Adapting WavLM for Vietnamese Speaker Diarization in Real-world Conversations

Tuan-Duy THANG
Ho Chi Minh City University of Technology
VNU-HCMC
Ho Chi Minh City, Vietnam
duy.thangkhmtk20@hcmut.edu.vn

Van-Huy NGUYEN
Thai Nguyen University of Technology
Thai Nguyen, Vietnam
huynguyen@tnut.edu.vn

Tri-Nhan DO
VinBigData, VinGroup
Ha Noi, Vietnam
dotrinhan99@gmail.com

Quoc-Khanh NGUYEN
VinBigData, VinGroup
Ha Noi, Vietnam
20521452@gm.uit.edu.vn

Trung-Kien PHAN
VinBigData, VinGroup
Ha Noi, Vietnam
v.kienpt13@vinbigdata.com

Dang-Khoa MAC
VinBigData, VinGroup
Ha Noi, Vietnam
v.khoamd@vinbigdata.org

***Abstract*—While end-to-end neural diarization (EEND) models have achieved state-of-the-art performance, their effectiveness in low-resource languages such as Vietnamese remains underexplored due to the lack of annotated conversational data. In this study, we adapt WavLM-based speaker diarization to Vietnamese by fine-tuning the DiariZen model with Vietnamese speech data. Additionally, we introduce ViYT-Diar, a high-quality benchmark of manually annotated Vietnamese dialogues. Experimental results show that our fine-tuned model achieves a DER of 11.76% on the English CALLHOME two-speaker test set and 2.38% on ViYT-Diar, significantly outperforming Pyannote 3.1 and Falcon API. Furthermore, our custom clustering pipeline maintains stable performance across chunk sizes (2.38% – 2.59% DER), whereas off-the-shelf clustering degrades from 9.73% to 3.75%. These results underscore the effectiveness of language-specific fine-tuning and tailored clustering for low-resource speaker diarization.**

***Index Terms*—WavLM, DiariZen, EEND-VC, Vietnamese Adaptation**

I. Introduction

Speaker diarization (SD) aims to answer the question "who spoke when?" in multi-speaker audio. It is essential for ASR, meeting transcription, and conversational analysis. Traditional SD systems follow a clustering-based pipeline: voice activity detection (VAD) identifies speech regions, speaker embeddings (e.g., i-vectors, x-vectors) are extracted, and clustering methods such as AHC [9] or VBx assign speaker labels. While effective in structured settings, these methods struggle with overlapping speech, domain shifts, and spontaneous conversations.

To overcome these limitations, end-to-end neural diarization (EEND) [6] reformulates SD as a multi-label classification problem, allowing direct prediction of frame-wise speaker activity and handling overlaps without explicit clustering. Extensions such as EEND-VC [11] and TS-VAD [15] integrate speaker embeddings or attention to further improve performance, achieving state-of-the-art results on datasets like AMI [12], CHiME-6 [26], and DIHARD III [19]. However, these models require large annotated datasets, which are limited for low-resource languages like Vietnamese.

Self-supervised learning (SSL) models, including Wav2Vec 2.0, HuBERT, and WavLM [4], address data scarcity by learning representations from unlabeled speech. Among them, WavLM has shown strong results across ASR, speaker verification, and diarization tasks. Building on this, DiariZen [8] combines WavLM embeddings with a Conformer architecture to better capture speaker turn dynamics and temporal dependencies. Though effective on standard benchmarks, its performance for Vietnamese — a tonal and phonetically distinct language — remains unexplored.

In this paper, we adapt DiariZen to Vietnamese speaker diarization in two-speaker telephone conversations on both simulated and real Vietnamese telephone data. Additionally, we introduce a benchmark dataset for Vietnamese diarization with curated annotations.

Our key contributions are:

- Adapting and fine-tuning DiariZen for Vietnamese speaker diarization using WavLM embeddings.
- Introducing ViYT-Diar, a real-world Vietnamese test set with expert-annotated speaker labels.

The remainder of the paper is organized as follows: Section II reviews related work. Section III presents our DiariZen-based pipeline. Section IV describes the datasets, training, and evaluation setup. Section V concludes the paper.

II. Related works

Speaker diarization has evolved significantly, transitioning from traditional clustering-based methods to deep learning-based solutions. This section reviews key developments in the field, focusing on clustering-based approaches, end-to-end neural diarization, and self-supervised learning for diarization.

Traditional speaker diarization systems follow a pipeline consisting of voice activity detection (VAD), speaker embedding extraction, and clustering algorithms. Early approaches

used i-vector-based embeddings, but recent advancements introduced x-vector-based embeddings, which improved speaker discrimination by leveraging deep neural networks. Clustering methods such as AHC [10], k-means, and VBx have been widely used to group similar speaker segments. While effective, these methods struggle with overlapping speech and require post-processing techniques such as resegmentation and re-assignment to improve accuracy.

To address the limitations of clustering-based methods, end-to-end neural diarization (EEND) was introduced, treating diarization as a multi-label classification problem [6]. Unlike traditional clustering-based approaches, EEND directly predicts speaker activity probabilities at each time step, allowing for better handling of overlapping speech. Further advancements, such as EEND-vector clustering (EEND-VC) [11] and TS-VAD [15], have improved diarization accuracy by integrating self-attention mechanisms and embedding-based speaker tracking. However, these methods require large annotated datasets, making them less practical for low-resource languages like Vietnamese.

Self-supervised learning (SSL) has revolutionized speech processing by allowing models to learn generalizable speech representations from unlabeled data. SSL models like Wav2Vec, HuBERT, and WavLM [4] have demonstrated superior performance in ASR, speaker verification, and diarization. Among these, WavLM has been particularly effective for diarization tasks due to its ability to capture long-term speaker dependencies and disentangle speaker identity from background noise.

DiariZen [8] is a self-supervised diarization model that integrates WavLM embeddings with a Conformer-based temporal encoder. It achieves state-of-the-art performance on AMI [3] and AISHELL-4 [5], demonstrating how transformer-based sequence modeling can refine speaker embeddings in multi-speaker scenarios. However, the model has not been adapted to Vietnamese, and its effectiveness in tonal languages remains an open research question.

Research on Vietnamese speaker diarization remains constrained by the scarcity of annotated corpora. To date, most efforts have relied on simulated dialogues generated by concatenating utterances from speaker-verification datasets such as VIVOS [14] and ZALO-400. These artificial conversations fail to reproduce key conversational phenomena—overlapping speech, spontaneous discourse markers, and natural turn-taking dynamics.

Nam and Huynh [16] were among the first to evaluate x-vector embeddings with clustering methods (AHC, k-means, mean-shift) on two-speaker simulations from VIVOS, reporting 89.29 % accuracy but suffering from limited real-world applicability. Nguyen et al. [17] subsequently compared x-vectors and ECAPA-TDNN on simulated Vietnamese telephone speech, demonstrating the superiority of ECAPA-TDNN for both verification and diarization. Nonetheless, their study likewise lacked evaluation of genuine multi-speaker interactions.

In view of these limitations and the absence of a standardized benchmark for Vietnamese diarization, we propose to fine-tune the DiariZen model [8] on Vietnamese data and to introduce ViYT-Diar, a real-world dataset designed for rigorous, language-specific evaluation of speaker diarization systems.

III. Methodology

Our approach follows the EEND-VC [11] principle by leveraging DiariZen's end-to-end neural diarization model [8]. Rather than relying on Pyannote's speaker embedding extraction and agglomerative hierarchical clustering modules [1], we adopt our own implementations for these components. Additionally, we adapt the entire system to both simulated and real-world Vietnamese data to better address the language-specific and domain-specific characteristics.

A. Vietnamese data simulation

Our simulation method is inspired by previous studies, but with modifications to enhance data quality. Specifically, we apply Voice Activity Detection (VAD) [23] to segment speech, ensuring that only speech-containing segments are used. These segments are then randomly selected and progressively filled into simulated two-speaker conversations. Additionally, we add background noise from the MUSAN dataset [22] to improve robustness against real-world conditions (Algorithm 1).

Algorithm 1 simulate_data

Require: speaker_data, max_duration, sr, noise
Ensure: audio, labels, annotations, num_speakers, speakers

```
speakers ← SelectSpeakers(speaker_data)
utterances ← AssignUtterances(speakers)
audio ← InitBackground(max_duration)
labels ← InitLabels(max_duration)
segments ← ExtractSegments(utterances)  ▷ using VAD()
Shuffle(segments),    t ← 0
for segment in segments do
    start_time ← GetStartTime(t)
    if Overflow(start_time, segment) then break
    end if
    Mix(audio, segment, start_time)
    Update(labels, segment, start_time)
    Annotate(segment, start_time)
    t ← start_time + segment.length
    if NoSpace(t) then break
    end if
end for
audio ← AddNoise(audio, noise)
return audio, labels, annotations, len(speakers), speakers
```

B. End-to-end neural diarization module

The diarization module is designed to handle overlapping speech and accurately segment multi-speaker conversations. It consists of three main components: the feature extraction block, a Conformer-based encoder, and a classification head (Fig. 1).

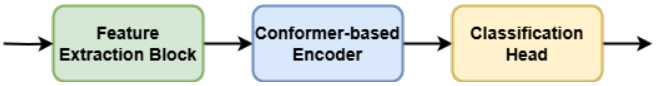

Fig. 1. End-to-end neural diarization module

1) Feature Extraction Block: To extract informative speech features, we adopt a self-supervised learning (SSL) approach based on WavLM [4], which is trained on large-scale speech corpora. This method leverages masked speech modeling, speech denoising, and gated relative position bias to enhance robustness in speaker diarization tasks.

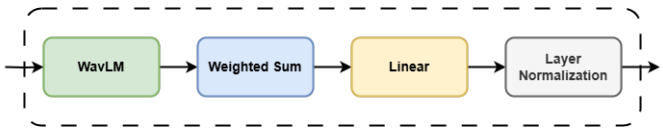

Fig. 2. Feature Extraction Block

Given an input speech signal $X = \{x_1, x_2, ..., x_T\}$, WavLM processes it through a convolutional feature encoder and transformer layers to produce a sequence of frame-level representations $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_T\}$. Instead of using only the final layer output, we employ a learnable weighted sum to aggregate information across multiple transformer layers:

$$\mathbf{h}_t = \sum_{l=1}^{L} \alpha_l \mathbf{W}_l(\mathbf{x}_t) \tag{1}$$

where $\mathbf{W}_l$ represents the transformation applied at layer l, and α_l is a trainable weight determining the contribution of each layer. The aggregated features are then projected into a lower-dimensional space:

$$\mathbf{z}_t = \mathbf{W}_{proj}\mathbf{h}_t + \mathbf{b}_{proj} \tag{2}$$

where $\mathbf{W}_{proj}$ and $\mathbf{b}_{proj}$ are the projection matrix and bias term, respectively. Layer normalization is applied before passing the features into the diarization model (Fig. 2).

2) Conformer-based Encoder: To model both short-term phonetic variations and long-term speaker dependencies, we employ a Conformer-based sequence encoder [7], consisting of 4 stacked Conformer blocks. The Conformer integrates self-attention mechanisms for global context modeling with convolutional layers for local feature extraction. Compared to the original Conformer, the version used in DiariZen removes the positional encoding in the multi-head self-attention (MHSA) module, as WavLM already encodes positional information. This modification prevents redundant encoding and allows the model to focus on speaker-related acoustic patterns. Each

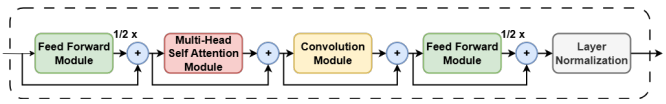

Fig. 3. Conformer Block

Conformer block refines frame-level features through four steps with residual connections. First, a feed-forward network (FFN_1) is applied to the input and scaled by one-half before being added back to the original input. This is followed by a multi-head self-attention (MHSA) layer, whose output is added to the previous result. Next, a depthwise convolution layer is applied and its output is added as well. Finally, another feed-forward network (FFN_2) is applied, also scaled by one-half, and the result is added before applying layer normalization (Fig. 3). This architecture allows the encoder to capture both global speaker dependencies and local speech features effectively.

3) Classification Head: The diarization decision is obtained through a classification head consisting of a linear transformation followed by a softmax activation function. Instead of predicting single-speaker labels, we employ a powerset-based classification approach [6], where each output state represents a unique combination of active speakers. The number of possible speaker states in a system supporting up to N speakers, including silence, is given by $C = 2^N$.

At each diarization frame, the model predicts a probability distribution over these C states and is trained using negative log-likelihood loss (NLL_Loss) [21], which encourages correct classification of the ground-truth state.

C. Speaker Embedding Extraction

Given a long audio recording, we first segment it into K non-overlapping chunks of equal duration L:

$$X = [X_1, X_2, \ldots, X_K], \quad X_k \in \mathbb{R}^{L \times D}, \tag{3}$$

where each X_k is processed independently by the EEND module to obtain frame-wise output probabilities $P(y_{k,t} \mid \mathbf{o}_{k,t})$ for each powerset class $y_{k,t}$.

From these outputs, we identify, for each chunk k and speaker s, the set of frames where only speaker s is active:

$$T_{k,s} = \left\{ t \;\middle|\; \arg\max_c P(y_{k,t} = c \mid \mathbf{o}_{k,t}) = s \right\} \tag{4}$$

where c is a singleton state

We then extract the corresponding audio segments $\{x_{k,t}\}_{t \in T_{k,s}}$ where only speaker s is active. These segments are concatenated to form a continuous speech segment $x_{k,s}$ for speaker s in chunk k:

$$x_{k,s} = \text{Concatenate}\left(\{x_{k,t}\}_{t \in T_{k,s}}\right) \tag{5}$$

To obtain speaker embeddings, we input $x_{k,s}$ into a pre-trained speaker verification model, such as WavLM [4] or ResNet34-LM [25], which outputs a fixed-dimensional embedding vector $\mathbf{e}_{k,s}$:

$$\mathbf{e}_{k,s} = \text{SpeakerEncoder}(x_{k,s}) \tag{6}$$

These embeddings $\{\mathbf{e}_{k,s}\}_{k=1}^{K}$ serve as speaker representations across chunks and are used in the clustering process to merge segments belonging to the same speaker. This approach ensures speaker identity consistency across different diarization segments.

D. Agglomerative Hierarchical Clustering

Using the speaker embeddings extracted in the Section III-C, we apply agglomerative hierarchical clustering (AHC) [10] with two target clusters, cosine affinity, and average linkage. Let

$$\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_M\} \tag{7}$$

be the set of all embeddings produced in the previous subsection. AHC begins by treating each $\mathbf{e}_i$ as its own cluster and then iteratively merges the two clusters with the smallest average pairwise cosine distance. For clusters A and B, this distance is computed as

$$d(A, B) = \frac{1}{|A|\,|B|} \sum_{\substack{\mathbf{e}_i \in A \\ \mathbf{e}_j \in B}} \big(1 - \mathrm{sim}(\mathbf{e}_i, \mathbf{e}_j)\big), \tag{8}$$

where

$$\mathrm{sim}(\mathbf{e}_i, \mathbf{e}_j) = \frac{\mathbf{e}_i^\top \mathbf{e}_j}{\|\mathbf{e}_i\|\,\|\mathbf{e}_j\|}. \tag{9}$$

Merging continues until exactly two clusters remain. Finally, each embedding—and its corresponding speech segment—is assigned the label of its cluster, ensuring consistent speaker identities across the entire recording.

IV. Experiments

A. Dataset Preparation

1) Training Datasets: To train our diarization model, we adopt the pretrained model weights from DiariZen [8], which was initially trained on a multilingual and multi-domain mix of three corpora: AMI [12], AISHELL-4 [5], and AliMeeting [27]. We then fine-tune the model on Vietnamese telephone speech using two datasets in parallel:

- A simulated corpus of 800 hours of enrollment utterances from 16000 unique speakers, constructed from a private in-house dataset as described in Section III-A;
- A real-world telephone dataset comprising approximately 6000 hours of two-speaker conversations.

All recordings were internally collected with proper consent and manually reviewed to ensure the exclusion of any personally identifiable information or sensitive content.

This dual-dataset fine-tuning enables the model to learn from both controlled conditions and natural conversational scenarios.

To analyze the impact of audio segment length on training and inference performance across different test datasets, we compare models trained with audio durations of 8s, 16s, and 24s. The training data is simulated on the fly, allowing the model to learn from dynamically varying segment structures.

2) Testing Datasets: For evaluation, we build ViYT-Diar, a Vietnamese test dataset sourced from various YouTube channels. The test set consists of 100 mono audio recordings, each with a duration of 60 seconds, sampled at 16kHz. Speech segments are manually annotated using LabelStudio [24] (Fig. 4), following strict criteria:

- Speaker voices must be clear,
- Silence between two segments shorter than 250ms is merged into a single segment,
- The dataset must include diverse conversational contexts to reflect real-world scenarios.

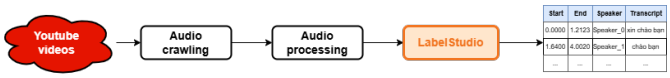

Fig. 4. Pipeline of building ViYT-Diar dataset

Additionally, we evaluate our model on the CALLHOME 2-speaker dataset [18] in English to compare its performance against other diarization models.

B. Experimental Setup

1) EEND model: For feature extraction, we utilize the pretrained WavLM Base+ model [4]. The outputs from its convolutional layers and 12 transformer encoders are combined using a weighted sum. A linear layer projects the features from 768 to 256 dimensions before passing them to the Conformer encoder. The Conformer consists of 4 blocks, each incorporating feed-forward, self-attention, and convolutional layers. The feed-forward network has input and hidden dimensions of 256 and 1024, respectively. The multi-headed self-attention mechanism employs 4 attention heads, while the convolution module uses a kernel size of 31. All dropout rates are set to 0.1. For classification, we adopt a powerset-based loss function, assuming a maximum of 4 speakers with up to 2 overlapping speakers per frame, plus an additional class for silence. This results in a total of 11 output classes. The final classification layer has an output dimension of 11. The total number of parameters for WavLM and Conformer is 94.7 million and 6.1 million, respectively.

2) Configuration: We optimize the model with AdamW [28], employing separate optimizers for WavLM (learning rate 1×10^{-5}) and all other components (learning rate 1×10^{-3}). Training proceeds for up to 100 epochs, with early stopping after 10 epochs without validation-loss improvement. Gradient clipping thresholds are set dynamically by AutoClip [20], using the 90th percentile of observed gradient norms. During inference, speaker embeddings are extracted from ResNet34-LM2 that trained via the WeSpeaker toolkit on VoxCeleb2 [25]. Finally, we apply `scikit-learn`'s agglomerative hierarchical clustering [10] with cosine distance and average linkage, fixing the number of clusters at two to match our two-speaker scenario.

C. Evaluation Metrics

To assess the performance of our speaker diarization system, we employ the Diarization Error Rate (DER) as the primary evaluation metric. DER quantifies the proportion of incorrectly classified speech time and is widely used for benchmarking diarization systems.

DER is computed as:

$$\mathrm{DER} = \frac{S + FA + MS}{T} \tag{10}$$

TABLE I
COMPARISON OF DER(%) ACROSS DIFFERENT CHUNK_SIZES

EEND	Clustering pipeline	Chunk_size	CALLHOME		ViYT-Diar	
			ignore_overlap	overlap	ignore_overlap	overlap
DiariZen	Pyannote	8	8.07	9.24	9.72	9.73
		16	**7.07**	**8.81**	4.23	4.24
		24	7.14	9.09	3.74	3.75
	Ours	8	9.71	11.76	**2.37**	**2.38**
		16	9.39	11.35	2.74	2.75
		24	9.17	11.30	2.58	2.59

TABLE II
COMPARISON OF DER(%) ACROSS DIFFERENT DIARIZATION MODEL

Model	Type	CALLHOME 2-speaker		ViYT-Diar	
		ignore_ovl	ovl	ignore_ovl	ovl
Pyannote 3.1	EN-pretrained	15.13	15.73	4.86	4.87
DiaPer	EN-pretrained	34.87	34.68	19.45	19.46
Falcon	API	34.88	36.55	5.17	5.18
DiariZen (original)	EN-pretrained	11.24	11.80	23.93	23.94
DiariZen (ours)	VI-pretrained	**9.71**	**11.76**	**2.37**	**2.38**

where:

- S represents speaker confusion, which occurs when a speaker is misclassified.
- FA (False Alarm) accounts for non-speech regions incorrectly classified as speech.
- MS (Missed Speech) denotes speech regions that are incorrectly classified as silence.
- T is the total duration of the reference speech.

A lower DER indicates better diarization performance. For evaluation, we follow the standard protocol where a collar tolerance of 0.25s is applied around speaker segment boundaries, and overlapping speech is considered in the scoring.

D. Results and Analysis

Table I reports DER results on two datasets: CALLHOME (English) and ViYT-Diar (Vietnamese). On CALLHOME, our clustering consistently underperforms Pyannote's pipeline across all chunk sizes, indicating that Pyannote is better suited to standard English conversations and benefits from longer chunks. In contrast, on ViYT-Diar our clustering excels and remains remarkably stable regardless of chunk size, whereas Pyannote only improves from 9.73% to 3.75% when increasing from 8s to 24s. These results demonstrate that while Pyannote clustering holds an advantage on English data, our pipeline is far more robust to chunk-size variation and better tailored to the characteristics of Vietnamese conversational speech. We hypothesize that the sensitivity to chunk size arises because different durations yield varying numbers of speaker embeddings for clustering (Section III-D). Models like Pyannote are more prone to instability with sparse or unbalanced embedding sequences, especially for shorter chunks, leading to inconsistent speaker grouping—particularly in tonal or overlapping Vietnamese speech.

Table II compares our diarization model with leading approaches including Pyannote 3.1 [2], the current state-of-the-art in speaker diarization; DiaPer [13], a previous EEND approach with perceiver-based attractors by the same authors; the Falcon API, a commercial private service; and the English pretrained DiariZen model [8]. Performance is reported on both the CALLHOME two-speaker dataset [18] and ViYT-Diar.

Our fine-tuned DiariZen model trained on Vietnamese data achieves the lowest DER on both datasets, with DERs of 11.76% on CALLHOME 2-speaker and 2.38% on ViYT-Diar. Compared to the pretrained English DiariZen model, which has a DER of 23.94% on ViYT-Diar, our fine-tuned version demonstrates a substantial improvement, confirming the effectiveness of adapting diarization models to the target language.

Additionally, Pyannote 3.1 and DiaPer, both pretrained on English datasets, show significantly higher DERs, particularly on Vietnamese audio, indicating that models trained on English data do not generalize well to Vietnamese conversations. The Falcon API performs poorly on CALLHOME 2-speaker but achieves a relatively lower DER of 5.18% on ViYT-Diar, suggesting that its performance varies across datasets.

Overall, our results highlight the importance of training diarization models with language-specific data. The fine-tuned DiariZen model consistently outperforms other approaches, demonstrating its robustness and effectiveness for Vietnamese

speaker diarization.

V. Conclusion

This paper presented an adaptation of DiariZen, based on WavLM embeddings, for Vietnamese speaker diarization. To address low-resource challenges, we fine-tuned the model on simulated and real Vietnamese telephone conversations, achieving strong diarization performance. We also introduced ViYT-Diar, a new benchmark for real-world Vietnamese speaker diarization, covering diverse conversational topics with manual speaker annotations.

On the CALLHOME two-speaker test, our fine-tuned DiariZen substantially reduces error compared to the English pre-trained baseline and other systems. On ViYT-Diar it achieves under 3% DER—about half the error of Pyannote, DiaPer, and Falcon API. Moreover, its clustering remains virtually unchanged across different chunk lengths, whereas competitor methods only improve noticeably with much longer segments. These results show that our pipeline is both more accurate for Vietnamese speech and more robust to varying conditions.

Our findings highlight the importance of language-specific fine-tuning and clustering design. Future work will explore scaling to multi-speaker conversations, improving inference speed, and extending to broader Vietnamese speech domains.

References

[1] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 7124–7128. IEEE, 2020.

[2] Hervé Bredin. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*, 2023.

[3] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer, 2005.

[4] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[5] Yihui Fu, Luyao Cheng, Shubo Lv, Yukai Jv, Yuxiang Kong, Zhuo Chen, Yanxin Hu, Lei Xie, Jian Wu, Hui Bu, et al. Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario. *arXiv preprint arXiv:2104.03603*, 2021.

[6] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Kenji Nagamatsu, and Shinji Watanabe. End-to-end neural speaker diarization with permutation-free objectives. *arXiv preprint arXiv:1909.05952*, 2019.

[7] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

[8] Jiangyu Han, Federico Landini, Johan Rohdin, Anna Silnova, Mireia Diez, and Lukáš Burget. Leveraging self-supervised learning for speaker diarization. In *Proc. ICASSP*, 2025.

[9] Kyu J Han, Samuel Kim, and Shrikanth S Narayanan. Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(8):1590–1601, 2008.

[10] Kyu Jeong Han and Shrikanth S Narayanan. Agglomerative hierarchical speaker clustering using incremental gaussian mixture cluster modeling. In *Interspeech*, pages 20–23, 2008.

[11] Keisuke Kinoshita, Marc Delcroix, and Naohiro Tawara. Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7198–7202. IEEE, 2021.

[12] Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. The ami meeting corpus. In *Proc. International Conference on Methods and Techniques in Behavioral Research*, pages 1–4, 2005.

[13] Federico Landini, Mireia Diez, Themos Stafylakis, and Lukáš Burget. Diaper: End-to-end neural diarization with perceiver-based attractors. *arXiv preprint arXiv:2312.04324*, 2023.

[14] Hieu-Thi Luong and Hai-Quan Vu. A non-expert kaldi recipe for vietnamese speech recognition system. In *Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016)*, pages 51–55, 2016.

[15] Ivan Medennikov, Maxim Korenevsky, Tatiana Prisyach, Yuri Khokhlov, Mariya Korenevskaya, Ivan Sorokin, Tatiana Timofeeva, Anton Mitrofanov, Andrei Andrusenko, Ivan Podluzhny, 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.

[16] Nguyen Duc Nam and Hieu Trung Huynh. Speaker diarization in vietnamese voice. In *Future Data and Security Engineering. Big Data, Security and Privacy, Smart City and Industry 4.0 Applications: 8th International Conference, FDSE 2021, Virtual Event, November 24–26, 2021, Proceedings 8*, pages 444–451. Springer, 2021.

[17] Tung Lam Nguyen, Bao Thang Ta, Thi Anh Xuan Tran, Nhat Minh Le, et al. Speaker diarization for vietnamese conversations using deep neural network embeddings. In *2022 IEEE Ninth International Conference on Communications and Electronics (ICCE)*, pages 300–305. IEEE, 2022.

[18] Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. Improved speech-to-text translation with the fisher and callhome spanish-english speech translation corpus. In *Proceedings of the 10th international workshop on spoken language translation: papers*, 2013.

[19] Neville Ryant, Prachi Singh, Venkat Krishnamohan, Rajat Varma, Kenneth Church, Christopher Cieri, Jun Du, Sriram Ganapathy, and Mark Liberman. The third dihard diarization challenge. *arXiv preprint arXiv:2012.01477*, 2020.

[20] Prem Seetharaman, Gordon Wichern, Bryan Pardo, and Jonathan Le Roux. Autoclip: Adaptive gradient clipping for source separation networks. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2020.

[21] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[22] David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*, 2015.

[23] Jongseo Sohn, Namsoo Kim, and Wonyong Sung. A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1):1–3, 1999.

[24] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. Label Studio: Data labeling software, 2020-2025. Open source software available from https://github.com/HumanSignal/label-studio.

[25] Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian. Wespeaker: A research and production oriented speaker embedding learning toolkit. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[26] Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, et al. Chime-6 challenge: Tackling multi-speaker speech recognition for unsegmented recordings. *arXiv preprint arXiv:2004.09249*, 2020.

[27] 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 *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6167–6171. IEEE, 2022.

[28] Pan Zhou, Xingyu Xie, Zhouchen Lin, and Shuicheng Yan. Towards understanding convergence and generalization of adamw. *IEEE transactions on pattern analysis and machine intelligence*, 2024.