All Neural Kronecker Product Beamforming for Speech Extraction with Large-scale Microphone Arrays

Weixin Meng[1,2], Xiaoyu Li[1], Andong Li[3], Jian Li[1,2], Xiaodong Li[1,2], Chengshi Zheng[1,2]
[1]Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences [2]University of Chinese Academy of Sciences [3] Tencent AI Lab, Beijing, China
{cszheng}@mail.ioa.ac.cn

Summary

- Existing frame-wise neural beamformers for speech extraction tasks can obtain promising performance in relatively high signal-to-noise ratio (SNR) scenarios using small microphone arrays, while they still suffer from performance degradation in relatively low SNR environments, e.g., SNR<-5 dB.
- We propose an all-neural beamformer based on Kronecker product decomposition, denoted by NeuKP-BF, for large-scale microphone arrays.
- The core idea is to incorporate the high spatial resolution of large microphone arrays and the powerful non-linear modeling capability of deep neural networks to improve speech extraction performance in challenging environments.
- In this paper, to reduce the feature representation redundancy and improve the interpretability, we used the Kronecker product rule to decompose the original large-scale array into two small virtual subarrays, and beamformers for the two subarrays were then designed and merged finally.

Introduction

Neural Beamforming Methods:

- T-F masking based beamforming methods: This category of methods utilize a hybrid cascade structure that combines neural networks with traditional signal processing beamformers
- Neural spatio-spectral filters: The neural spatio-spectral filters employs an end-to-end strategy, in which DNNs serve as a spatio-spectral filter to explicitly or implicitly extract and fuse the spectral and spatial cues and then output the target speech
- All-neural beamforming methods: The beamforming weight vectors are estimated in an end-to-end manner, and the beamforming operations are implemented frame-by-frame. In addition, if the network is designed and trained to obey some beamforming optimizing criterions, these all neural beamformers can guarantee a performance upper bound, while maintaing the interpretability of the trained models.

Proposed Method:

- Despite the significant progress made by the aforementioned beamforming methods, their performance significantly degrades in relatively low signal-to-noise ratio (SNR) scenarios, e.g., SNR<-5 dB.
- A reasonable solution to this problem is to utilize DNN-based beamformers deployed on large-scale microphone arrays, consisting of dozens or hundreds of microphones.
- We propose a fully neural beamformer based on Kronecker Product (KP) decomposition for speech extraction in relatively low SNR environments using large-scale microphone arrays inspired by traditional KP based beamformers.
- This dimension reduction method can be seen as a pre-beamforming process of the received signal, which can effectively improve the SNR of the input features before being fed into the network, thereby strengthening the ability of the beamformer to extract spectral features in low SNR scenarios.

Signal Model and Problem Formulation

KP-MVDR:

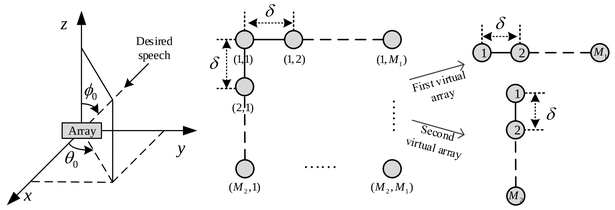

Fig.1: Uniform rectangular array

- Some assumptions:

$$\mathbf{h}(k) = \mathbf{h}_1(k) \otimes \mathbf{h}_2(k), \tag{1}$$
$$\mathbf{w}(k,l) = \mathbf{w}_1(k,l) \otimes \mathbf{w}_2(k,l). \tag{2}$$

- Optimization problem:

$$\min_{\mathbf{w}(k,l)} E\left\{ |(\mathbf{w}_1(k,l) \otimes \mathbf{w}_2(k,l))^H \mathbf{v}(k,l)|^2 \right\}, \quad s.t. \quad \mathbf{w}_1^H(k,l)\mathbf{h}_1(k) = 1, \quad s.t. \quad \mathbf{w}_2^H(k,l)\mathbf{h}_2(k) = 1. \tag{3}$$

- The iterative solution of KP-MVDR:

$$\mathbf{w}_{1,i}(k,l) = \frac{\mathbf{R}_{\mathbf{v}_{1,i-1}}^{-1}(k,l)\mathbf{h}_1(k)}{\mathbf{h}_1{}^H(k)\mathbf{R}_{\mathbf{v}_{1,i-1}}^{-1}(k,l)\mathbf{h}_1(k)}. \tag{4}$$

$$\mathbf{w}_{2,i}(k,l) = \frac{\mathbf{R}_{\mathbf{v}_{2,i}}^{-1}(k,l)\mathbf{h}_2(k)}{\mathbf{h}_2{}^H(k)\mathbf{R}_{\mathbf{v}_{2,i}}^{-1}(k,l)\mathbf{h}_2(k)}, \tag{5}$$

NeuKP-BF

Network Structure:

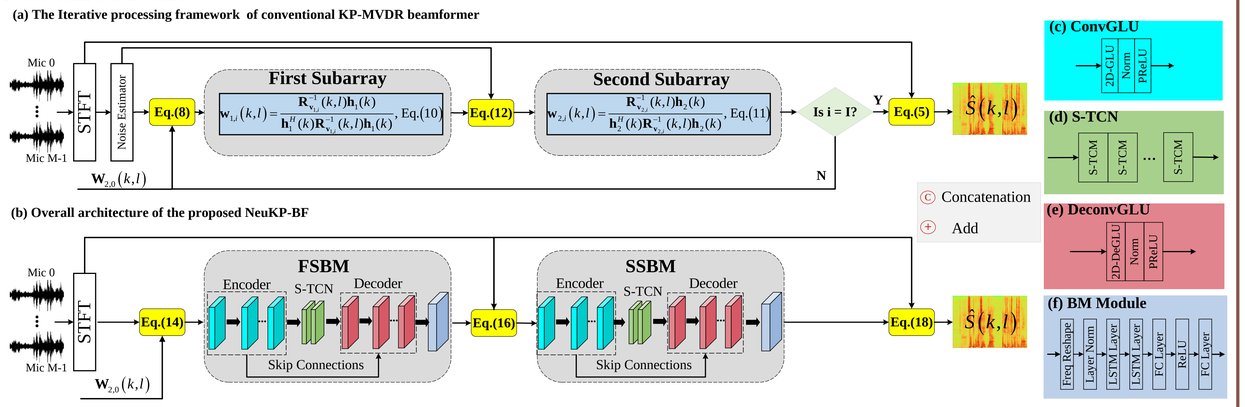

Fig.2: The diagram of the conventional KP-MVDR and the proposed NeuKP-BF

- The overall architecture consists two major parts, namely the first virtual subarray beamformer module (FSBM) and the second virtual subarray beamformer module (SSBM).
- In the FSBM, we aimed to simulate the process of the conventional KP-MVDR beamformer to estimate the weight vector of the first subarray with the initialized beamformer weight vector of the second virtual subarray $\mathbf{w}_{2,0}(k,l)$.
- In the SSBM, we aimed to simulate the process of the conventional KP-MVDR beamformer to estimate the weight vector of the second virtual subarray with the estimated beamformer weight vector of the first virtual subarray $\mathbf{w}_{\mathrm{FS},1}(k,l)$.

Loss Function:

- The whole loss function contains two terms, which can be expressed as:

$$\mathcal{L} = \alpha\mathcal{L}_{sub} + \beta\mathcal{L}_{final} \tag{6}$$

- where $\mathcal{L}_{sub}$ and $\mathcal{L}_{final}$ denote the loss function of the outputs of the FSBM and the SSBM, respectively.

Experimental Resutls

Dataset:

- We also generate a large noisy-clean training set with 100 hours duration based on the DNS-Challenge dataset, where around 90,000 provided noises are utilized and the SNR ranges from -20dB to 10dB.
- All utterances are convolved with 100,000 provided room impulse responses, whose T_{60} ranges from 0.1 to 0.3 second.

Ablation Study:

Table 1: Ablation study on NeuKP-BF

Model	Param.(M)	Metrics		
		PESQ	ESTOI	DNSMOS
Noisy	-	1.48	0.36	1.20
FSBM only	1.17	2.29	0.63	2.17
FSBM++	2.46	2.48	0.68	2.31
DAS+FSBM	1.08	2.25	0.63	2.15
NeuKP-BF	2.18	**2.80**	**0.77**	**2.50**

Comparison with baseline systems

Table 2: Quantitative comparisons with advanced baselines

Model	Causality	DOA	Param.(M)	SNR(dB)					
				-15	-10	-5	0	5	10
Noisy	-	-	-	1.12/0.08/1.09	1.14/0.16/1.09	1.20/0.27/1.10	1.50/0.42/1.11	1.79/0.55/1.25	2.11/0.67/1.53
SMF-MPDR	✗	✓	-	1.36/0.29/1.11	1.54/0.40/1.15	1.82/0.50/1.25	2.14/0.63/1.44	2.40/0.72/1.81	2.71/0.81/2.34
KPOB-MLDR	✗	✓	-	1.40/0.33/1.18	1.70/0.46/1.23	1.97/0.57/1.39	2.31/0.70/1.72	2.57/0.78/2.01	2.90/0.85/2.51
KP-MVDR	✗	✓	-	2.39/0.67/1.89	2.57/0.72/2.25	2.63/0.75/2.26	2.79/0.78/2.43	2.83/0.79/2.55	2.94/0.80/2.72
Oracle-MVDR	✗	✗	-	**2.68/0.84/2.32**	**2.78/0.86/2.52**	**2.80/0.86/2.56**	2.86/**0.87**/2.63	2.91/0.88/2.65	3.08/0.90/2.77
TaylorBF	✓	✗	2.46	1.48/0.34/1.62	1.96/0.53/1.87	2.32/0.67/2.21	2.80/0.79/2.53	3.02/0.86/2.73	3.32/0.90/2.91
COSPA	✓	✗	3.20	1.49/0.10/1.10	1.47/0.17/1.14	1.67/0.30/1.27	2.00/0.48/1.64	2.31/0.62/2.00	2.62/0.73/2.38
GCRN (72ch)	✓	✗	9.79	1.17/0.16/1.60	1.25/0.28/1.57	1.69/0.45/1.78	2.16/0.63/2.18	2.48/0.73/2.47	2.83/0.81/2.68
Conv-TasNet (72ch)	✓	✗	9.97	1.30/0.28/1.74	1.64/0.44/1.99	1.88/0.57/2.24	2.24/0.68/2.44	2.45/0.75/2.60	2.71/0.81/2.73
NeuKP-BF	✓	✓	2.18	1.88/0.51/1.80	2.33/0.67/2.12	2.71/0.78/2.46	**3.09**/0.86/**2.74**	**3.27/0.89/2.86**	**3.53/0.92/3.01**

Conclusions

- The proposed system used the Kronecker product rule to decompose the original large array into two small virtual subarrays, an DNN-based beamformer for each subarray was designed and merged to extract the target speech.
- The proposed system surpasses previous state-of-the-art speech enhancement systems in low SNR scenarios, indicating the superiority of the proposed decomposition paradigm.