COMPLEX-VALUED SPATIAL AUTOENCODERS FOR MULTICHANNEL SPEECH ENHANCEMENT

Mhd Modar Halimeh, and Walter Kellermann

Multimedia Communications and Signal Processing,
Friedrich-Alexander University Erlangen-Nürnberg, Cauerstr. 7, 91058 Erlangen, Germany,
mhd.m.halimeh@fau.de

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 manipulating both the phase and the amplitude of the microphone signals and hence, the network 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.

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

1. INTRODUCTION

Multichannel speech enhancement aims at extracting the clean speech signal from a set of noisy microphone signals. Due to the widespread availability of devices with multiple microphones, interest in multichannel speech enhancement techniques for, e.g., source separation, source extraction, or noise suppression [1,2], has greatly increased over recent years.

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 from other directions [3–5]. Many beamformers can be found in the literature 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 Multichannel Non-negative Matrix Factorization (MNMF) to decompose time-frequency bins into speech and noise components for obtaining the necessary statistics for an MVDR beamformer. MNMF is replaced by a DNN-based speech prior in [14] to estimate clean speech statistics, and a DNN-driven MWF is proposed in [15].

Alternatively, departing from statistically optimum beamformers, a beamformer's weights can be directly estimated using DNNs. The authors in [16] 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 [17] for better ASR performance. Robust speech recognition was also the aim in [18] 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 [19] by employing a shared LSTM network across subbands and in [20] using a convolutional recurrent network. Sinc and dilated convolutional layers were used in [21] to perform waveform mapping. The authors in [22] proposed graph neural networks based on a U-net architecture to estimate the clean speech signal.

In this paper, we present a novel approach, termed the Complex-valued Spatial Autoencoder (COSPA), to data-driven online multichannel spatiospectral filtering using complex-valued DNNs. The COSPA adopts the filter-and-sum technique from conventional beamforming, allowing for phase-aligned superposition of the desired signal, in contrast to, e.g., [19]. Moreover, unlike, e.g., [16], the network is trained for speech quality enhancement and is not a preprocessing block for an ASR system. Due to its complex-valued nature, the COSPA is capable of processing spatiospectral information directly resulting in an end-to-end approach which eliminates the need for separate processing to, e.g., localize the desired source or estimate the necessary statistics, as in DNN-supported model-based methods, e.g., in [12, 15]. In particular, the COSPA uses a proper complex-valued network which processes complex-valued inputs by applying complex-valued weights and activation functions, in contrast to heuristically processing a concatenation of the input's real and imaginary components by applying real-valued networks, e.g., as in [22,23].

In the following, signals in the Short Time Fourier Transform (STFT) domain are denoted by uppercase letters and signals in the time domain are denoted by lowercase letters. Furthermore, transposition is denoted by $(\cdot)^{\mathrm{T}}$, '*' denotes conjugate complex and vectors are denoted by boldface letters. Finally, signal frames in the STFT domain are denoted by uppercase boldface letters and indexed only by their time index.

2. PROBLEM DESCRIPTION

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

$$X_m(\tau, f) = D_m(\tau, f) + N_m(\tau, f), \tag{1}$$

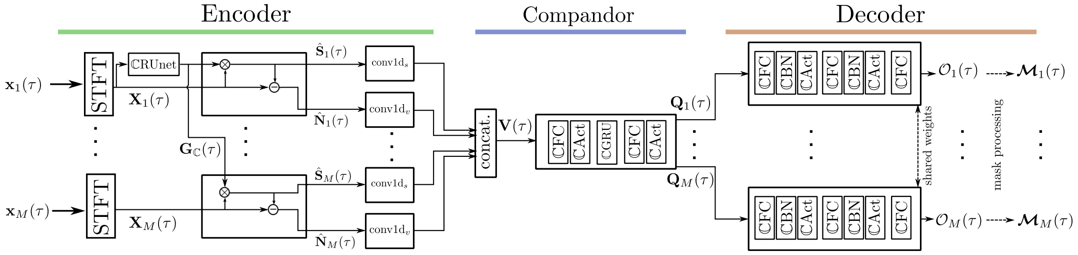

Fig. 1: The proposed complex-valued spatial autoencoder structure.

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 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 signals, i.e., interfering speakers are not 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.

3. COMPLEX-VALUED SPATIAL AUTOENCODERS

Complex-valued DNNs [24] have shown convincing results in single-channel speech enhancement [25, 26] as well as in echo suppression [27]. Their ability to manipulate and exploit phase information makes them a natural candidate for our multichannel signal processing task.

The COSPA 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., $\mathbf{x}_m(\tau) = [x_m(\tau), ..., x_m(\tau - L + 1)]^{\mathrm{T}}$ and outputs one complex-valued mask $\boldsymbol{\mathcal{M}}_m(\tau)$ per channel. For each time-frequency bin (τ, f), an estimate of the desired source signal is obtained as follows

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

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*.

3.1. Spatial Encoders

As seen in Fig. 1, M frames of length L of the M microphone signals $\{\mathbf{x}_m(\tau)\}_{m=1}^{M}$, i.e., one frame per channel, 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 [27] 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 $\mathbf{G}_{\mathbb{C}}(\tau)$ 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

$$\hat{\mathbf{S}}_m(\tau) = \mathbf{G}_{\mathbb{C}}(\tau) \odot \mathbf{X}_m(\tau), \tag{3}$$

$$\hat{\mathbf{N}}_m(\tau) = (1 - \mathbf{G}_{\mathbb{C}}(\tau)) \odot \mathbf{X}_m(\tau), \tag{4}$$

where $\odot$ denotes the Hadamard product operator.

The use of the same mask 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. However, one should acknowledge that using a single complex-valued mask across the different channels cannot effectuate spatially selective filtering, and can be disadvantageous for scenarios where the different microphones face different noise conditions.

The initial signal components estimates are downsampled using two single-dimensional convolutional layers denotes by conv1d$_s$ and conv1d$_n$ to reduce their dimensionality to $L_1 < L$. More specifically, the 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 computational purposes as a certain degree of redundancy is expected in the signals $\hat{\mathbf{S}}_m(\tau)$ and $\hat{\mathbf{N}}_m(\tau)$.

3.2. 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. 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 $\mathbf{V}(\tau) \in \mathbb{C}^{2ML_1}$. Therefore, $\mathbf{V}(\tau)$ encapsulates both spatial and spectral information regarding the desired source and any active noise sources.

The vector $\mathbf{V}(\tau)$ 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 [25] 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 $\mathbf{Q}(\tau) \in \mathbb{C}^{ML_4}$, which is decomposed into M excitation vectors $\{\mathbf{Q}_m(\tau)\}_{m=1}^{M}$ of length L_4 such that each vector is used to construct a complex-valued mask at the decoding stage.

3.3. Spatial Decoders

Following the compandor unit is the decoding stage, where each excitation vector $\mathbf{Q}_m(\tau)$ is fed into a decoder network consisting of a $\mathbb{C}$FC, a complex-valued Batch Normalization (BN), and a $\mathbb{C}$Act. This cascade is repeated once more and then followed by the final $\mathbb{C}$FC 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 [25]

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

and

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

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 $\{\mathbf{Q}_m(\tau)\}_{m=1}^{M}$ rather than channel-specific decoder networks. Finally, the STFT-domain estimate of the source signal, $\hat{\mathbf{S}}(\tau)$, is obtained using Eq. (2), while the corresponding time-domain signal frame $\hat{\mathbf{s}}(\tau)$ is obtained using the inverse STFT.

3.4. Training and Optimization

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

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

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 the source's ground truth Direction Of Arrival (DOA). The training target Eq. (7) is used to implicitly constrain the network to not severely distort signals arriving from the source's direction. A simple rearrangement of Eq. (7) as a function of the microphone signals yields

$$S_{\text{target}}(\tau, f) = \sum_{m=1}^{M} W_m^*(\tau, f) \left(R_{\mathbb{C},m}(\tau, f) X_m(\tau, f) \right), \tag{8}$$

where $R_{\mathbb{C},m}(\tau, f)$ denotes the ideal complex-valued ratio mask at microphone m and time-frequency bin (τ, f) that does not take into account, e.g., phase-alignment across channels. Clearly, this target is not attainable using only a spatial filter, i.e., $W_m(\tau, f)$, but instead, spectral filtering as represented by $R_{\mathbb{C},m}(\tau, f)$ 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 passed through an MVDR beamformer, and therefore, explicit dereverberation is not targeted by the network.

To optimize the network's weights, the Signal-to-Noise-Ratio (SNR) loss function is used [28]

$$\mathcal{J}_{\text{SNR}}(\mathbf{s}_{\text{target}}(\tau), \hat{\mathbf{s}}(\tau)) = -10 \log_{10} \left(\frac{\|\mathbf{s}_{\text{target}}(\tau)\|^2}{\|\mathbf{s}_{\text{target}}(\tau) - \hat{\mathbf{s}}(\tau)\|^2} \right), \tag{9}$$

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.

4. EXPERIMENTAL RESULTS

For evaluation, we compare the COSPA to five benchmarks:

Table 1: Average performance of the various approaches.

	ΔSINR [dB]	SDR [dB]	ΔPESQ	ΔSTOI
$\mathbb{C}$RUnet	7.7	4.2	0.16	0.03
DNN-MVDR	5.3	-	0.08	0.07
DNN-MWF	5.1	5.0	0.16	0.04
OMVDR	5.0	-	0.1	0.09
OGMVDR	14.3	-	0.24	0.12
COSPA	7.5	5.3	0.23	0.09

(I) The $\mathbb{C}$RUnet as a DNN-based single-channel speech enhancement method. This network with approximately 0.5 M parameters 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 [28]. When applied to all M microphones, this approach provides one source signal estimate per microphone signal and therefore, its performance metrics were averaged over the M channels.

(II) 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 from the estimated noise microphone signals. The noise signals are obtained using complex-valued masks estimated by a pre-trained $\mathbb{C}$RUnet. This approach is similar to [29] which uses a real-valued DNN instead. Moreover, while [29] estimates the source's DOA from the microphone signals, the DNN-MVDR is provided ground truth DOAs and as a consequence, is expected to outperform [29]. This approach is used to represent the performance of DNN-supported, MVDR-based, spatial filtering methods.

(III) A DNN-driven MWF (denoted DNN-MWF), where the noise and speech spatial covariance matrices are recursively estimated from estimated noise and speech microphone signals. These signals are estimated using complex-valued masks by a pre-trained $\mathbb{C}$RUnet. This approach is similar to [15] where noise and speech statistics are estimated by a real-valued DNN, and is used as a benchmark for the performance of spatiospectral filtering via DNN-supported MWF-based methods.

(IV) An oracle knowledge MVDR beamformer [6] (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.

(V) An oracle knowledge Generalized MVDR (GMVDR) beamformer (denoted OGMVDR) [7] 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.

The hyperparameters of the different algorithms were optimized w.r.t. a subset of the test dataset such that the best performance of each method was pursued and the benchmark methods were all normalized for a fair comparison. 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 [30] with disjoint speakers for training and testing. The noise and music sequences were obtained from the MUSAN dataset [31], which includes singing voices among other types of noise, and for which training and testing interferers 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 [32].

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 used[1], averaged over time and scenarios:

- ΔSINR: 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 [33].
- ΔPESQ: describes the PESQ (Perceptual Evaluation of Speech Quality [34]) difference between the unprocessed first microphone signal and the enhanced signal.
- ΔSTOI: describes the STOI (Short-Time Objective Intelligibility [35]) 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 1, where (V) performs best as it utilizes perfect spatial and spectral knowledge. When comparing the COSPA to the single-channel approach (I), clear gains are observed due to the utilization of spatiospectral 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. The comparison places COSPA in terms of performance in-between the two oracle knowledge methods, (IV) and (V), even though COSPA does not exploit any side information such as source DOA. The comparison between the COSPA and (III) is especially interesting, as both methods estimate the spatiospectral filters using channel-wise signal components estimates. However, while (III) follows a model-based approach to obtain the spatiospectral filters, COSPA is completely data-driven and does not assume a model for the different signal components. It is worth mentioning that no SDR values are provided for the different MVDR beamformer variants, as distortionless response is guaranteed in the source's direction.

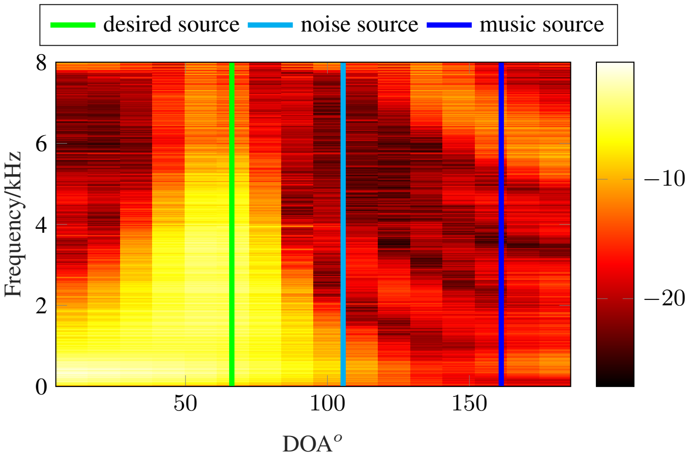

Fig. 2: An examplary COSPA beampattern.

To better examine the spatial selectivity of the proposed approach, the average results in Table 1 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 $\{\mathcal{M}_m(\tau); \quad \tau = 1, 2, ...\}_{m=1}^{M}$ 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 $\{\mathcal{M}_m(\tau); \quad \tau = 1, 2, ...\}_{m=1}^{M}$, 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.

5. 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. This end-to-end approach represents a competitive alternative to conventional beamforming methods and achieves source localization and the estimation of spatiospectral filters by a single system. Experiments demonstrate that COSPA's spatiospectral 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.

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

6. 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. Van Veen and K. 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 et al., *Array Processing: Kronecker Product Beamforming*, Springer International Publishing, 2019.

[6] J. Benesty et al., *Microphone Array Signal Processing*, Springer-Verlag Berlin Heidelberg, 2008.

[7] S. Gannot et al., “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 et al., “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 et al., “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. Martín-Doñas et al., “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 et al., “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 et al., “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] Y. Masuyama et al., “Consistency-aware multi-channel speech enhancement using deep neural networks,” in *arXiv:2002.05831*, Feb. 2020.

[16] X. Xiao et al., “Deep beamforming networks for multi-channel speech recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 5745–5749.

[17] B. Li et al., “Neural network adaptive beamforming for robust multichannel speech recognition,” in *Interspeech Conf.*, San Francisco, USA, Sep. 2016, p. 1976–1980.

[18] Z. Meng et al., “Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition,” in *ICASSP*, New Orleans, USA, Mar. 2017, pp. 271–275.

[19] 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.

[20] S. Chakrabarty and E. 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.

[21] C. Liu et al., “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.

[22] P. Tzirakis et al., “Multi-channel speech enhancement using graph neural networks,” in *ICASSP*, Toronto, Canada, June 2021, pp. 3415–3419.

[23] Z. Wang and D. Wang, “Multi-microphone complex spectral mapping for speech dereverberation,” in *arXiv:2003.01861*, Mar. 2020.

[24] C. Trabelsi et al., “Deep complex networks,” in *Proc. Int. Conf. Learning Representations*, Vancouver, BC, Feb. 2018.

[25] H. Choi et al., “Phase-aware speech enhancement with deep complex U-net,” in *arXiv:1903.03107*, Feb. 2019.

[26] Y. Hu et al., “DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement,” in *arXiv:2008.00264*, Aug. 2020.

[27] M. Halimeh et al., “Combining Adaptive Filtering and Complex-valued Deep Postfiltering for Acoustic Echo Cancellation,” in *ICASSP*, Toronto, Canada, 2021.

[28] J. Le Roux et al., “SDR - half-baked or well done?,” in *ICASSP*, Brighton, UK, May 2019.

[29] H. Erdogan et al., “Improved MVDR beamforming using single-channel mask prediction networks,” in *Interspeech*, 2016, p. 1981–1985.

[30] J. Garofolo et al., “TIMIT acoustic-phonetic continuous speech corpus LDC93S1,” Web Download. Philadelphia: Linguistic Data Consortium.

[31] D. Snyder et al., “MUSAN: A Music, Speech, and Noise Corpus,” 2015, arXiv:1510.08484v1.

[32] E. Habets, “Room impulse response generator,” Tech. Rep., Technische Universiteit Eindhoven, The Netherlands, May 2006.

[33] E. Vincent et al., “Performance measurement in blind audio source separation,” pp. 1462–1469, 2006.

[34] 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.

[35] C. Taal et al., “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in *ICASSP*, Dallas, USA, Mar. 2010, pp. 4214–4217.