A Study of Multichannel Spatiotemporal Features and Knowledge Distillation on Robust Target Speaker Extraction

Yichi Wang[1], *Jie Zhang*[1], *Shihao Chen*[1], *Weitai Zhang*[2], *Zhongyi Ye*[2], *Xinyuan Zhou*[1], *Lirong Dai*[1]
[1]NERC-SLIP, University of Science and Technology of China, Hefei, China
[2]iFLYTEK Research, iFLYTEK CO. LTD., Hefei, China.

Introduction

- Target speaker extraction (TSE) using direction of arrival (DOA) has a wide range of applications. Due to the inherent phase uncertainty, existing TSE methods often suffer from speaker confusion within specific frequency bands. Imprecise DOA clues can also deteriorate the TSE performance.
- In this work, we propose several new multichannel spatiotemporal features. The narrow-band Conformer [1] model is applied in combination with the designed features for TSE. We apply knowledge distillation to improve the model robustness against DOA mismatches.

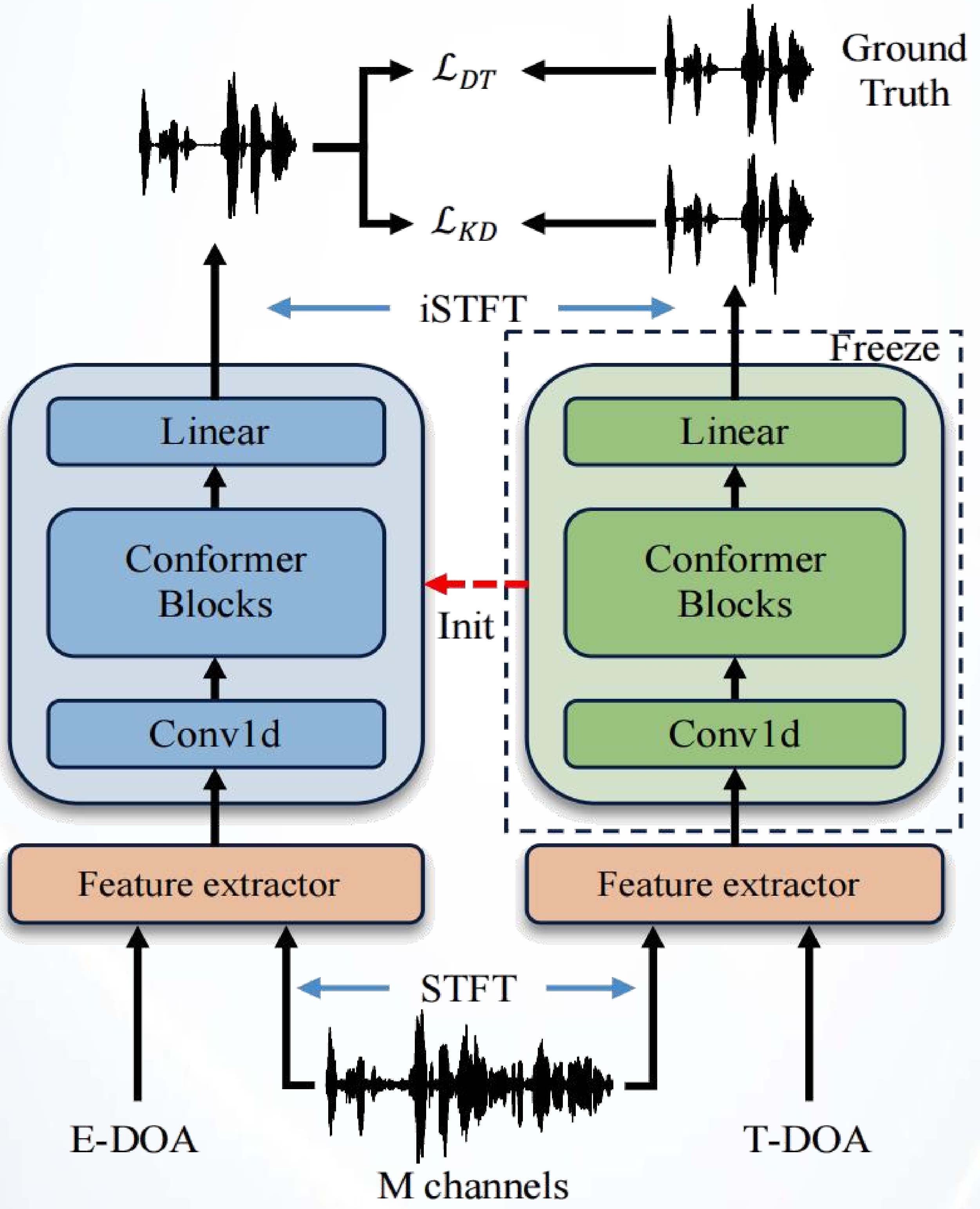

Figure 1. The paradigm of the proposed DOA-assisted TSE model

Visualization: Each cluster of lines denotes a distinct speaker, and we can theoretically prove that each line corresponds to a specific frequency bin, where the frequency-dependent slope is related to the time difference of arrival (TDOA).

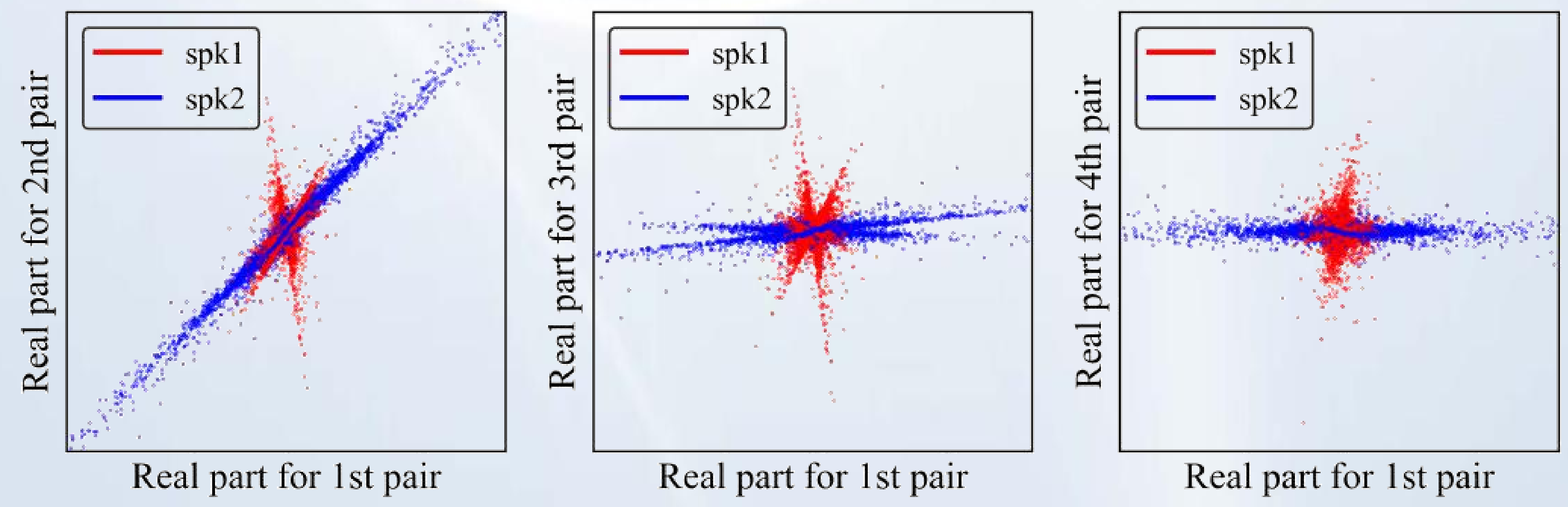

Figure 2. ΔSTFT across different microphone pairs (with a DOA gap of 67°).

Multichannel Spatiotemporal Features

- **Target-dependent Phase Difference (TPD) & Cosine directional function (CDF)** [2]:

$$\mathrm{TPD}_{p_1,p_2}(f) = \frac{2\pi f}{2c(F-1)} f_s d_p cos\varphi_p cos\theta_p$$

$$\mathrm{CDF}_{p_1,p_2}(t,f) = cos(\mathrm{IPD}_{p_1,p_2}(t,f) - \mathrm{TPD}_{p_1,p_2}(f))$$

- **Sine directional function (SDF):**

$$\mathrm{SDF}_{p_1,p_2}(t,f) = sin(\mathrm{IPD}_{p_1,p_2}(t,f) - \mathrm{TPD}_{p_1,p_2}(f))$$

- **ΔShort-Time Fourier Transform(STFT)**:

$$\Delta\mathrm{STFT}_{p_1,p_2} = Y_{p_1}(t,f) - Y_{p_2}(t,f)$$

- **Spatial Correlation(SC):**

$$\mathrm{SC}_{p_1,p_2}(t,f) = Y_{p_1}(t,f) * Y_{p_2}^{*}(t,f)$$

- **Normalized Spatial Correlation (NSC):**

$$\mathrm{NSC}_{p_1,p_2}(t,f) = \frac{\mathrm{SC}_{p_1,p_2}(t,f)}{\mathrm{SC}_{p_1,p_1}(t,f) * \mathrm{SC}_{p_2,p_2}(t,f)}$$

Experiments and Results

Dataset: We consider a single interfering speaker, (i.e., I = 1), and employ a circular array consisting of M = 6 microphone with a radius of 5 cm. The dataset configuration keeps fully the same as the fixed microphone geometry used for FaSNet-TAC [3].

Table 1: The TSE performance of different feature combinations in terms of SI-SNRi (in dB) and PESQ scores (last two columns).

	Feature Combination	Speaker angle(°)				Overlap ratio(%)				Average	WB-PESQ	NB-PESQ
		<15	15-45	45-90	>90	<25	25-50	50-75	>75			
	Mixture	-5.35	-5.37	-5.61	-5.48	-5.61	-5.25	-5.6	-5.4	-5.47	1.29	1.66
	CDF(1-fold)	0.88	14.11	16.50	18.22	16.54	14.34	12.38	10.48	13.43	2.83	3.27
	CDF(6-fold)	5.76	15.16	16.69	17.76	17.62	15.16	13.54	11.91	14.56	2.88	3.34
	CDF(1-fold)+IPD	10.49	17.92	19.06	20.19	21.06	18.63	15.74	14.45	17.46	3.20	3.60
CDF(6-fold)+	ILD	6.87	15.33	16.64	17.56	17.91	15.38	13.80	11.90	14.75	2.86	3.34
	SC	10.13	17.82	19.12	20.21	21.25	18.23	15.99	14.13	17.40	3.21	3.60
	NSC	10.77	18.19	19.19	20.30	20.96	18.68	16.28	14.74	17.66	3.23	3.62
	IPD	10.52	18.29	19.43	20.47	21.21	18.57	16.51	14.77	17.76	3.26	3.65
	IPD+ILD	10.41	17.03	17.82	18.72	20.22	17.25	15.12	13.38	16.49	3.13	3.56
	IPD+SC	11.10	18.40	19.50	20.61	21.17	18.84	16.61	15.18	17.95	3.28	3.66
	IPD+NSC	10.58	18.75	19.85	20.95	21.74	19.08	16.59	15.18	18.14	3.29	3.67
	IPD+SC+NSC	11.50	18.78	19.84	20.91	21.83	19.30	16.95	15.13	18.30	3.30	3.68
	IPD+SDF(6-fold)	10.37	**19.27**	**20.36**	**21.39**	**22.16**	19.06	17.16	15.69	18.52	**3.38**	**3.73**
	IPD+ΔSTFT	11.72	18.86	19.98	21.03	21.82	19.44	17.00	15.48	18.43	3.34	3.70
	IPD+SDF(6-fold)+ΔSTFT	**12.23**	19.15	20.18	21.29	22.04	**19.61**	**17.31**	**15.96**	**18.73**	3.35	3.70

Table 2: The performance in SI-SNRi with DOA mis-matches, where columns and rows denote training and testing conditions, respectively.

Condition	true DOA				[-5°, 5°]		[-10°, 10°]		[-15°, 15°]	
	prop1	**prop2**	**baseline**	**TAC**	**DT**	**KD**	**DT**	**KD**	**DT**	**KD**
true DOA	**18.73**	18.30	17.46	11.17	18.58	**18.63**	18.00	18.31	17.24	18.27
[-5°, 5°]	**17.88**	17.51	16.67	10.60	18.01	**18.11**	17.59	18.02	16.99	17.97
[-10°, 10°]	**15.58**	15.27	14.68	8.79	16.06	16.20	16.53	16.75	16.31	**16.79**
[-15°, 15°]	**12.49**	12.68	12.04	5.93	13.91	13.60	14.74	14.76	15.28	**15.33**
Average	**16.17**	15.94	15.21	9.12	16.64	16.64	16.72	16.96	16.46	**17.09**

Conclusion

- The combination of CDF(6-fold)+IPD+SDF(6-fold)+ΔSTFT achieves the best performance.
- The application of knowledge distillation shows a more clear superiority over other methods in the existence of large DOA mis-matches.
- The proposed spatiotemporal features are compatible with the existing IPD and CDF.

Reference

① C. Quan and X. Li, "NBC2: Multichannel speech separation with revised narrow-band conformer," arXiv preprint arXiv:2212.02076, 2022.
② Z.-Q. Wang and D. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," IEEE/ACM ASLP, 27(2): 457–468, 2018.
③ Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in ICASSP, 2020, pp. 6394–6398.