Localizing Spatial Information in Neural Spatiospectral Filters

Annika Briegleb   Thomas Haubner   Vasileios Belagiannis   Walter Kellermann

Multimedia Communications and Signal Processing

Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

{annika.briegleb, thomas.haubner, vasileios.belagiannis, walter.kellermann}@fau.de

Abstract—Beamforming for multichannel speech enhancement relies on the estimation of spatial characteristics of the acoustic scene. In its simplest form, the delay-and-sum beamformer (DSB) introduces a time delay to all channels to align the desired signal components for constructive superposition. Recent investigations of neural spatiospectral filtering revealed that these filters can be characterized by a beampattern similar to one of traditional beamformers, which shows that artificial neural networks can learn and explicitly represent spatial structure. Using the Complex-valued Spatial Autoencoder (COSPA) as an exemplary neural spatiospectral filter for multichannel speech enhancement, we investigate where and how such networks represent spatial information. We show via clustering that for COSPA the spatial information is represented by the features generated by a gated recurrent unit (GRU) layer that has access to all channels simultaneously and that these features are not source- but only direction of arrival-dependent.

Index Terms—spatial filtering, multichannel speech enhancement, DNN interpretability

I. Introduction

Spatial filtering for speech and audio signals usually relies on signal-independent and -dependent beamformers such as the delay-and-sum beamformer (DSB) or the minimum variance distortionless response (MVDR) beamformer. Currently, (deep) neural networks that can support [1]–[5] or substitute [6]–[10] conventional beamformers are moving into the focus for speech enhancement and target speech extraction in noisy and reverberant scenarios. We denote the latter approaches as *neural spatiospectral filters*.

The simplest method for spatial filtering in speech enhancement, the DSB, solely relies on phase alignment of the signals captured by the different channels to enhance the desired signal. For this, traditional beamformers require knowledge of the direction of arrival (DoA) of the desired source. On the other hand, for a single desired speaker in the presence of non-speech or nondirectional background noise, neural spatiospectral filters, often do not require this information [6], [7]. While beampatterns demonstrate that neural spatiospectral filters can represent spatial information [6], [11], it still needs to be clarified how the spatial information contained in the multichannel input is processed and represented by these neural spatiospectral filters. In this contribution, we investigate where and how spatial information is captured in an exemplary neural spatiospectral filter. For this, we analyze the Complex-valued Spatial Autoencoder (COSPA) [6], which estimates a complex-valued multichannel mask for speech enhancement.

While each neural spatiospectral filter has its own architecture design, most filters include temporal processing. We choose COSPA because the joint temporal processing of all channels is localized in one gated recurrent unit (GRU) layer in the bottleneck of the network and the specific architecture of COSPA (cf. Sec. II) allows to analyze the spatial processing conveniently. Consequently, we investigate the spatial information contained in the features of the network before and after this GRU layer as it is the only network part with memory that can modify all channels individually, which is required to align the phases of the different channel signals.

In Sec. II, we discuss the signal model, the architecture of COSPA, and why its GRU layer is of special interest for spatial filtering. In Sec. III, we explain how we measure the spatial information contained in the network's features. We describe our experimental setup in Sec. IV-A and present and discuss our results in Sec. IV-B. Sec. V concludes the paper.

II. Spatial filtering with COSPA

In this contribution, we consider the task of extracting the signal of a point-like speech source from a multichannel signal recorded by M microphones in a reverberant room using a neural spatiospectral filter that estimates a complex-valued mask for speech enhancement. The filter can thus modify both the magnitude and the phase of the input signal. In the short-time Fourier transform (STFT) domain, we consider the frame-wise single-channel speech signal estimate $\hat{S}(\tau)$, the multichannel signal $X_m(\tau), m = 1, ..., M$, and the complex-valued multichannel filter mask $\mathcal{M}_m(\tau)$ estimated by the neural filter. Frames are half-overlapping and indexed by τ in both the time and the STFT domain. We omit the frequency index for brevity. The estimated speech signal is obtained by

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

As neural spatiospectral filter for our investigations we choose COSPA as introduced in [6], which estimates the complex-valued mask $\mathcal{M}_m$ for each channel m. A simplified structure of COSPA is depicted in Fig. 1. COSPA consists of an encoder and a decoder, which both process all signal channels equally, and a compandor, which can process each channel individually. As spatial information is obtained from the differences between signals captured by spatially distributed
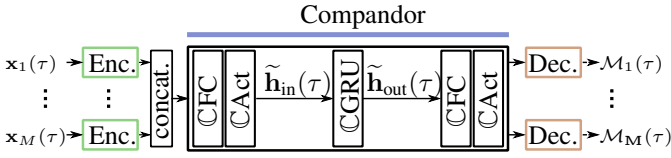
Fig. 1: Architecture of COSPA (adapted from [6]).

sensors, and processing it usually includes phase alignment, i.e., delays, the compandor is of specific interest for investigating the representation of spatial information inside COSPA. As shown in Fig. 1, the compandor consists of a complex-valued GRU layer [12], [13] in between two complex-valued fully connected layers. The GRU layer consisting of U_{out} complex-valued GRUs (see [6]) provides memory to the network, which is expected to facilitate the phase alignment of the channels and is thus in the focus of our analysis. The relevance of the GRU layer is also supported by [14], where the GRU layer is initialized with features derived from the DoA of the desired speaker to guide the network for target speaker extraction. Note that, here, for the considered scenario of speech enhancement without speech interferers COSPA does not receive any explicit information about the DoA of the desired source.

As shown in Fig. 1, we denote the features at the input of the GRU layer as $\widetilde{\mathbf{h}}_{\text{in}}(\tau) = [\widetilde{h}_{\text{in}}(\tau, 1), ..., \widetilde{h}_{\text{in}}(\tau, U_{\text{in}})]^\top$, with U_{in} being the number of input channels of the GRU layer, and $\widetilde{h}_{\text{in}}(\tau, u) \in \mathbb{C}$ being a single feature for unit u and time frame τ. These features are obtained from the first fully connected layer in the compandor, which is the first layer in COSPA that can fuse all channels. Similarly, we define $\widetilde{\mathbf{h}}_{\text{out}}(\tau) = [\widetilde{h}_{\text{out}}(\tau, 1), ..., \widetilde{h}_{\text{out}}(\tau, U_{\text{out}})]^\top$ as features at the output of the GRU layer with U_{out} being the number of GRUs in the GRU layer.

III. QUANTIFYING SPATIAL INFORMATION

As explained in Sec. II, we expect that the features generated by COSPA's GRU layer contain spatial information about the desired source. Hence, for a test signal in which the speaker position changes discretely with time, we expect that output features of the GRU layer corresponding to frames coming from the same source position are similar to each other and thereby express DoA information resulting from phase alignment. In contrast, at the input of the GRU layer we do not expect to find such similarities in the feature vectors. To identify frames characterized by similar feature vectors and to quantify the amount of spatial information contained in the features, we use the k-means clustering algorithm [15] on the features $\widetilde{\mathbf{h}}_{\text{in/out}}(\tau)$ at the input and the output of the GRU layer in the compandor. As these features can have very different value ranges per unit, we normalize the features of each unit per test sequence to have an amplitude range of -1 to 1 for our clustering analysis. Hence, for clustering, each frame is characterized by its normalized feature vector $\mathbf{h}_{\text{in/out}}(\tau)$.

Due to the expected discriminative nature of the features, we use the L1 distance $\text{d}(\cdot, \cdot)$ for the k-means algorithm. Given Q

sources, which can be different speakers or the same speaker at different positions, we use $k = Q + 1$ cluster centers, to allow one cluster for each of the Q sources in a test sequence, and one cluster for signal pauses as we assume that spatial information cannot be extracted from silent signal frames. We assign the pause label to the cluster that has the lowest signal energy in the corresponding target signal segments. The labels for the Q sources are assigned by a majority vote of the assigned frames.

We quantify the clustering results in terms of the *grouping success*, the average distance $\bar{d}$ of the features to their assigned cluster center, and the overall percentage of frames assigned to the pause cluster. The grouping success measures in percent how many of the frames belonging to one source are assigned to the same cluster. For $Q = 2$, a grouping success of 50% means that no distinction can be made between the two sources, while a grouping success of 100% means that for both sources all frames belonging to the active segments of the corresponding source are assigned to the same cluster. For the calculation of the grouping success we only consider frames assigned to one of the active source clusters and disregard all frames assigned to the pause cluster. Since pauses can also appear while a speaker is active, the assignment to the pause cluster is not necessarily wrong but uninformative. Hence, we disregard these frames.

To account for the pause cluster, we measure the overall percentage of frames assigned to the pause cluster as we expect the features after the GRU layer to be more expressive with respect to the corresponding DoA and to benefit from the memory of the GRU layer. Hence, we expect that the features $\mathbf{h}_{\text{out}}(\tau)$ at the output of the GRU layer can be clustered more consistently than the features $\mathbf{h}_{\text{in}}(\tau)$ at the input, and therefore the pause cluster should exhibit a lower percentage.

Finally, the average distance of the features to the cluster centers $\bar{d}$ is computed as

$$\bar{d} = \frac{1}{S} \sum_{s=1}^{S} \left(\frac{1}{k} \sum_{n=1}^{k} \left(\frac{1}{T_{n,s}} \sum_{\tau_{n,s}=1}^{T_n} \text{d}(\mathbf{c}_{n,s}, \mathbf{h}(\tau_{n,s})) \right) \right), \quad (2)$$

where $\tau_{n,s} = 1, ..., T_{n,s}$ indexes the frames assigned to cluster n in sequence $s, s = 1, ..., S$, and $\text{d}(\mathbf{c}_{n,s}, \mathbf{h}(\tau_{n,s}))$ describes the L1 distance from feature vector $\mathbf{h}(\tau_{n,s})$ to cluster center $\mathbf{c}_{n,s}$. $\bar{d}$ describes how dense the clusters are: a smaller value means that the assigned features are closer to their cluster center and hence more similar to each other.

IV. EXPERIMENTAL VALIDATION

So far, experiments with COSPA have only been conducted for scenarios with one spatially static desired source [6]. To investigate the exploitation of the spatial information contained in the multichannel signal, we perform experiments with changing speaker positions within one signal sequence. The purpose of our experiments is to verify the hypothesis that a layer with memory that can access all channels individually in a neural spatial filter, i.e., the GRU layer in COSPA's compandor, explicitly represents spatial information about

the acoustic scene at its output. We first examine COSPA's masking behavior for changing speaker positions in a scenario without background noise and with only one active speaker at any time. We analyze the features before and after the GRU layer and measure their discriminative power via clustering. Then, we investigate whether the features distinguish between speakers or between DoAs, and, finally, we assess how robust the spatial processing is against spatially and spectrally white background noise.

A. Experimental setup

For all experiments, signals are generated by simulating a room with dimensions sampled uniformly from [4-8, 4-8, 1-4] m and a reverberation time between 200 ms and 500 ms. A uniformly spaced linear microphone array (ULA) with $M = 3$ microphones and element spacing of 0.04 m, is randomly placed in the room with uniformly distributed positions. Defining the array's endfire directions as 0° and 180°, speaker positions are chosen between 0° and 180°. The speech data is taken from the TIMIT database [16] with its default split between training data and test data, and convolved with the corresponding room impulse response generated according to [17] for the respective speaker positions. All training and test sequences have a length of 7 s at 16 kHz sampling rate.

We create two training datasets of approximately 8 h each for our experiments. In the first dataset, denoted by *DS-clean*, each training sequence contains two speakers which are placed randomly at $Q = 2$ different positions in the room and are talking alternatingly. In each sequence, there are three segments, where Speaker 1 is active in the first and third segment and Speaker 2 is active in the middle segment. The time instants of the two speaker changes are chosen randomly, uniformly distributed between 1 and 3 s, and 5 and 6 s respectively. The angular distance of the two speakers' DoAs is at least 20° and their distance to the array and the walls is at least 0.3 m. For creating the target signal, the multichannel input speech signal image corresponding to each speaker separately is filtered with a DSB steered towards the position of the respective speaker. The two signals for the respective speakers are combined to obtain the target signal for the end-to-end training of COSPA. Note that this dataset does not contain any background noise and only serves the purpose of investigating the phase alignment capabilities of COSPA.

The second training dataset, denoted by *DS-WGN*, has the same setup as DS-clean but the signals consist of the speech signals and additive spatially and spectrally white Gaussian noise at signal-to-noise ratios (SNRs) between -10 dB and 50 dB in steps of 5 dB. We train COSPA individually on both training datasets and denote the resulting trained networks as COSPA_{clean} and COSPA_{WGN} respectively.

For testing, we define four datasets. *DST-clean* has the same setup as the training dataset DS-clean and consists of 50 test sequences used for the verification of the hypothesis that the GRU layer output $\widetilde{\mathbf{h}}_{out}(\tau)$ represents spatial information about the source. *DST-WGN* has 50 test sequences
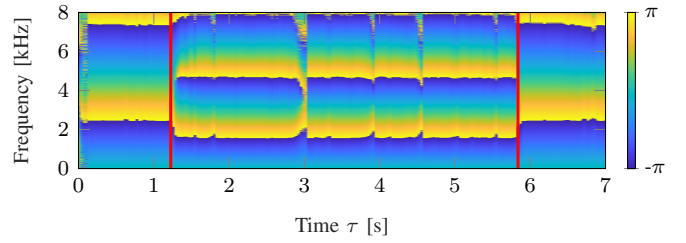

Fig. 2: Exemplary phase component of the estimated mask $\mathcal{M}_2$. Speaker 1 is at 103° and Speaker 2 at 43°. The red lines indicate the change of the active speaker.

per SNR $\in [-10, 5, 0, 5, 10, 20, 30, 50]$ dB but otherwise has the same setup as the training dataset DS-WGN. This test dataset is used to test the robustness against noise of the spatial representation in the GRU layer. To show that the features $\mathbf{h}_{out}(\tau)$ encode spatial and not speaker information, we use *DST-1Pos*, which contains 50 sequences with two alternating speakers placed at the same position in the room, and *DST-1Spk*, which contains 50 signals from a single speaker including one position change. $Q = 2$ and $k = 3$ for all experiments.

We parameterize COSPA as published in [6]. Especially, the fully connected layer before the GRU layer in the compandor and the GRU layer itself both have $U_{in} = U_{out} = 128$ units generating both the real and imaginary part of the $U_{in/out}$ complex-valued features $\widetilde{h}_{in/out}(\tau, u)$. For computing the STFT features as input to COSPA, we use frames of 1024 time-domain signal samples and a frame shift of 512 samples.

Since clustering is initialization-dependent, we report the results for the clustering as average over five trials, where in each trial the best out of five k-means clustering attempts is used to compute the evaluation metrics.

B. Results

Spatial adjustment: We first investigate whether COSPA can react to changing spatial conditions within an acoustic scene without guiding information during testing. As a typical example, in Fig. 2, it can be seen that for the test dataset DST-clean the phase mask $\mathcal{M}_2$ generated by COSPA_{clean} indeed reflects the DoA change following the changing speaker activity. From this it can be concluded that spatial information of the received signal is used by COSPA to generate the phase component of the estimated mask $\mathcal{M}_m$.

Based on this observation and the special role of the GRU layer in COSPA's compandor (cf. Sec. II), we investigate whether the features generated before and after the GRU layer also reflect the positional changes in the signal.

Features: Figs. 3a and b show the feature vectors $\mathbf{h}_{in}(\tau)$ and $\mathbf{h}_{out}(\tau)$ at the input and output of the GRU layer, respectively, corresponding to the time-domain signal in Fig. 3e. In Fig. 3a, speech pauses are already clearly discernible from speech for $\mathbf{h}_{in}(\tau)$, especially around $\tau = 5s$ and following $\tau = 6s$. Otherwise, no distinctive differences between the two speakers are apparent. In Fig. 3b, it can be observed that the features $\mathbf{h}_{out}(\tau)$ are more consistent along the time axis than
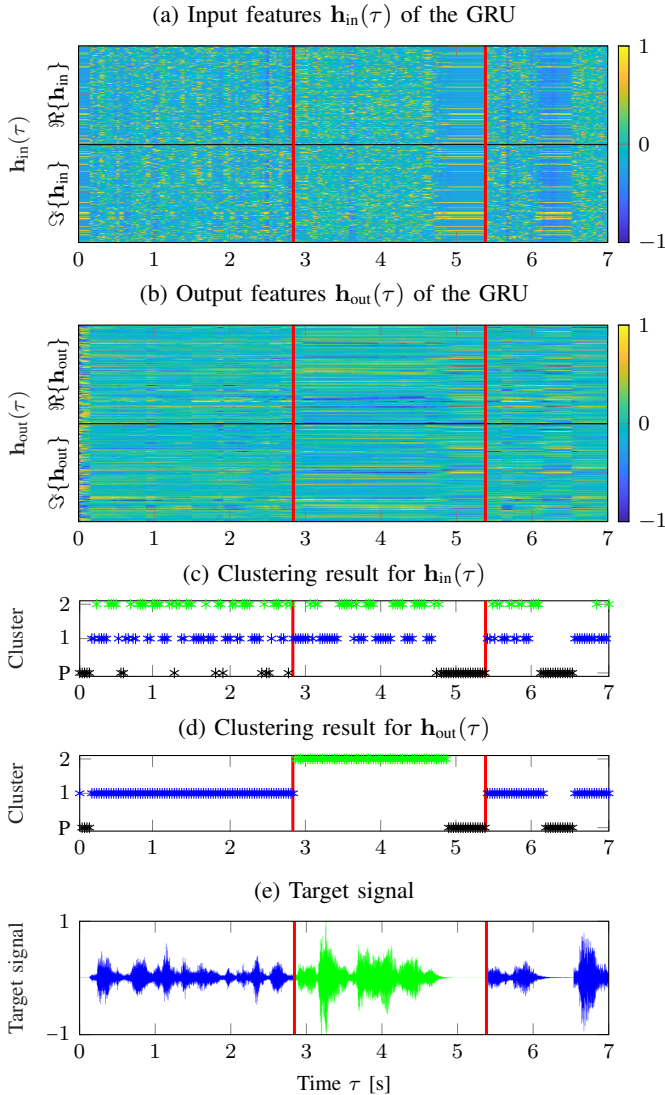
(a) Input features $\mathbf{h}_{in}(\tau)$ of the GRU

(b) Output features $\mathbf{h}_{out}(\tau)$ of the GRU

(c) Clustering result for $\mathbf{h}_{in}(\tau)$

(d) Clustering result for $\mathbf{h}_{out}(\tau)$

(e) Target signal

Fig. 3: Stacked real and imaginary part of exemplary feature vectors $\mathbf{h}(\tau)$ in COSPA$_{clean}$ before (a) and after (b) the GRU layer, and the respective clustering result (P: Pause, 1: Position 1, 2: Position 2) (c, d) and time-domain target signal (e). Speaker 1 is at $51°$, Speaker 2 at $78°$. The red lines indicate the position change.

$\mathbf{h}_{in}(\tau)$, that pauses are still discernible, and that there is clearly a specific feature pattern for each speaker.

Clustering results: Figs. 3c,d show the clustering results according to the features presented in Figs. 3a,b. Clusters are colored according to the signal part that is represented. It can be clearly seen that for the input features $\mathbf{h}_{in}(\tau)$ clustering appears to be uninformative concerning the speaker position, while the pauses are detected quite well. For the output feature vectors $\mathbf{h}_{out}(\tau)$ clustering is very informative regarding the activity of a source from a certain DoA.

Table I provides the clustering results for features obtained for COSPA$_{WGN}$ for the four test datasets. For DST-clean, it can be seen that 95.1% of frames belonging to either speaker position are assigned to the corresponding cluster at the output

TABLE I: Grouping success and characteristics of clusters for the COSPA$_{WGN}$ features before and after the GRU layer.

Test dataset	SNR [dB]	grouping success [%] ↑		average distance to cluster centers $\bar{d}$ ↓		relative size of pause cluster [%] ↓	
		$\mathbf{h}_{in}$	$\mathbf{h}_{out}$	$\mathbf{h}_{in}$	$\mathbf{h}_{out}$	$\mathbf{h}_{in}$	$\mathbf{h}_{out}$
DST-clean	∞	78.3	95.1	37.5	33.3	12.5	8.9
DST-1Spk	∞	71.0	94.6	39.4	34.6	28.8	19.1
DST-1Pos	∞	57.7	66.3	37.4	32.6	30.2	24.9
DST-WGN	50	80.7	94.5	37.6	33.2	32.0	21.5
	30	80.4	95.3	38.9	34.2	31.7	21.9
	20	79.1	94.6	41.2	35.4	33.0	24.8
	10	74.3	93.3	41.6	34.6	41.7	27.3
	5	68.9	90.1	42.0	33.7	45.2	31.4
	0	66.0	84.4	42.6	32.6	48.6	36.6
	-5	63.5	82.0	43.6	31.8	45.0	38.9
	-10	61.1	77.4	44.9	32.5	32.6	29.4

of the GRU layer, whereas only 78.3% of corresponding frames are grouped before the GRU layer. This shows that the memory in the GRU layer supports extracting the spatial information significantly, but also that the features after the first fully-connected layer in the compandor already contain a limited amount of spatial information.

Table I also gives the average distance of the data points to the respective cluster centers $\bar{d}$, which confirms that the clusters for $\mathbf{h}_{out}(\tau)$ are more compact than those for $\mathbf{h}_{in}(\tau)$, since $\bar{d}$ is smaller. This also points to more informative and discriminative features after the GRU layer.

The pause cluster: It can be noted from both Fig. 3 and Table I that fewer frames are assigned to the pause cluster at the output of the GRU layer than at the input, and that those frames belong to longer speech pauses and are not just single frames. This can be explained by the fact that at the input of the GRU layer, the pause cluster aside from the silent frames also contains several undecided frames and therefore has relatively many frames assigned to it. At the output of the GRU layer this cluster contains only frames with actual speech pauses and hence also has a lower number of frames assigned to it. This also shows the beneficial effect of the GRU layer's memory, since it creates, based on previous signal parts, more meaningful features for frames where signal energy is low (cf. Figs. 3a,b). Before the GRU layer, these frames would be assigned to the pause cluster or, even worse, to the wrong speaker cluster. Note that since the grouping success is only computed based on frames sorted into one of the speaker clusters, erroneously assigning many low-level speech frames to the pause cluster can lead to relatively high grouping success scores even at the input of the GRU layer.

Adaptation to DoA changes: We observed in our experiments that the grouping success for the frames directly following a position change depends on whether the position change coincides with a speech pause or not. As an example, in Fig. 3d the second switch around $\tau = 5.4\,\text{s}$ coincides with a switch from a pause to a speech signal. COSPA is able to identify this change immediately. If the change had happened at $\tau = 4.9\,\text{s}$, COSPA would not know that the position of the desired speaker has changed, because this will only become apparent when the speaker starts talking. Unsurprisingly and similar to acoustic source localization, adaptation to a new

DoA cannot be expected during pauses. A similar effect can be observed for a switch between two DoAs during speech activity. In some scenarios the assignment of the input features after this DoA switch changes almost instantaneously, while for the output features changing the cluster assignment takes a few frames when speech is present. This is expected behavior, since the memory provided by the GRU will remember the spatial setup of the scene and hence will wait for more evidence for a DoA change before reassigning the features. The delay in changing the features is beneficial for increasing robustness against directional interferers, e.g., when applying COSPA to target speaker extraction and speech enhancement in the presence of multiple sources. (Note that this behavior depends also on the differences in amplitude of the signals and cannot be observed in Fig. 3d at $\tau = 2.9\,\text{s}$ as the signal amplitudes differ notably.)

Source independence of features: To validate that the features generated by the GRU layer indeed reflect different spatial setups and not different speakers as they are present in DST-clean, we also test COSPA$_{\text{WGN}}$ on DST-1Spk and DST-1Pos, which contain signals from a single speaker with changing position or from two speakers at the same position respectively (cf. Sec. IV-A). As can be seen in Table I, for DST-1Spk the clustering scores for $\mathbf{h}_{\text{out}}$ are good even though only one speaker is present. On the other hand, the grouping success for $\mathbf{h}_{\text{out}}$ for DST-1Pos is really low, which shows that the two speakers at the same position cannot be distinguished by the features of the GRU layer. These results show that COSPA indeed reacts to a change of spatial setup and not to the characteristics of a different speaker.

Apart from the results shown here, we find that COSPA can also adapt to a different number of position changes during testing than it was trained on. This underscores the flexibility of such a neural spatiospectral filter.

Robustness against noise: So far, the results have been presented for COSPA tested on signals without any background noise. To demonstrate how well the spatial information is captured in the GRU layer in the presence of noise and for the task of noise reduction rather than just phase alignment, we provide the clustering test results for COSPA$_{\text{WGN}}$, which was trained for a wide range of SNR levels, for test sequences with various SNR levels in Table I. It can be seen that for all SNRs both the grouping success and the cluster density are notably better for the output than for the input features of the GRU layer. As expected, the grouping success decreases for lower SNRs. The number of frames assigned to the pause cluster also decreases from input to output of the GRU layer, which points to more discriminative features and a good feature memory in the GRU layer.

In summary it can be stated that the spatial information exploited for spatiospectral filtering with COSPA is processed by the GRU layer in the compandor and clearly expressed in the output features of this GRU layer. From the exemplary character of COSPA's single GRU layer in an otherwise memoryless neural network, we can infer that the GRU layer indeed implements a noise-robust and source signal-independent phase equalization for coherent multichannel signals.

V. Conclusion

In this paper, we investigated, through the example of COSPA, the capabilities of a neural spatiospectral filter to represent and exploit spatial information. We found that the complex-valued GRU layer in COSPA's compandor processes the spatial information and that the features obtained at the output of the GRU layer provide discriminative information about the DoA of a signal. Different from a DSB and as long as only one point-like speech source is active, COSPA does not require information about the DoA of the signal to adapt to positional changes of the source. Furthermore, we showed that clustering of input and output features is an appropriate tool for assessing the capabilities of a neural network layer for the task of spatial filtering.

References

[1] X. Zhang, Z.-Q. Wang, and D. Wang, "A speech enhancement algorithm by iterating single- and multi-microphone processing and its application to robust ASR," in *IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2017, pp. 276–280.

[2] Z.-Q. Wang and D. Wang, "All-Neural Multi-Channel Speech Enhancement," in *Proc. Interspeech*, 2018, pp. 3234–3238.

[3] 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. Audio, Speech and Lang. Process.*, vol. 28, pp. 3080–3094, 2020.

[4] Y. Masuyama, M. Togami, and T. Komatsu, "Consistency-aware multi-channel speech enhancement using deep neural networks," in *IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2020, pp. 821–825.

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

[6] M. M. Halimeh and W. Kellermann, "Complex-valued spatial autoencoders for multichannel speech enhancement," in *IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2022, pp. 261–265.

[7] A. Li, W. Liu, C. Zheng, and X. Li, "Embedding and beamforming: All-neural causal beamformer for multichannel speech enhancement," in *IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2022, pp. 6487–6491.

[8] 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 Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2017, pp. 271–275.

[9] X. Xiao et al., "Deep beamforming networks for multi-channel speech recognition," in *IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2016, pp. 5745–5749.

[10] B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, and M. Bacchiani, "Neural network adaptive beamforming for robust multichannel speech recognition," in *Interspeech*, 2016.

[11] K. Tesch and T. Gerkmann, "Insights into deep non-linear filters for improved multi-channel speech enhancement," *arXiv:2206.13310*, 2022.

[12] K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *2014 Conf. Empirical Methods in Natural Lang. Process. (EMNLP)*, 2014, pp. 1724–1734.

[13] 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 *IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2021.

[14] K. Tesch and T. Gerkmann, "Spatially Selective Deep Non-linear Filters for Speaker Extraction," *arXiv:2211.02420*, 2022.

[15] S. Lloyd, "Least squares quantization in pcm," *IEEE Trans. Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[16] J. Garofolo et al., "Timit acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, Nov 1992.

[17] E. Habets, "Room impulse response generator," Tech. Rep., Technische Universiteit Eindhoven, The Netherlands, 2006.