Enhancing Low-latency Speaker Diarization with Spatial Dictionary Learning

Weiguang Chen[1], Tran The Anh[2], Xionghu Zhong[1], Eng Siong Chng[2]

[1]College of Computer Science and Electronic Engineering, Hunan University, China

[2]School of Computer Science and Engineering, Nanyang Technological University, Singapore

ICASSP 2024 KOREA

PAPER ID: 6144

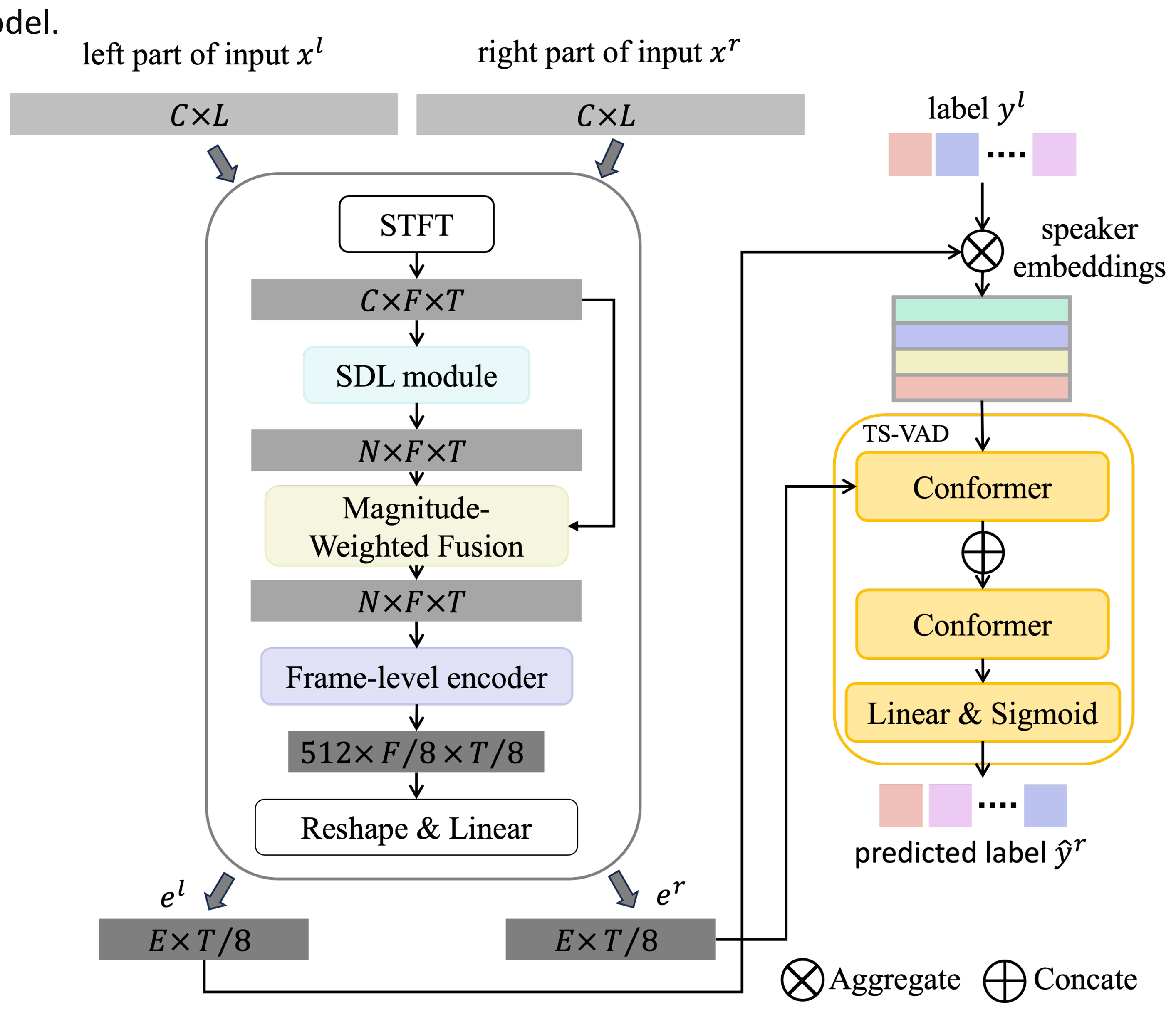
INTRODUCTION

- Speaker diarization serves as a crucial precursor for downstream applications, such as multi-speaker automatic speech recognition.
- Recent advancements have focused on adapting existing frameworks to online diarization by buffering speakers' acoustic representations.
- Despite this, online diarization based on acoustic features encounters several challenges:
  - The latency in acquiring a reliable acoustic embedding for a new speaker can extend up to several seconds.
  - As the number of speakers increases, distinguishing between speakers with similar timbres becomes more challenging.
- To address these issues, this paper introduces a novel spatial dictionary learning method for online speaker diarization.

METHOD

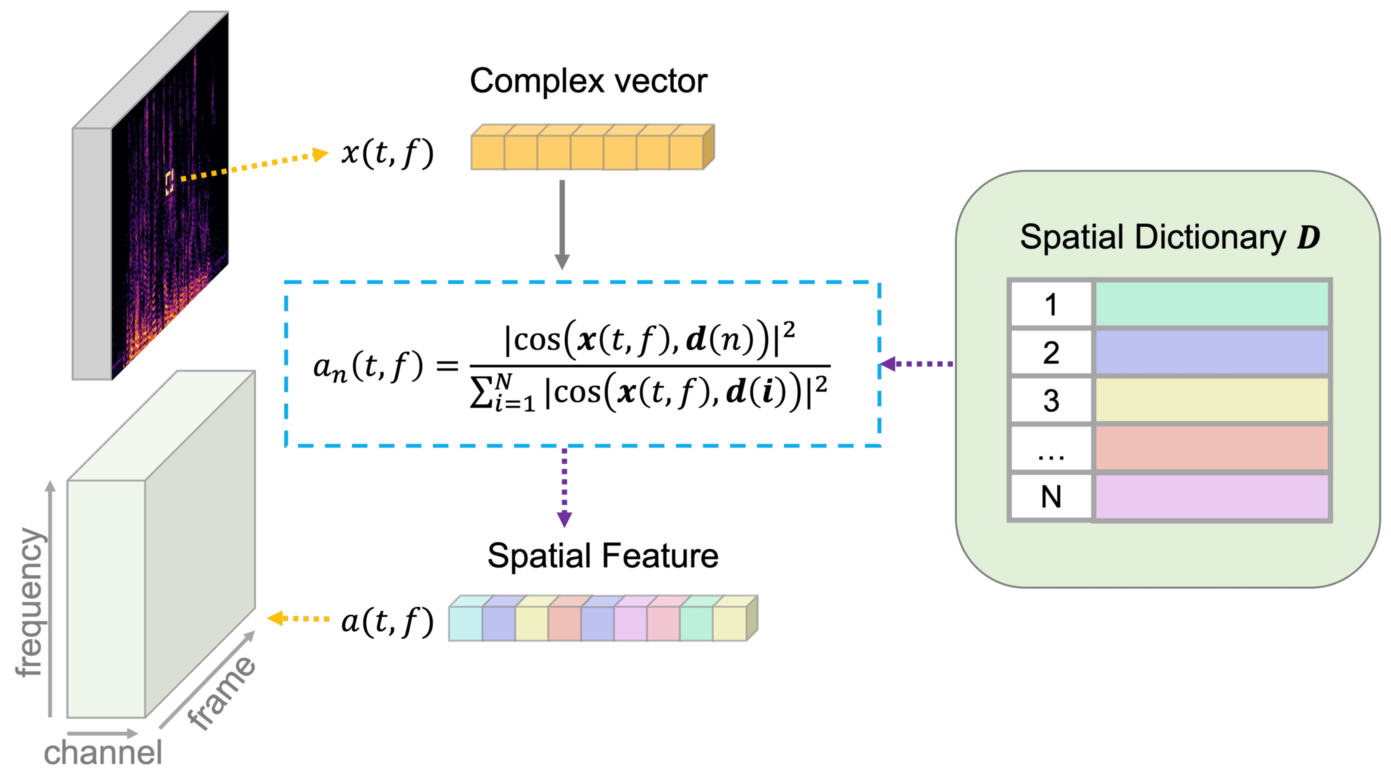
- **Overall Framework --- SDL-TS-VAD**

- During training, the left part of the input is utilized to generate speaker embeddings, which are then used for predicting the right portion via the Conformer based TS-VAD model.



- Two novel modules are proposed --- Spatial dictionary learning (SDL) and Magnitude-weighted fusion (MWF)
  - SDL projects the complex vector at each time-frequency point onto a hypersphere.
  - MWF fuses the magnitude spectrum with the result obtained from SDL.

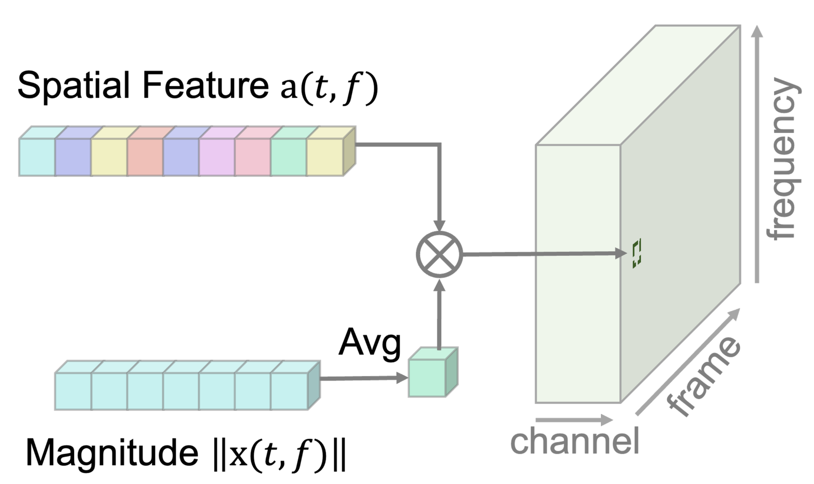
- **Spatial Dictionary Learning**



- The spatial dictionary is composed of a complex tensor with dimensions $N\times M(Size_D\times Channel)$.
- The input at each time-frequency point is a complex vector, encompassing both amplitude and phase information.
- The cosine similarity is calculated between the input complex vector and the items in the spatial dictionary.

- **Magnitude-Weighted Fusion**

- Average pooling is applied to multi-channel magnitude spectra.
- The features from SDL and the averaged magnitude are then fused by multiplication.



RESULTS AND CONCLUSION

- We evaluated our SDL-TS-VAD on the AliMeeting[1] dataset. The results for Diarization Error Rate (DER) are presented in Table 1.

Table 1. The performance of different systems on Alimeeting Eval and Test set in terms of DER (%). † and ‡ denote the 1st-ranked and 2nd-ranked methods from Alimeeting challenge respectively. Results marked with ^ are sourced from [2] .

Model	Type	Input Channels	Eval				Test
			FA (%)	MISS (%)	SC (%)	DER (%)	DER (%)
Official baseline[1]	offline	single-	-	-	-	15.24	15.60
BeamformIt [3] +VBx [4] ^	offline	multi-	0.00	13.10	0.47	13.57	13.51
BeamformIt [3] +VBx [4] +OSD [5] ^	offline	multi-	1.09	4.17	2.84	8.10	9.45
Target-DOA [6] ^	offline	multi-	1.17	3.92	4.15	9.23	11.95
TS-VAD [7] †	offline	single-	1.00	2.50	0.70	4.12	-
TS-VAD [7] †	offline	multi-	1.10	1.10	0.10	2.26	2.98
FFM-TS-VAD [2] ‡	offline	multi-	0.83	2.55	0.26	3.64	5.63
Online TS-VAD [8]	online	single-	-	-	-	**8.14**	**11.42**
SDL-TS-VAD	online	multi-	2.35	2.73	0.95	**6.04**	**6.19**

- We also conducted ablation studies on the SDL and MWF modules, with the results summarized in Table 2. "MagPhase" denotes the combination of magnitude and phase spectra, while "RealImag" signifies the concatenation of real and imaginary spectra.

Table 2. Ablation studies on SDL and MWF

ID	Input Feature	SDL	N	MWF	DER (%)
1	MagPhase	✗	-	✗	9.09
2	RealImag	✗	-	✗	8.55
3	RealImag	✓	32	✗	12.23
4	RealImag	✓	16	✓	6.80
5	RealImag	✓	32	✓	6.19
6	RealImag	✓	64	✓	6.67

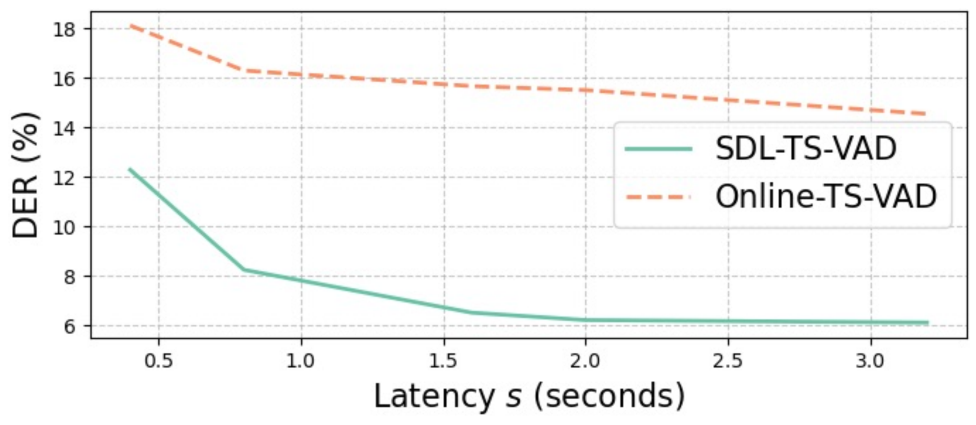
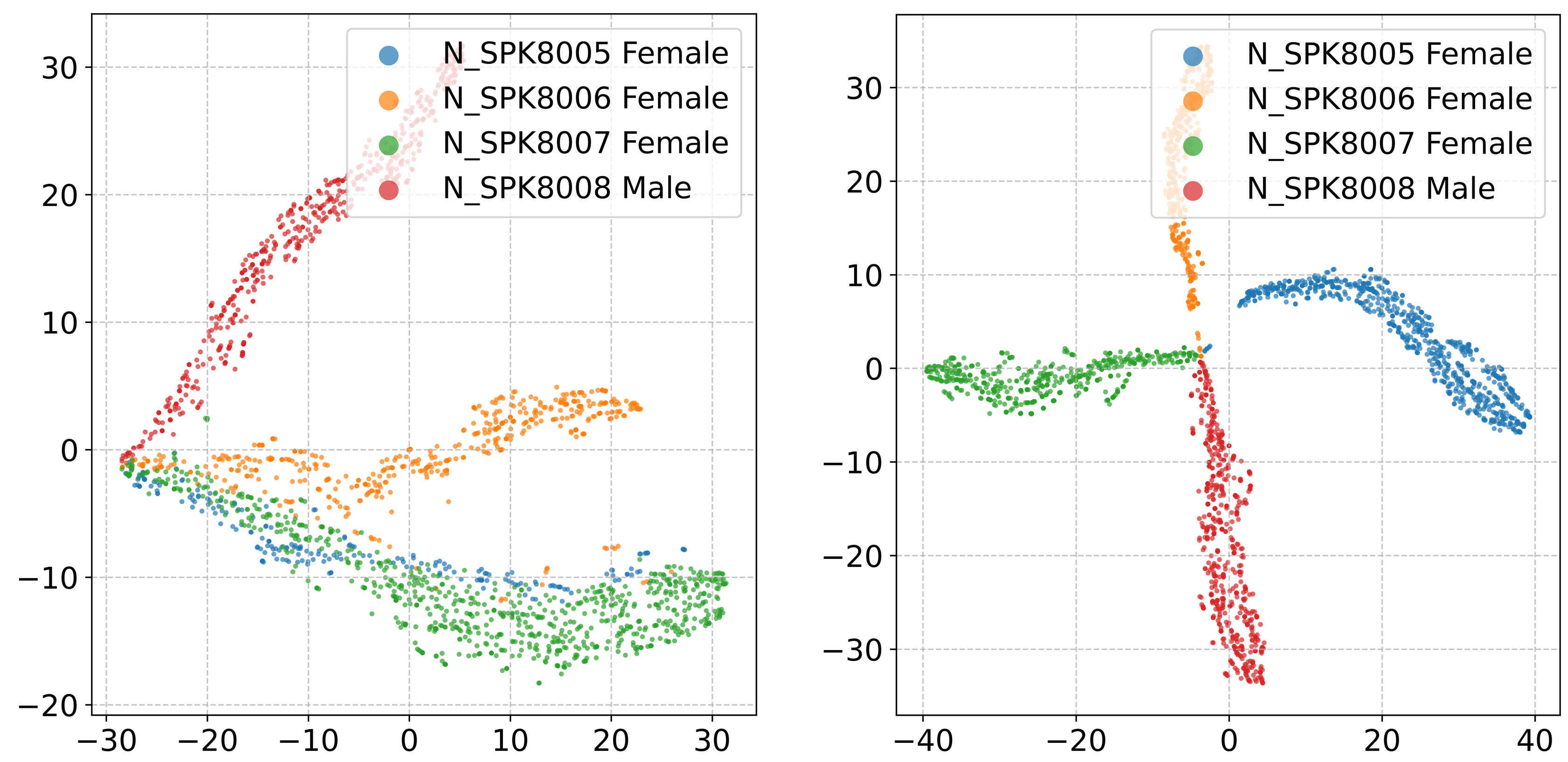


Figure 1. The DERs (%) of our SDL-TS-VAD and online TS-VAD with different latencies.

- We compared our method with re-implemented Online TS-VAD under various latencies, as shown in Figure 1.
- Additionally, we visualized the speaker embeddings aggregated from each block of session "R8002_M8002_MS802" in the Alimeeting test set, as depicted in Figure 2.



(a) Online TS-VAD (b) SDL-TS-VAD

Figure 2. The t-SNE plots of aggregated speaker embeddings from each block of session "R8002 M8002 MS802".

- The experiments demonstrated that our method achieved a significant improvement compared to the single-channel method, and SDL-TS-VAD exhibited performance comparable to the second-ranked offline method of the AliMeeting challenge[1].
- In the future, we will extend our method to scenarios where speakers can move in the meeting and the number of speakers is greater.

REFERENCES

[1] Fan Yu, Shiliang Zhang, Yihui Fu, Lei Xie, Siqi Zheng, Zhihao Du, Weilong Huang, Pengcheng Guo, Zhijie Yan, Bin Ma, et al., "M2met: The icassp 2022 multi-channel multi-party meeting transcription challenge," in *Proc. ICASSP*, 2022, pp.6167–6171.

[2] Naijun Zheng, Na Li, Xixin Wu, Lingwei Meng, Jiawen Kang, Haibin Wu, Chao Weng, Dan Su, and Helen Meng, "The cuhk-tencent speaker diarization system for the icassp 2022 multi-channel multi-party meeting transcription challenge," in *Proc. ICASSP*, 2022, pp. 9161–9165.

[3] Xavier Anguera, Chuck Wooters, and Javier Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[4] Mireia Diez, Luk´aˇs Burget, Federico Landini, and Jan Cernocky, "Analysis of speaker diarization based on Bayesian hmm with eigenvoice priors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 355–368, 2019.

[5] Herve Bredin and Antoine Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *Proc. Interspeech*, 2021, pp. 3111–3115.

[6] Naijun Zheng, Na Li, JianWei Yu, Chao Weng, Dan Su, XunYing Liu, and Helen Meng, "Multi-channel speaker diarization using spatial features for meetings," in *Proc. ICASSP*, 2022, pp. 7337–7341.

[7] Weiqing Wang, Xiaoyi Qin, and Ming Li, "Cross-channel attention-based target speaker voice activity detection: Experimental results for the m2met challenge," in *Proc. ICASSP*, 2022, pp. 9171–9175.

[8] Weiqing Wang, Qingjian Lin, and Ming Li, "Online target speaker voice activity detection for speaker diarization," in *Proc. Interspeech*, 2022, pp. 1441–1445.