Array Agnostic Multi-channel Speech Presence Probability Estimation

Shuai Tao*‡, Kaixuan Yang†‡, Stijn Kindt†, Jesper Rindom Jensen*, Mads Græsbøll Christensen* and Nilesh Madhu†
†*IDLab, Ghent University-imec*, Ghent, Belgium
**Department of Electronic Systems*, *Aalborg University*, Aalborg, Denmark
*stao@es.aau.dk, †kaixuan.yang@ugent.be, †stijn.kindt@ugent.be, *jrj@es.aau.dk, *mgc@es.aau.dk, †nilesh.madhu@ugent.be
‡ Equal contribution

***Abstract*—In this work, a novel array-agnostic approach is proposed for multi-channel speech presence probability (MC-SPP) estimation. A neural architecture used in our previous work for array-fixed MC-SPP estimation is adapted to accommodate a variable number of microphone channels and guarantee permutation invariance of the inputs. Specifically, convolution and Transformer-based layers are modified to perform channel-wise spectral and temporal processing, followed by Mean Pooling for channel fusion. Transform-Average-Concatenate layers are inserted to effectively aggregate array-level information added to channel-wise independent features. The previously proposed modified minimum variance distortionless response beamformer is then cascaded to produce spatially filtered outputs. Our benchmarking results demonstrate that the proposed approach achieves performance highly comparable to the array-fixed counterpart on known array geometries, while generalizing better to unseen array geometries. Notably, under microphone index permutation conditions, our method significantly outperforms the array-fixed approach, maintaining a much lower complexity in terms of model size and MACs.**

***Index Terms*—array-agnostic approach, multi-channel speech presence probability, MVDR beamforming, Transform-Average-Concatenate**

I. Introduction

For a fixed microphone array with known geometry and constant microphone spacing and number, most existing multi-channel speech enhancement methods effectively reduce noise and restore speech [1], [2]. These methods are widely used in applications like hearing aids and voice communication [3], [4]. However, since the fixed array cannot be modified, specifically the spacing and the number of microphones, it poses significant challenges for algorithm transplantation. To address this issue, the array-agnostic method has recently garnered considerable research interest [5]–[7].

One commonly used multi-channel speech enhancement approach is the minimum variance distortionless response (MVDR) beamforming [8] which can preserve the target speech while minimizing the background noise [9]. To perform MVDR beamforming, the noise power spectral density (PSD) matrix and the steering vector are required. However, estimating these statistics is particularly challenging in complex acoustic environments, such as those with low signal-to-noise ratios (SNR) and significant reverberation. To achieve accurate statistics estimation, the multi-channel speech presence probability (MC-SPP) [10] leverages spatial information to precisely detect speech components, providing a soft decision for speech presence and absence. Recently, to further enhance multi-channel speech performance, deep neural networks (DNNs) have been employed to estimate the MC-SPP. This estimate serves as a mask in a modified MVDR beamformer, resulting in superior performance compared to baselines [11]. However, since the DNN model is explicitly designed and trained with an array-fixed geometry, DNN-based MC-SPP estimation cannot be performed with agnostic arrays without retraining the model for the specific geometry.

In this work, we extend our previously proposed array-fixed approach [11] by introducing a new DNN model designed to estimate the MC-SPP from unknown geometries. Firstly, we introduce the Transform-Average-Concatenate (TAC) layer [12], which can share information between channels in a permutation-invariant manner, enabling the DNN model to accommodate a variable number of microphone channels and ensure input permutation invariance. Additionally, given the high performance of the Transformer-based layers in [13] for extracting time-frequency information, the T-Transformer and F-Conformer are introduced. Therefore, the DNN model consists of the convolution, TAC, and Transformer-based layer to jointly process spatial and time-frequency information in the agnostic array. Finally, Mean Pooling is employed for channel fusion. Experiments were conducted on simulated acoustic datasets, enabling us to benchmark performance on both known and unseen array geometries. For evaluation, the array-agnostic MC-SPP estimate is applied to guide modified MVDR beamforming proposed in [11], and a set of evaluation metrics is used to evaluate our proposed method performance, including speech enhancement performance and model complexity.

II. Signal model and problem formulation

Given a microphone array with arbitrary geometry, the observed signal in a noisy and reverberant environment, in the short-time Fourier transform (STFT) domain, is given by

$$\mathbf{y}(k,l) = \mathbf{x}(k,l) + \mathbf{x}_r(k,l) + \mathbf{n}(k,l), \tag{1}$$

where $k \in [0,\dots,K-1]$ is the frequency index, $l \in [0,\dots,L-1]$ is the time frame index, $\mathbf{y}(k,l) =$

$[Y_1(k,l),\ldots,Y_M(k,l)]^T$, M is the number of microphones, $\mathbf{x}(k,l) = [X_1(k,l),\ldots,X_M(k,l)]^T$ is the direct speech, $\mathbf{x}_r(k,l) = [X_1(k,l),\ldots,X_M(k,l)]^T$ is the reverberant speech, and $\mathbf{n}(k,l) = [N_1(k,l),\ldots,N_M(k,l)]^T$ is the background noise. This work aims to extract the direct speech from the observed signal so the signal model can be expressed by

$$\mathbf{y}(k,l) = \mathbf{x}(k,l) + \mathbf{v}(k,l), \tag{2}$$

where $\mathbf{v}(k,l) = \mathbf{x}_r(k,l) + \mathbf{n}(k,l)$ is the background interference.

MVDR beamforming is performed to extract the target speech, $X_1(k,l)$, from the observed signal. Firstly, using the covariance subtraction method [14], the steering vector can be obtained by

$$\mathbf{d}(k,l) = \frac{\mathbf{\Phi}_{xx}(k,l)\mathbf{e}_1}{\mathbf{e}_1^H\mathbf{\Phi}_{xx}(k,l)\mathbf{e}_1}, \tag{3}$$

where $\mathbf{e}_1 = [1,0,\ldots,0]^T$ is an M dimensional selection vector and $\mathbf{\Phi}_{xx}(k,l) = \mathrm{E}[\mathbf{x}(k,l)\mathbf{x}(k,l)^H]$ is the clean speech PSD matrix.

Subsequently, the MVDR beamforming weights are given by

$$\mathbf{h}(k,l) = \frac{\mathbf{\Phi}_{vv}^{-1}(k,l)\mathbf{d}(k,l)}{\mathbf{d}^H(k,l)\mathbf{\Phi}_{vv}^{-1}\mathbf{d}(k,l)}. \tag{4}$$

where $\mathbf{\Phi}_{vv}(k,l) = \mathrm{E}[\mathbf{v}(k,l)\mathbf{v}(k,l)^H]$ is the noise PSD matrix.

Finally, with the MVDR weights, the MVDR beamforming can be performed as

$$\hat{X}_1(k,l) = \mathbf{h}^H(k,l)\mathbf{y}(k,l), \tag{5}$$

where $\hat{X}_1(k,l)$ is the enhanced speech, i.e., an estimate of the desired speech signal at the first microphone.

III. SPP-based Statistics Estimation

Since, in (4), $\mathbf{\Phi}_{vv}(k,l)$ and $\mathbf{\Phi}_{xx}(k,l)$ are required, the MC-SPP [10] can be employed to estimate these statistics. With two hypotheses: $\mathcal{H}_0$ represents speech absence and $\mathcal{H}_1$ represents speech presence, the observed signal can be defined as: $\mathbf{y}(k,l) = \mathbf{v}(k,l)$, and $\mathbf{y}(k,l) = \mathbf{x}(k,l) + \mathbf{v}(k,l)$, respectively. The likelihood function of speech and noise can be derived assuming that both components follow a multivariate Gaussian distribution and are statistically independent [10].

In this way and using the Bayes' theorem, the *a posteriori* MC-SPP is given by

$$p(k,l) = \left\{1 + \frac{q(k,l)}{1-q(k,l)}[1+\xi(k,l)]\exp\left[-\frac{\beta(\mathrm{k,l})}{1+\xi(\mathrm{k,l})}\right]\right\}^{-1}, \tag{6}$$

where $p(k,l) = p[\mathbf{y}(k,l)|\mathcal{H}_1]$, $q(k,l)$ is the *a priori* multi-channel speech absence probability (MC-SAP), $\xi(k,l)$ is the *a priori* signal-to-noise ratio (SNR) which is defined as

$$\xi(k,l) = \mathrm{tr}[\mathbf{\Phi}_{vv}^{-1}(k,l)\mathbf{\Phi}_{xx}(k,l)], \tag{7}$$

and $\beta(k,l)$ is defined as

$$\beta(k,l) = \mathbf{y}^H(k,l)\mathbf{\Phi}_{vv}^{-1}(k,l)\mathbf{\Phi}_{xx}(k,l)\mathbf{\Phi}_{vv}^{-1}(k,l)\mathbf{y}(k,l). \tag{8}$$

In [11], one DNN model is employed to estimate $p(k,l)$ to improve its estimation accuracy. During training, the actual noise and clean speech are used to compute $q(k,l)$, $\xi(k,l)$, and $\beta(k,l)$ for the learning target. Then, at inference time, the array-fixed MC-SPP estimate $\hat{p}(k,l)$ is obtained from the output of the trained DNN model.

Given $\hat{p}(k,l)$, the noise PSD matrix estimate, $\hat{\mathbf{\Phi}}_{vv}(k,l)$, and the clean speech PSD matrix estimate, $\hat{\mathbf{\Phi}}_{xx}(k,l)$, are recursively updated over time [11]. The so-obtained estimates are then used in (4), to compute an estimate of the MVDR beamforming weights, $\hat{\mathbf{h}}(k,l)$. As shown in [11], a set of modified MVDR weights $\mathbf{h}_m(k,l)$ can be derived using $\hat{p}(k,l)$ as

$$\mathbf{h}_m(k,l) = \hat{p}(k,l)\hat{\mathbf{h}}(k,l). \tag{9}$$

Like the conventional MVDR beamforming, the modified MVDR weights, $\mathbf{h}_m(k,l)$, can then be applied for beamforming in (5).

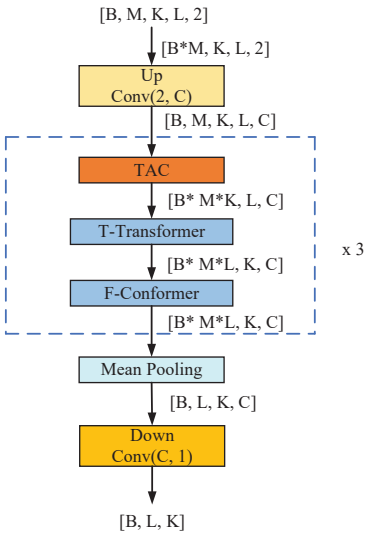


Fig. 1: Proposed model structure consisting of convolution (Conv), TAC, T-Transformer, and F-Conformer layers.

IV. Proposed Method

In this work, a novel DNN model that can operate with an unknown microphone geometry is proposed to achieve array-agnostic MC-SPP estimation and then guide MVDR beamforming. Fig. 1 depicts the proposed neural architecture derived from the DeFT-AN [15], which showed the best performance for MC-SPP estimation in [11].

The neural network incorporates the real and imaginary parts of the STFT-domain signal as input features. To accommodate a variable number of microphone channels as input, an input convolution (Up-Conv) performs spectral and temporal processing of the input features independently for each channel. Correspondingly, channel-wise mean pooling is applied before the output convolution (Down-Conv) for fusion. All stacked dual-path transformer blocks consistently process speech features in a channel-wise manner, while two major adaptations of neural architecture are considered to improve computational efficiency.

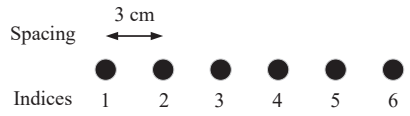


Fig. 2: Uniform Linear Array of 6 Microphones (ULA-6).

Firstly, compared to the original DeFT-AN [15], there is no need to leverage dense convolutional blocks to extract multi-channel spatial information. Instead, TAC layers [7] are inserted in place of the dense blocks for spatial aggregation through linear transformation, channel-wise averaging, and concatenation operations. Secondly, we swap the roles of the time and frequency modules—replacing the time-domain Conformer with a Transformer, and the frequency-domain Transformer with a Conformer. This architectural change empirically enhances performance. The resulting Time-Frequency Attentive block is repeated three times in a cascading fashion, enabling efficient latent-space modeling of speech patterns.

Using the proposed DNN model, the array-agnostic MC-SPP estimate $\tilde{p}(k,l)$ can be obtained. Then, the Kullback-Leibler [16] divergence is employed to optimize the model parameters during training [11]:

$$\mathcal{L}(p(k,l),\tilde{p}(k,l)) = p(k,l)\log\left(\frac{p(k,l)}{\tilde{p}(k,l)}\right). \tag{10}$$

Given $\tilde{p}(k,l)$ and the modified MVDR weights in (9), MVDR beamforming can be performed with an agnostic array.

V. Experimental settings

A. Datasets and Acoustic Parameters

For the training and validation datasets, clean reading speeches were sourced from the DNS Challenge [17]. The testing sources were obtained from the TSP database [18]. Specifically, 7,000 and 3,000 samples of 2-second duration (totaling 3.89 hours and 1.67 hours, respectively) were used for training and validation. For testing, 160 samples were prepared, each with a duration of 10 seconds, accounting for a total of 0.44 hours. Noise samples were sourced from the Audioset [19], Freesound [20], and Demand [21] datasets, and were rendered as an isotropic (diffuse) noise field [22].

All experiments are performed based on data simulated according to the signal model in (2). Pyroomacoustics[1] is used to generate acoustic scenarios for training, validation, and testing. As shown in Fig.2, a basic Uniform Linear Array (ULA) is assumed in general, with a target speaker source in the broadside region (i.e., a limited region in front of the array). The room dimensions and source position change randomly for each sample within a limited range. Detailed configurations of the acoustic parameters are presented in Table I.

[1]https://pyroomacoustics.readthedocs.io/en/pypi-release/index.html

TABLE I: Configurations of Acoustic Parameters

Room size	Length: $\mathcal{U}(3,5)$ m; width: $\mathcal{U}(7,9)$ m; height: $\mathcal{U}(3,4)$ m
Microphone Array	Linear array with 4 microphones
Array position	First microphone: [1.5, 2, 1.7] 3 cm distance with others
Source position	$[\mathcal{U}(1.4,1.7), \mathcal{U}(2.5,3), 1.7]$ m
RT_{60}	$\mathcal{U}(0.2,0.5)$ s
Input SNR	$\mathcal{U}(-10,10)$ dB

*$\mathcal{U}(a,b)$ stands for uniformly sampling over the interval $[a,b]$.

TABLE II: Microphone Array Settings with Various Number of Microphones for Training, Validation, and Evaluation

# Mics	Mics' Indices	
	Training & Validation	Evaluation
2	[1, 2]	[1, 2], [1, 3], [1, 4], [1, 5], [1, 6]
3	[1, 2, 3]	[1, 2, 3], [1, 2, 4], [1, 3, 4], [1, 2, 5], [1, 3, 6]
4	[1, 2, 3, 4]	[1, 2, 3, 4], [1, 2, 4, 5], [1, 2, 3, 6], [1, 2, 4, 6], [1, 2, 5, 6]

B. Benchmarks and Training Procedure

The proposed array agnostic approach (Agnostic) is compared to its counterpart for an array-fixed geometry. For the array-fixed method (Fixed), DeFT-AN [15] showed the best performance in [11], therefore, it is used to estimate array-fixed MC-SPP as the state-of-the-art baseline. Additionally, an ablation study to investigate the impact of the TAC layer (Agnostic (No TAC)) is also conducted.

To train and validate the DNN models for the array-fixed and array-agnostic methods, considered microphone array settings are shown in Table II, and explained as follows:

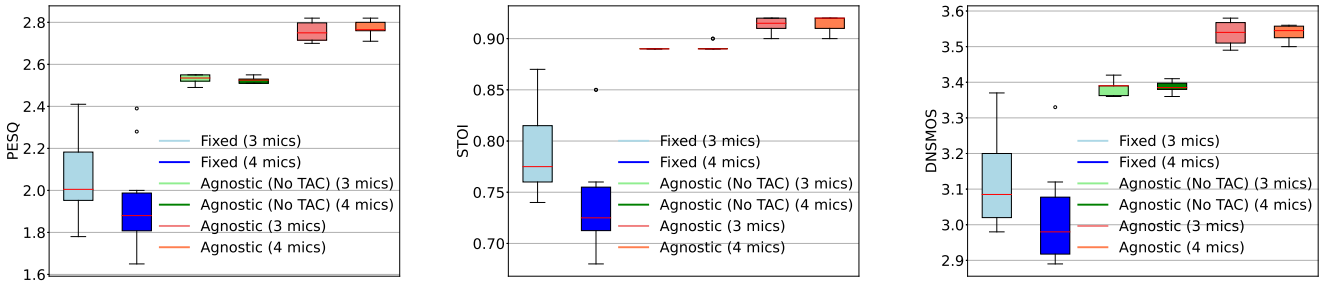
- **Array-fixed** Model: Sub-arrays of 2, 3, and 4 microphones are selected from the ULA-6 for **independent training** of **three** DNN models.
- **Array-agnostic** Model: Sub-arrays of 2, 3, and 4 microphones are randomly selected from the ULA-6 for **multitask training** of **one** DNN model.

When testing for known and unseen array geometries, we take the following into account:

- **Known** array geometries: Testing samples are generated using the same array geometries and indexes as the training data.
- **Unseen** array geometries: While the number of microphones remains the same as in the training data, the spacing is different. The indexing of the microphones can be:
  - **In Order**: Microphones are arranged sequentially according to their indexes.
  - **With Permutation**: Two random shuffles of microphone indices are considered and evaluated for cases involving 3 and 4 microphones.

Since the first microphone is chosen as the reference microphone, it will remain consistent in any testing geometry. This ensures the benchmarking process is straightforward and maintains generality, due to the translational and reflectional symmetry of the sub-arrays on ULA-6.

As for audio pre-processing, speech signals sampled at 16 kHz are transformed to the STFT domain with a 16 ms Hamming window and 8 ms overlaps, resulting in $K = 129$ for



(a) PESQ (b) STOI (c) DNSMOS

Fig. 3: Permutation test results. For the 3 and 4 microphone conditions, two shuffled versions of the microphone indices were used, resulting in a total of 10 Unseen-With Permutation testing arrays. Box plot distribution reflects different array geometries.

frequency dimension. As short sequence of $L = 5$ is used for temporal modelling as in [14], with a latent feature dimension of $C = 6$. For training of the DNN models, Adam optimizer [23] with a learning rate of 0.01 is deployed for minimizing loss in (10) with batch size $B = 16$. To prevent overfitting, the weight decay is set to 0.00001. According to the validation loss curve, the best models are saved within 200 epochs.

C. Evaluation Metrics

To evaluate the enhancement performance, perceptual evaluation of speech quality (PESQ) [24], short-time objective intelligibility (STOI) [25], and deep noise suppression mean opinion score (DNSMOS) [26] are measured. For model complexity evaluation, the Python library `ptflops` [2] is used to measure the number of parameters (Params) and Multiply-ACcumulate operations (MACs) per second.

VI. Results and Discussion

This section presents the speech enhancement results and model complexity assessment to draw conclusions.

Comparison on Known and Unseen-In Order: Table III presents the numerical results for the speech enhancement performance evaluation on both known and unseen arrays, with microphone indices ordered as in Table II. Compared to the array-fixed MC-SPP estimation-based method, the proposed method generally demonstrates highly comparable performance. While the PESQ scores are slightly lower for known arrays, the proposed method achieves almost same STOI and DNSMOS scores. Notably, our proposed method demonstrates better generalization in unseen-in order scenarios and consistently outperforms the array-fixed approach. The array-agnostic method exhibits improved performance as the number of microphones increases across all evaluation metrics, indicating effective utilization of multi-channel spatial information.

Comparison on Unseen-With Permutation: Subsequently, Fig. 3 illustrates the results of the permutation test, revealing

[2]https://pypi.org/project/ptflops/

TABLE III: Comparison of Speech Enhancement Performance on Known and Unseen-In Order Arrays.

# Mics	Index	Methods	PESQ	STOI	DNSMOS
		Unprocessed	2.00	0.85	2.95
2	Known	Fixed	**2.79**	**0.91**	**3.54**
		Agnostic (No TAC)	2.45	0.88	3.35
		Agnostic	2.71	0.90	3.53
	Unseen-In Order	Fixed	2.52	0.88	3.36
		Agnostic (No TAC)	2.52	**0.89**	3.33
		Agnostic	**2.58**	0.88	**3.42**
3	Known	Fixed	**2.87**	**0.92**	3.55
		Agnostic (No TAC)	2.49	0.89	3.38
		Agnostic	2.81	**0.92**	**3.56**
	Unseen-In Order	Fixed	2.73	**0.91**	3.48
		Agnostic (No TAC)	2.54	0.89	3.38
		Agnostic	**2.74**	**0.91**	**3.52**
4	Known	Fixed	**2.90**	**0.93**	**3.55**
		Agnostic (No TAC)	2.52	0.90	3.41
		Agnostic	2.82	0.92	**3.56**
	Unseen-In Order	Fixed	**2.76**	**0.91**	3.48
		Agnostic (No TAC)	2.53	0.89	3.38
		Agnostic	**2.76**	**0.91**	**3.53**

* In the Unseen-In Order condition, the metric score is the average.

that the array-fixed method failed to enhance speech. In contrast, the array-agnostic method maintained high performance, attributed to the permutation invariance of its neural architecture leading to consistent performance of mask-based MVDR beamformer. Additionally, smaller variations across different microphone groups were observed for the array-agnostic methods, further demonstrating their robustness with respect to microphone spacing.

Ablation Study of the TAC Layers: In the ablation study of the array-agnostic model without TAC layers, the model (No TAC) consistently exhibits lower performance across all test conditions. This suggests an inability to capture spatial information without the TAC layers, as it essentially learns gradient descent on an 'aggregated' single-channel representation.

Comparison on Model Complexity: Finally, Table IV demonstrates that the array-agnostic method consistently exhibits a smaller size and lower computational cost compared to the array-fixed method. Regarding MACs, the computational cost increases proportionally with the number of microphones for configurations with 2, 3, and 4 microphones. Notably, the

TABLE IV: Comparison of Model Complexity

# Mics	Methods	Params	MACs
2	Fixed	630.74 k	53.43 G/s
	Agnostic (No TAC)	104.04 k	18.41 G/s
	Agnostic	107.41 k	18.95 G/s
3	Fixed	666.96 k	56.36 G/s
	Agnostic (No TAC)	104.04 k	27.42 G/s
	Agnostic	107.41 k	28.23 G/s
4	Fixed	670.16 k	56.61 G/s
	Agnostic (No TAC)	104.04 k	36.43 G/s
	Agnostic	107.40 k	37.51 G/s

introduction of TAC layers results in a significant performance improvement, while only causing a marginal increase in complexity.

VII. Conclusion

In this work, we proposed a novel array-agnostic approach for MC-SPP estimation. We adapted a neural architecture from our previous work on array-fixed MC-SPP estimation to accommodate a variable number of microphone channels and ensure permutation invariance of the inputs. Experiments conducted on simulated acoustic datasets confirmed three key findings. Firstly, in known scenarios, the proposed method achieved high speech enhancement performance comparable to the array-fixed method, while maintaining lower model complexity. Secondly, in unseen scenarios with agnostic array geometry, our method demonstrated robustness against changes in microphone geometry. Notably, under microphone index permutation conditions, our method significantly outperformed the array-fixed method. Finally, the ablation study revealed that the TAC layer significantly improves MC-SPP estimation accuracy by effectively capturing spatial information from variable input channels.

References

[1] A. Herzog and E. A. Habets, "Signal-dependent mixing for direction-preserving multichannel noise reduction," in *2021 29th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 96–100.

[2] K. Tesch and T. Gerkmann, "Insights into deep non-linear filters for improved multi-channel speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 563–575, 2022.

[3] K. Ngo, M. Moonen, S. H. Jensen, and J. Wouters, "A flexible speech distortion weighted multi-channel wiener filter for noise reduction in hearing aids," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 2528–2531.

[4] W. Jin, M. J. Taghizadeh, K. Chen, and W. Xiao, "Multi-channel noise reduction for hands-free voice communication on mobile phones," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 506–510.

[5] H. Taherian, S. E. Eskimez, T. Yoshioka, H. Wang, Z. Chen, and X. Huang, "One model to enhance them all: array geometry agnostic multi-channel personalized speech enhancement," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 271–275.

[6] A. Mannanova, K. Tesch, J.-M. Lemercier, and T. Gerkmann, "Meta-learning for variable array configurations in end-to-end few-shot multi-channel speech enhancement," in *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2024, pp. 200–204.

[7] T. Yoshioka, X. Wang, D. Wang, M. Tang, Z. Zhu, Z. Chen, and N. Kanda, "Vararray: Array-geometry-agnostic continuous speech separation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6027–6031.

[8] J. Benesty, *Microphone array signal processing*. Springer Verlag, 2008.

[9] 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 Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 3080–3094, 2020.

[10] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2159–2169, 2011.

[11] S. Tao, P. Mowlaee, J. R. Jensen, and M. G. Christensen, "Learning-based multi-channel speech presence probability estimation using a low-parameter model and integration with mvdr beamforming for multi-channel speech enhancement," in *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2024, pp. 100–104.

[12] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6394–6398.

[13] D. Lee and J.-W. Choi, "DeFT-AN: Dense frequency-time attentive network for multichannel speech enhancement," *IEEE Signal Processing Letters*, vol. 30, pp. 155–159, 2023.

[14] S. Chakrabarty and E. A. 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.

[15] D. Lee and J.-W. Choi, "Deft-an: Dense frequency-time attentive network for multichannel speech enhancement," *IEEE Signal Processing Letters*, vol. 30, pp. 155–159, 2023.

[16] J. R. Hershey and P. A. Olsen, "Approximating the kullback leibler divergence between gaussian mixture models," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, 2007, pp. IV–317.

[17] C. K. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "ICASSP 2021 deep noise suppression challenge," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6623–6627.

[18] P. Kabal, "TSP speech database," *McGill University, Database Version*, vol. 1, no. 0, pp. 09–02, 2002.

[19] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2017, pp. 776–780.

[20] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 411–412.

[21] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics*, vol. 19, no. 1. AIP Publishing, 2013.

[22] E. A. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 2911–2917, 2008.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[25] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2010, pp. 4214–4217.

[26] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6493–6497.