TS-VAD+: Modularized Target-Speaker Voice Activity Detection for Robust Speaker Diarization

Tran The Anh, Azmat Adnan, Wu Yihao, and Chng Eng Siong
College of Computing and Data Science, Nanyang Technological University, Singapore
E-mail: tranthea001@e.ntu.edu.sg, azmat.adnan@ntu.edu.sg, yihao005@e.ntu.edu.sg, aseschng@ntu.edu.sg

***Abstract*—Target-speaker voice activity detection (TS-VAD) improves speaker diarization by modeling speaker activity using prior speaker embeddings. We present TS-VAD+, a modular and scalable extension of TS-VAD enhancing robustness and generalization across diverse acoustic and domain conditions. TS-VAD+ integrates a transformer-based architecture, replacing traditional BLSTM, and is modularized to adapt any large-scale self-supervised speech representations or pretrained speaker encoders, improving embedding quality. To further strengthen diarization performance, TS-VAD+ is employed with two additional modules: a profile enhancement module employing a denoising model to suppress noise and embedding leakage, and a post-processing module refining speech activity boundaries via an external VAD. Our evaluation on the AliMeeting dataset and the cross-domain DIHARD-III benchmark shows TS-VAD+ outperforms conventional TS-VAD. Notably, our enhancements reduce DIHARD-III DER from 19.53% to 17.97%. Scalability analysis indicates optimal performance when limiting the maximum number of speakers to eight. The TS-VAD+ code is released, supporting community research on this modularized approach.**[1]

I. Introduction

Speaker diarization, the task of determining "who spoke when" in an audio recording, is a fundamental component for processing multi-speaker interactions. Its utility spans a wide range of speech applications like meeting transcription, broadcast media analysis, and conversational AI, where accurately segmenting and attributing individual speaker turns is crucial for ASR and speaker recognition. However, this task presents significant challenges in precisely identifying speaker transitions, accurately attributing speech segments, and robustly handling adverse acoustic conditions such as overlapping speech, background noise, and varying channel characteristics.

Traditionally, diarization relied on multi-stage clustering approaches, encompassing Voice Activity Detection (VAD), speaker feature extraction (e.g., i-vectors, x-vectors), and subsequent clustering (e.g., Agglomerative Hierarchical Clustering (AHC), Spectral Clustering). The Bayesian HMM clustering of x-vector sequences (VBx) [1] system offers a robust probabilistic framework for segmenting and clustering speaker turns. Despite their widespread use, these multi-stage approaches frequently suffer from error propagation across sequential components, notably struggling with accurately separating speakers during overlapping speech, and require extensive hyperparameter tuning for optimal performance.

[1]The code is available at https://github.com/TonnyTran/TSVAD

End-to-End Neural Diarization (EEND) models directly map audio features to diarization outputs, overcoming multi-stage pipeline limitations. Early EENDs, like BLSTM-EEND [2], used permutation-free objectives for speaker label assignment. Subsequent advancements, such as Encoder-Decoder Based Attractor (EDA) calculation for EEND (EEND-EDA) [3], significantly improved speaker representation learning. More recent research explores advanced neural architectures including DiaPer [4], EEND-M2F [5], and novel Mamba-based Segmentation Models [6]. While EEND simplifies pipelines and often performs well, their scalability with many speakers and robustness in highly noisy or reverberant environments remain critical research areas.

Target-Speaker Voice Activity Detection (TS-VAD) explicitly models speech activity for specific speakers using enrollment embeddings, significantly enhancing performance in multi-speaker overlapping speech where conventional VAD struggles [7]. Subsequent TS-VAD variants include Seq2Seq-TS-VAD [8], PET-TSVAD [9], EDA-TS-VAD [10], and ANSD-MA-MSE [11]. However, existing TS-VAD systems face challenges: their performance degrades under diverse acoustic conditions and noise due to poor domain generalization. They are also highly sensitive to initial speaker embedding quality, with issues like embedding leakage or noisy embeddings degrading performance. Achieving optimal Diarization Error Rates (DER) further necessitates refined speech activity boundaries and additional post-processing.

To address these shortcomings and enhance target-speaker diarization robustness and generalization, we propose TS-VAD+, a modular, adaptable framework. Extending original TS-VAD, this work includes key enhancements: a transformer-based architecture for improved temporal modeling; large-scale self-supervised speech representations; pretrained speaker encoders to boost embedding discriminability and acoustic resilience; a profile enhancement module explicitly utilizes a powerful denoising model (e.g., DEMUCS [12]) to purify speaker embeddings and mitigate cross-speaker leakage; a dedicated post-processing module leverages external, highly accurate VAD systems (e.g., Pyannote VAD [13]) to refine speech activity segmentation, reducing false alarms and ensuring precise boundary detection. TS-VAD+'s modularity provides exceptional flexibility. We conduct comprehensive experimental evaluations on AliMeeting and DIHARD-III benchmarks, and perform scalability analysis for optimal speaker configurations.

979-8-3315-7206-8/25/$31.00

II. Prior Work

Target Speaker Voice Activity Detection (TS-VAD) [10] is a technique designed to identify and segment the speech activities of specific target speakers within a given audio stream. It operates based on the profiles of these target speakers, such as i-vectors, which serve as representations of their unique characteristics. The typical TS-VAD architecture consists of three primary modules: a Convolutional Neural Network (CNN) based encoder, a Bidirectional Long Short-Term Memory (BLSTM) based independent-speaker-detection (ISD) module, and a BLSTM-based joint-speaker-detection (JSD) module.

The CNN-based encoder extracts high-level embeddings from the audio input, capturing features crucial for speaker identity. These embeddings are then fed into the ISD module, which processes each target speaker independently, leveraging their profile to accurately identify their individual speech activities. Concurrently, the JSD module performs joint modeling to predict the activities of all speakers simultaneously, capturing both temporal and cross-speaker correlations. However, this BLSTM-based joint modeling can limit capturing long-range temporal dependencies and complex inter-speaker dynamics, impacting scalability and performance in dynamic or dense speaker environments.

III. The Proposed Framework

We present TS-VAD+, a modular TS-VAD with several key enhancements, including a transformer-based TS-VAD design for improved temporal modeling, robust speech representations from large-scale self-supervised models, advanced pretrained speaker encoders, a profile enhancement module for purifying speaker embeddings, and a post-processing module for refining speech activity segmentation. The overall architecture of TS-VAD+ is illustrated in Fig.1, showcasing its modular design.

A. Overall Architecture

The TS-VAD+ framework detects simultaneous voice activity for up to N target speakers from raw audio, optionally using an initial RTTM file. The pipeline begins with an SSL speech representation module, extracting frame-level embeddings $E \in \mathbb{R}^{T \times E}$ (T: sequence length, E: embedding dimension). Concurrently, a pretrained speaker encoder generates N speaker profiles, P_n (dimension P).

Selecting target speech for the fixed N speakers is crucial. During training, ground truth extracts active speaker segments; if fewer than N speakers are active, we either randomly augment profiles from other utterances or pad with zero tensors. In inference, an x-vector clustering system selects target speakers, as ground truth is unavailable.

Each speaker's profile P_n is duplicated T times and concatenated with E, forming N mixed embeddings $Q_n \in \mathbb{R}^{T \times (E+P)}$. These mixed embeddings integrate speaker identity and audio content for detection.

TS-VAD+ significantly enhances temporal modeling and sequence-level reasoning through a transformer-based architecture. This approach, unlike simpler recurrent or feed-forward networks, excels at capturing long-range dependencies via self-attention. As shown in Fig.1, the central detection mechanism uses two cascaded transformer models. First, a 6-layer independent-speaker-detection module processes each Q_n independently, using shared weights for efficiency. It outputs N distinct feature sequences, $F_n \in \mathbb{R}^{T \times F}$ (F: new feature dimension). These N features are concatenated into a combined $\mathbb{R}^{T \times (N \cdot F)}$ tensor, then fed into the second transformer: a joint speaker detection module. This module leverages contextual information across all speakers for a final, refined decision. The binary output, $T \times N$, indicates per-frame speaker activity. This hierarchical transformer design enables comprehensive understanding of multi-speaker interactions and precise temporal localization. The model is trained by minimizing the sum of binary cross-entropy losses across all speakers.

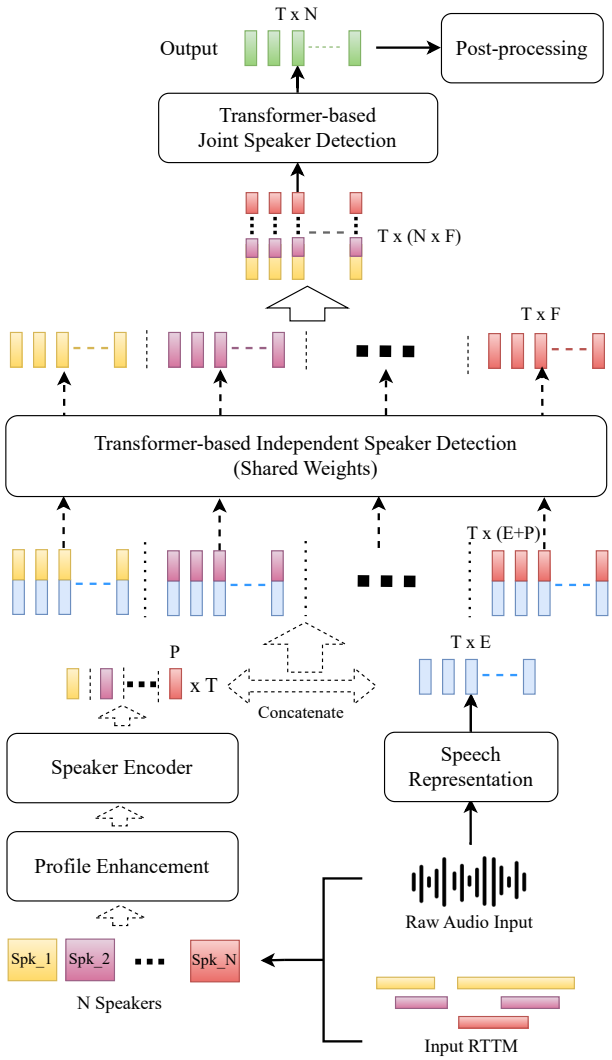

Fig. 1. Modularizing TS-VAD+ model

B. Speech Representation

To ensure robustness, TS-VAD+ incorporates large-scale self-supervised speech representations. These advanced mod-

els, trained on vast unlabeled speech data, learn highly robust and discriminative features less susceptible to noise or environmental variations. Examples include:

- wav2vec2 [14]: Learns representations by masking audio and predicting content using a transformer.
- UniSpeech [15]: Extends wav2vec2 with multi-modal pretraining (speech and text).
- Whisper Encoder [16]: A strong self-supervised representation learner, trained on diverse audio and transcript data.
- WavLM [17]: Builds on HuBERT and wav2vec2, enhancing robustness by focusing on both content and speaker information via a masked prediction task.

The module outputs a $T \times E$ tensor.

C. Speaker Encoder

For speaker profile generation, TS-VAD+ employs well-established, pretrained speaker encoders. Trained on extensive speaker identification/verification datasets, these encoders produce highly discriminative and robust speaker embeddings across various acoustic conditions. Examples include:

- ECAPA-TDNN [18]: An advanced Time-Delay Neural Network (TDNN) for speaker verification, known for robustness with squeeze-and-excitation and channel-wise attention.
- CAM++ [19]: A state-of-the-art Transformer-based speaker embedding model using additive angular margin (AAM) loss.
- WavLM-sv [17]: A specialized WavLM variant for speaker verification, leveraging WavLM's self-supervised representations fine-tuned for speaker-specific features.

The encoder outputs a vector of dimension P for each speaker, representing their unique voice characteristics.

D. Profile Enhancement

TS-VAD+ introduces a critical profile enhancement module utilizing a speech enhancement model to purify speaker embeddings and suppress cross-speaker leakage. Contaminated speaker profiles, common in multi-speaker conversations due to other speakers or noise, can degrade VAD performance. Our module leverages advanced speech enhancement models such as DEMUCS [12] to clean input speaker profiles (Fig.1), isolating the target speaker's vocal characteristics. This results in purer, more discriminative embeddings, leading to accurate and reliable VAD decisions.

E. Post-processing

Finally, TS-VAD+ includes a post-processing module that refines speech activity segmentation and reduces false alarms. The raw joint speaker detection output provides frame-level activity scores, which can be noisy. We integrate external, optimized VAD systems, such as PyAnnote VAD [13], as a refinement layer. This system smooths output, corrects spurious detections, and fills small gaps. This step significantly improves temporal accuracy and reduces false alarms, yielding cleaner, more reliable speech activity segmentation.

IV. Experimental Setup

A. Datasets

1) AliMeeting: The AliMeeting dataset [20] presents a comprehensive resource for evaluating speaker diarization systems in Mandarin meetings, featuring over 118.75 hours of real-world recordings. It captures diverse scenarios with 15-30 minute discussions among 2-4 participants in conference rooms and offices. Offering balanced participant representation and covering various room sizes, noise levels, and microphone-speaker distances, AliMeeting allows for detailed study of the impact of acoustic conditions on performance. Importantly, it incorporates both near-field and far-field recordings.

2) DIHARD-III: The DIHARD III dataset [21] is a rich and diverse corpus of audio recordings across various domains and conditions. The dataset consists of 5-10 minute segments sampled from 11 conversational domains, which cover a wide range of recording equipment, environments, noise levels, speaker demographics and interaction types. The Third DIHARD Challenge includes two tracks: diarization from reference speech activity detection and diarization from scratch. The DIHARD III Challenge is designed to support speaker diarization research. It does not provide a fixed training set, allowing participants to train their systems on any proprietary and/or public data.

3) Simulated Data: For TS-VAD training, a 1000-hour wideband (16kHz) simulated dataset was created, primarily using the Librispeech corpus for speech. Following the methodology from Landini et al. [22] this dataset meticulously mimics speaker turns, silent periods, and overlaps based on RTTM statistics to realistically generate multi-speaker conversations. Each simulated conversation segment ranges from 5 to 20 minutes and features 1 to 10 speakers. To ensure acoustic diversity and realism, the speech is combined with background noise from the MUSAN corpus and convolved with varied room impulse responses, such as those from the Simulated Room Impulse Response Database. Signal-to-Noise Ratios (SNRs) are sampled across 0, 5, 10, 15, and 20 dBs, providing a comprehensive and acoustically rich training resource that closely approximates real-world conversational audio.

B. Configurations

While experimenting on the Alimeeting dataset, TS-VAD+ was trained over 40 epochs (the speech representation model was frozen for the initial 10 epochs). On DIHARD-III, the TS-VAD pipeline was first trained for 30 epochs on a mixture of 1000-hour simulated data, VoxConverse, and MSDWild (with the speech representation model frozen for the initial 10 epochs), then fine-tuned for 20 epochs on the DIHARD-III development set.

Audio was processed at 16 kHz in 4 s (100-frame) windows and fed to the network in batches of 160. On-the-fly augmentation mixed MUSAN noise [23] and simulated RIRs [24] to inject real-world reverberation and clutter. Optimization utilized AdamW (initial LR 1×10^{-4}, 0.9 step decay per epoch), and a binary-cross-entropy-with-logits loss was employed.

TABLE I
TS-VAD+ WITH DIFFERENT SPEECH REPRESENTATIONS AND SPEAKER ENCODERS ON ALIMEETING

ID	Speech Representation	Speaker Encoder	DER (c=0.00s)	DER (c=0.25s)
A1	wav2vec2 Base	ECAPA TDNN	16.9	9.79
A2	wav2vec2 Large	ECAPA TDNN	16.51	9.55
A3	Unispeech	ECAPA TDNN	16.41	9.49
A4	Whisper Encoder	ECAPA TDNN	17.53	10.29
A5	WavLM Large	ECAPA TDNN	16.37	9.49
A6	WavLM Base +	ECAPA TDNN	17.70	10.93
A7	finetuned WavLM Base +	ECAPA TDNN	**15.94**	**9.05**
A8	finetuned WavLM Base +	CAM++	17.67	10.17
A9	finetuned WavLM Base +	WavLM-sv	17.67	10.24
VBx+OSD+VAD [4]			28.84	15.60
PyAnnote3.1 [13]			24.40	15.51
DiaPer [4]			26.27	18.82
Mamba-based Segmentation Model [6]			16.20	-
EEND-M2F [5]			13.20	5.87

TABLE II
TS-VAD+ WITH PROFILE ENHANCEMENT AND POST-PROCESSING ON DIHARD-III

ID	Model	Profile Enhancement	Post Processing	DER	MS	FA	SC
D1	TS-VAD+	-	-	19.53	9.67	7.15	2.72
D2	TS-VAD+	DEMUCS	-	19.39	9.90	6.83	2.67
D3	TS-VAD+	-	VAD	18.11	11.13	4.44	2.54
D4	TS-VAD+	DEMUCS	VAD	17.97	11.36	4.11	**2.49**
TS-VAD [11]				20.12	-	-	-
VBx+OSD+VAD [4]				20.28	10.14	**3.93**	6.15
PyAnnote3.1 [13]				21.70	**8.15**	6.17	7.29
DiaPer [4]				22.77	11.31	5.83	5.58
EEND-EDA [3]				21.94	-	-	-
ANSD-MA-MSE [11]				16.76	-	-	-
Mamba-based Segmentation Model [6]				16.70	-	-	-
EEND-M2F [5]				**16.07**	-	-	-

C. Evaluation metric

The performance of the TS-VAD model was evaluated using the Diarization Error Rate (DER), calculated using the md-eval perl script with two collar conditions: $c = 0.00s$ (oracle boundaries) and $c = 0.25s$ (standard tolerance for speaker transitions). DER is further broken down into three components: Miss Speech (MS - segments where ground truth speech is not detected), False Alarm (FA - segments where non-speech is incorrectly detected as speech), and Speaker Confusion (SC - segments where speech is correctly detected but attributed to the wrong speaker).

V. RESULTS AND ANALYSIS

We comprehensively evaluate TS-VAD+'s effectiveness across three modular aspects. First, we explore the most effective speech representation and speaker encoder configuration using the AliMeeting dataset. Second, we assess the benefits of profile enhancement and post-processing on the challenging DIHARD-III dataset. Finally, we examine TS-VAD+'s scalability by varying target speaker density.

A. Speech Representation and Speaker Encoder

We evaluated TS-VAD+'s performance using a variety of SSL speech representations, specifically wav2vec2 Base[2], wav2vec2 Large[3], Unispeech[4], Whisper Encoder[5], WavLM Large[6], WavLM Base+[7], and a Librispeech-finetuned WavLM Base+[8]. We also assessed its performance with several pretrained speaker encoders, namely ECAPA TDNN[9], CAM++[10], and WavLM-sv[11]. All evaluations were conducted on the AliMeeting dataset. Models were trained on the AliMeeting training set, with optimal configurations determined using the development set and final testing performed on the test set. Due to AliMeeting sessions typically containing 2-4 speakers, we set N to 4. During inference, TS-VAD+ enhances initial coarse speaker segments, which are generated by VBx RTTM, through the application of target-speaker modeling.

As shown in Table I, the best performance is achieved when combining finetuned WavLM Base + features with the ECAPA-TDNN encoder, yielding a DER of 15.94% (c = 0.00s) and 9.05% (c = 0.25s). Other strong configurations include WavLM Large (16.37%, 9.49%) and Unispeech (16.41%, 9.49%), all paired with ECAPA-TDNN. In contrast, models using CAM++ or WavLM-sv encoders exhibit significantly higher DERs, indicating poor compatibility with TS-VAD+.

Compared to recent results, our TS-VAD+ method significantly improves upon VBx+OSD+VAD [4], PyAnnotate3.1 [13], DiaPer [4], and Mamba-based Segmentation Model [6]. Furthermore, TS-VAD+ closely matches EEND-M2F [5] (13.20% DER at c=0.00s, 5.87% at c=0.25s), positioning our approach as a highly competitive solution for diarization.

B. Profile Enhancement and Post-processing

We further assess TS-VAD+ on the DIHARD-III corpus (full set), which presents significant domain mismatch due to its lack of training data and diversity across 11 domains. Our base model is pretrained on a mixture of 1000-hour simulated data, VoxConverse, and MSDWild, then fine-tuned on the DIHARD-III development set. The maximum number of speakers N is set to 8, with a collar size $c = 0.00s$, and no oracle VAD is applied during inference. TS-VAD+ is equipped with finetuned WavLM Base + and ECAPA TDNN, which is the best set up from above experiment.

Table II illustrates that TS-VAD+ surpasses conventional TS-VAD, achieving a 0.59% absolute improvement in DER. We then analyzed the impact of two distinct modules: profile enhancement via DEMUCS, and post-processing with PyAnnote's neural VAD. DEMUCS effectively enhances speaker embedding quality, while the PyAnnote VAD mitigates speech activity detection errors. The combined application of both modules yields the most significant performance gain, reducing the DER from 19.53% to 17.97%. This setup also demonstrates a decrease in false alarms (FA) from 7.15% to 4.11% and a reduction in speaker confusion (SC) from 2.72% to 2.49%.

[2] https://huggingface.co/facebook/wav2vec2-base
[3] https://huggingface.co/facebook/wav2vec2-large
[4] https://huggingface.co/microsoft/unispeech-large-1500h-cv
[5] https://huggingface.co/openai/whisper-large-v3
[6] https://huggingface.co/microsoft/wavlm-large
[7] https://huggingface.co/microsoft/wavlm-base-plus
[8] https://drive.google.com/file/d/1-zlAj2SyVJVsbhifwpTlAfrgc9qu-HDb
[9] https://github.com/TaoRuijie/ECAPA-TDNN
[10] https://www.modelscope.cn/models/iic/speech_campplus_sv_zh-cn_3dspeaker_16k
[11] https://huggingface.co/microsoft/wavlm-base-plus-sv

TABLE III
PERFORMANCE OF TS-VAD+ IN SPECIFIC DOMAINS ON DIHARD-III

ID	Domain	Num speakers	Overlap Ratio (%)	DiaPer	PyAnnote3.1	VBx+OSD+VAD	TS-VAD+
1	audiobooks	1	0.00	**3.32**	4.07	4.52	3.91
2	broadcast_interview	3-4	1.61	18.94	10.56	**8.86**	9.12
3	clinical	2-3	3.21	18.36	24.79	20.51	**16.56**
4	court	3,6-9	1.79	34.42	11.49	**6.59**	9.27
5	cts	2	11.77	12.06	14.44	14.22	**10.93**
6	maptask	2	1.83	**8.21**	10.59	9.63	10.79
7	meeting	3-6	21.30	43.63	38.08	35.67	**31.21**
8	restaurant	4-8	26.77	63.66	49.85	52.05	**46.33**
9	socio_field	2-3	4.53	**15.34**	20.86	18.47	15.81
10	socio_lab	2	3.58	11.02	14.96	**10.50**	11.77
11	webvideo	1-7	17.57	48.72	49.99	48.22	**47.75**
All	All	1-9	9.37	22.77	21.70	20.28	**17.97**

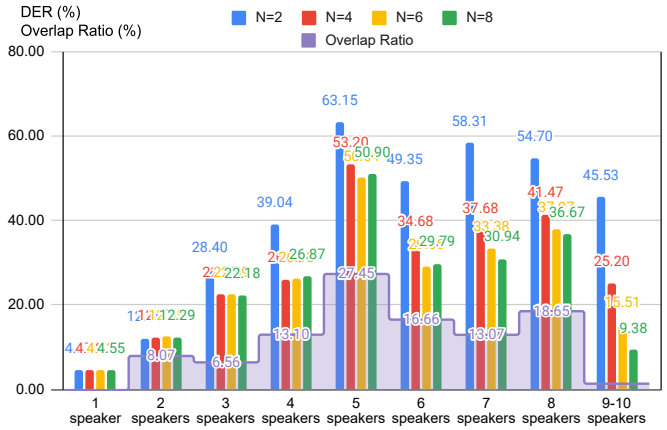

Fig. 2. TS-VAD+ with different maximum number of speakers

While EEND-M2F remains the top-performing model with a DER of 16.07%, our best TS-VAD+ configuration is competitive despite relying on a modular architecture. Furthermore, TS-VAD+ significantly outperforms recent baselines including VBx+OSD+VAD, PyAnnote3.1, and DiaPer, achieving absolute DER improvements of 2.3% to 4.8%. These results underscore the advantage of incorporating DEMUCS and post-processing into TS-VAD+ for enhanced robustness under challenging acoustic and domain conditions.

A domain-wise analysis (Table III) further illustrates the generalization capacity of TS-VAD+ across a variety of real-world scenarios compared to popular systems including Diaper, PyAnnote3.1 and VBx+OSD+VAD. TS-VAD+ consistently outperforms all baselines in most domains, with notable gains in clinical and cts, highlighting its strength in conversational and structured dialogue settings. Especially in highly overlapping, large number of speakers and acoustically complex environments such as meeting, restaurant and webvideo, TS-VAD+ achieves lower DERs than all other methods. These results demonstrate that TS-VAD+ not only delivers strong average performance but also maintains stable effectiveness across diverse and challenging acoustic domains.

C. Scalability based on maximum number of speakers

We assess the scalability of TS-VAD+ on the DIHARD-III evaluation set by varying the maximum number of target speakers $N \in \{2, 4, 6, 8\}$. Recordings are grouped by actual speaker count, and DER is analyzed in relation to speaker overlap. As shown in 2, higher speaker counts generally correspond to higher overlap ratios, for example, the 5-speaker group has an overlap ratio of 27.45%, while the 8-speaker group reaches 18.65% resulting in more challenging diarization conditions with high DER.

With lower N values (e.g., 2 or 4), the model struggles to accommodate complex multi-speaker scenarios, resulting in high DERs. Raising N to 6 yields noticeable improvements, particularly for recordings with more than four speakers. However, setting $N = 8$ delivers the most consistent and significant gains. For instance, DER in the 9–10 speaker group drops from 45.53% at $N = 2$ to 9.38% at $N = 8$, with similar improvements seen in other high-overlap groups. Importantly, $N = 8$ also maintains competitive performance in low-speaker conditions, such as single- and two-speaker recordings, showing no notable degradation compared to smaller N values. These findings confirm that $N = 8$ provides robust coverage for both simple and complex speaker configurations.

VI. CONCLUSIONS

We introduce TS-VAD+, a modular, adaptable framework enhancing target-speaker diarization robustness and generalization. Extending TS-VAD, it employs a transformer-based architecture for temporal modeling and integrates large-scale self-supervised speech representations/pretrained speaker encoders, boosting embedding discriminability and acoustic resilience. TS-VAD+ also includes profile enhancement (purifying embeddings, mitigating leakage) and post-processing (refining speech activity boundaries via external VADs). Its modularity offers exceptional flexibility. Evaluations on AliMeeting and DIHARD-III demonstrate effectiveness; scalability analysis indicates optimal performance up to eight speakers. This highlights modular design and advanced deep learning potential in robust, generalized target-speaker diarization; future work could explore module fusion and performance in extreme noise or more speakers.

VII. ACKNOWLEDGEMENT

The computational work for this article was fully performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg/).

References

[1] F. Landini, J. Profant, M. Diez, and L. Burget, "Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: Theory, implementation and analysis on standard tasks," *Computer Speech & Language*, vol. 71, p. 101 254, 2022.

[2] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," *arXiv preprint arXiv:1909.05952*, 2019.

[3] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and P. García, "Encoder-decoder based attractor calculation for end-to-end neural diarization," *arXiv preprint arXiv:2106.10654*, 2021.

[4] F. Landini, T. Stafylakis, L. Burget, *et al.*, "Diaper: End-to-end neural diarization with perceiver-based attractors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[5] M. Härkönen, S. J. Broughton, and L. Samarakoon, "Eend-m2f: Masked-attention mask transformers for speaker diarization," *arXiv preprint arXiv:2401.12600*, 2024.

[6] A. Plaquet, N. Tawara, M. Delcroix, S. Horiguchi, A. Ando, and S. Araki, "Mamba-based segmentation model for speaker diarization," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2025, pp. 1–5.

[7] I. Medennikov, M. Korenevsky, T. Prisyach, *et al.*, "Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario," *arXiv preprint arXiv:2005.07272*, 2020.

[8] M. Cheng, W. Wang, Y. Zhang, X. Qin, and M. Li, "Target-speaker voice activity detection via sequence-to-sequence prediction," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[9] D. Wang, X. Xiao, N. Kanda, M. Yousefi, T. Yoshioka, and J. Wu, "Profile-error-tolerant target-speaker voice activity detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 11 906–11 910.

[10] D. Wang, X. Xiao, N. Kanda, T. Yoshioka, and J. Wu, "Target speaker voice activity detection with transformers and its integration with end-to-end neural diarization," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[11] M.-K. He, J. Du, Q.-F. Liu, and C.-H. Lee, "Ansd-mamse: Adaptive neural speaker diarization using memory-aware multi-speaker embedding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1561–1573, 2023.

[12] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *arXiv preprint arXiv:2006.12847*, 2020.

[13] H. Bredin, R. Yin, J. M. Coria, *et al.*, "Pyannote. audio: Neural building blocks for speaker diarization," in *ICASSP 2020-2020 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2020, pp. 7124–7128.

[14] N. Vaessen and D. A. Van Leeuwen, "Fine-tuning wav2vec2 for speaker recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7967–7971.

[15] C. Wang, Y. Wu, Y. Qian, *et al.*, "Unispeech: Unified speech representation learning with labeled and unlabeled data," in *International Conference on Machine Learning*, PMLR, 2021, pp. 10 937–10 947.

[16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*, PMLR, 2023, pp. 28 492–28 518.

[17] S. Chen, C. Wang, Z. Chen, *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1505–1518, 2021.

[18] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *ArXiv*, vol. abs/2005.07143, 2020.

[19] H. Wang, S. Zheng, Y. Chen, L. Cheng, and Q. Chen, "Cam++: A fast and efficient network for speaker verification using context-aware masking," *arXiv preprint arXiv:2303.00332*, 2023.

[20] F. Yu, S. Zhang, Y. Fu, *et al.*, "M2met: The icassp 2022 multi-channel multi-party meeting transcription challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6167–6171.

[21] N. Ryant, P. Singh, V. Krishnamohan, *et al.*, *The third dihard diarization challenge*, 2021. arXiv: 2012.01477 [eess.AS].

[22] F. Landini, M. Diez, A. Lozano-Diez, and L. Burget, "Multi-speaker and wide-band simulated conversations as training data for end-to-end neural diarization," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[23] D. Snyder, G. Chen, and D. Povey, *MUSAN: A Music, Speech, and Noise Corpus*, arXiv:1510.08484v1, 2015. eprint: 1510.08484.

[24] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224. DOI: 10.1109/ICASSP.2017.7953152.