Fig. 2 in [dB] after being normalized to a maximum of 0 dB. As shown by the beampattern, the proposed COSPA is able to successfully localize the desired source as well as being spatially selective to emphasize signals coming from the source’s direction. One must point out that since Fig. 2 is averaged over time, an unseen aspect is the time-varying nature of the produced masks that, e.g., can exploit the different sources’ activity patterns. Finally, it is worth noting that unlike MVDR-based approaches, the COSPA does not guarantee a distortionless response in the source’s direction which can be seen as a result of the spectral filtering side of the method.

IV. CONCLUSION AND FINAL REMARKS

In this paper we introduced a novel data-driven approach to multichannel signal enhancement. This approach utilizes a complex-valued DNN, termed Complex-valued Spatial Autoencoder, to estimate complex-valued masks that are applied to the microphone signals. The proposed approach is compared to different single and multichannel approaches under different acoustic conditions, where the COSPA’s spatio-spectral filtering capabilities reflect physically plausible spatial selectivity and result in superior speech quality. Finally, encouraged by the results achieved in denoising, we plan on extending COSPA to the task of source extraction, where multiple interfering speakers are considered.

¹Audio examples and source code implementation can also be found at <https://github.com/ModarHalimeh/COSPA>

REFERENCES

- [1] Tie-Jun Shan and T. Kailath, "Adaptive beamforming for coherent signals and interference," *IEEE Trans. on Acoust., Speech, and Signal Process.*, vol. 33, no. 3, pp. 527–536, 1985.
- [2] J. Benesty and Y. Huang, *Adaptive Signal Processing: Applications to Real-World Problems*, Springer-Verlag Berlin Heidelberg, 2003.
- [3] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.
- [4] H. Van Trees, *Optimum Array Processing*, Wiley-Interscience, New York, 2002.
- [5] J. Benesty, I. Cohen, and J. Chen, *Array Processing: Kronecker Product Beamforming*, Springer International Publishing, 2019.
- [6] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer-Verlag Berlin Heidelberg, 2008.
- [7] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. on Signal Process.*, vol. 49, no. 8, pp. 1614–1626, 2001.
- [8] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized Eigenvalue decomposition," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 15, no. 5, pp. 1529–1539, 2007.
- [9] L. Pfeifenberger, M. Zöhrer, and F. Pernkopf, "Eigenvector-based speech mask estimation for multi-channel speech enhancement," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 27, no. 12, pp. 2162–2172, 2019.
- [10] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. on Signal Process.*, vol. 50, no. 9, pp. 2230–2244, 2002.
- [11] W. Xue, A. H. Moore, M. Brookes, and P. A. Naylor, "Modulation-domain multichannel Kalman filtering for speech enhancement," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 26, no. 10, pp. 1833–1847, 2018.
- [12] J. M. Martín-Doñas, J. Jensen, Z. H. Tan, A. M. Gomez, and A. M. Peinado, "Online multichannel speech enhancement based on recursive EM and DNN-based speech presence estimation," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 28, pp. 3080–3094, 2020.
- [13] K. Shimada, Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Unsupervised speech enhancement based on multichannel nmf-informed beamforming for noise-robust automatic speech recognition," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 27, no. 5, pp. 960–971, 2019.
- [14] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Semi-supervised multichannel speech enhancement with a deep speech prior," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 27, no. 12, pp. 2197–2212, 2019.
- [15] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *IEEE International*
- [16] Bo Li, Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, and Michiel Bacchiani, "Neural network adaptive beamforming for robust multi-channel speech recognition," in *Interspeech Conf.*, San Francisco, USA, Sep. 2016, p. 1976–1980.
- [17] Z. Meng, S. Watanabe, J. R. Hershey, and H. Erdogan, "Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition," in *IEEE International Conf. on Acoust., Speech and Signal Process.*, New Orleans, USA, Mar. 2017, pp. 271–275.
- [18] X. Li and R. Horaud, "Multichannel speech enhancement based on time-frequency masking using subband long short-term memory," in *IEEE Workshop on Applications of Signal Process. to Audio and Acoust.*, New Paltz, NY, USA, Oct. 2019, pp. 298–302.
- [19] S. Chakrabarty and E. A. P. Habets, "Time-frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 787–799, 2019.
- [20] C. Liu, S. Fu, Y. Li, J. Huang, H. Wang, and Y. Tsao, "Multichannel speech enhancement by raw waveform-mapping using fully convolutional networks," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 28, pp. 1888–1900, 2020.
- [21] C. Trabelsi *et al.*, "Deep complex networks," in *Proc. Int. Conf. Learning Representations*, Vancouver, BC, Feb. 2018.
- [22] H. Choi *et al.*, "Phase-aware speech enhancement with deep complex U-net," in *arXiv:1903.03107*, Feb. 2019.
- [24] M. Halimeh *et al.*, "Combining Adaptive Filtering and Complex-valued Deep Postfiltering for Acoustic Echo Cancellation," in *IEEE International Conference on Acoustics, Speech and Signal Processing Proceedings*, 2021.
- [25] J. Le Roux *et al.*, "SDR - half-baked or well done?," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Brighton, UK, May 2019.
- [26] J. Garofolo *et al.*, "TIMIT acoustic-phonetic continuous speech corpus LDC93S1," Web Download. Philadelphia: Linguistic Data Consortium.
- [27] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.
- [28] E. Habets, "Room impulse response generator," Tech. Rep., Technische Universiteit Eindhoven, The Netherlands, May 2006.
- [29] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," pp. 1462–1469, 2006.
- [30] ITU-T Recommendation P.862.2, "Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs," Recommendation, ITU, Nov. 2007.
- [31] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *IEEE International Conf. on Acoust., Speech and Signal Process*, Dallas, USA, Mar. 2010, pp. 4214–4217.