MAPR 2025 THE 8TH INTERNATIONAL CONFERENCE ON MULTIMEDIA ANALYSIS AND PATTERN RECOGNITION

Adapting WavLM for Vietnamese Speaker Diarization in Real-world Conversations

Tuan-Duy Thang[1,2], Van-Huy Nguyen[3,4], Tri-Nhan Do[3], Quoc-Khanh Nguyen[3], Trung-Kien Phan[3], Dang-Khoa Mac[3]
[1]Ho Chi Minh City University of Technology, [2]VNU-HCMC, [3]VinBigData, [4]Thai Nguyen University of Technology

INTRODUCTION

Speaker diarization answers the question: **"who spoke when"** in multi-speaker conversations.

Traditional pipelines face challenges: Overlapping speech, Domain mismatch.

Vietnamese is **a low-resource language** for diarization:

- Lacks large annotated datasets
- Pretrained models are ineffective without adaptation.

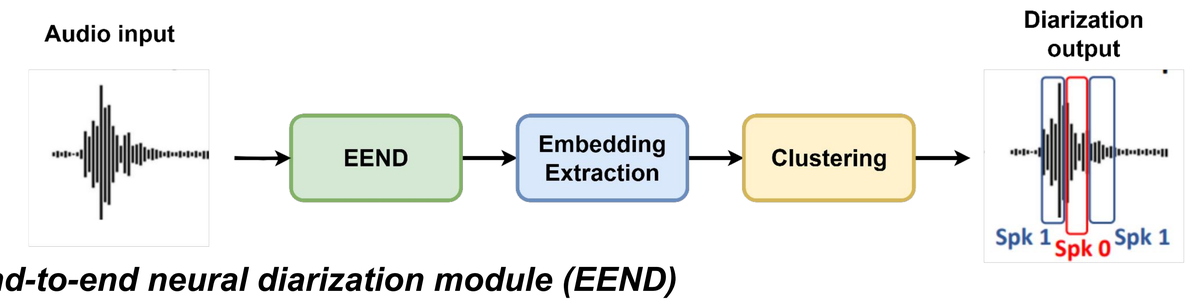
Ours contributions:

- Fine-tune SOTA diarization model on Vietnamese speech: DiariZen.
- Develop ViYT-Diar dataset: 100 real-world Vietnamese conversations with manual annotations.
- Design robust speaker embedding and clustering pipeline for Vietnamese tonal speech

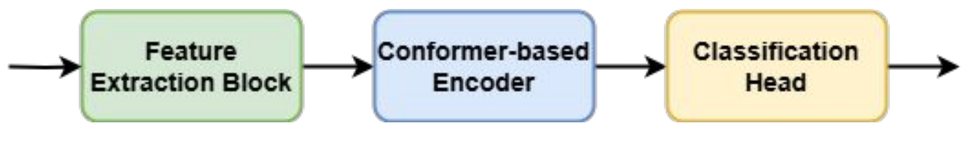
METHOD

A. Diarization Pipeline

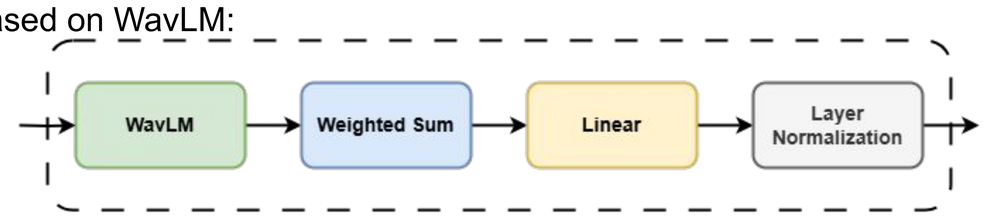
Our pipeline follows the EEND-VC principle by leveraging DiariZen's end-to-end neural diarization model:



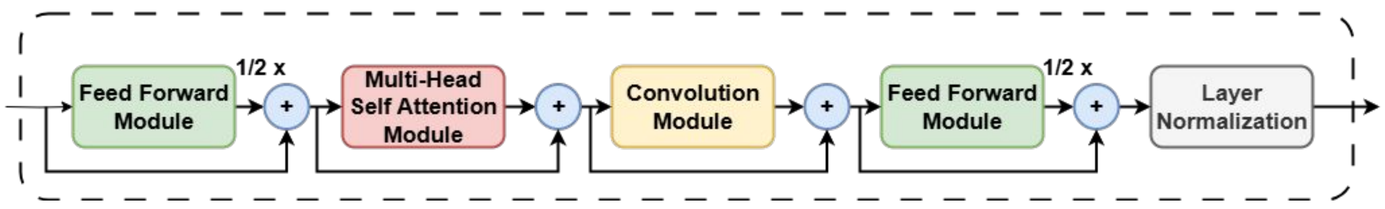
1. End-to-end neural diarization module (EEND)



To extract informative speech features, we adopt a self-supervised learning (SSL) approach based on WavLM:



We use a *4-layer Conformer encoder* that combines self-attention for global context and convolution for local feature extraction:



Classification head: A linear layer followed by softmax forms a powerset-based classifier (2^n states) trained with NLL loss to predict active speaker combinations per frame.

2. Speaker Embedding Extraction

The input audio is segmented into chunks X_k. From each chunk, we extract single-speaker segments $T_{k,s}$ and encode them into embeddings $e_{k,s}$ using a speaker encoder (e.g., ResNet34)

$$X = [X_1, X_2, \ldots, X_K], \quad X_k \in \mathbb{R}^{L \times D},$$
$$T_{k,s} = \left\{ t \;\middle|\; \arg\max_c P(y_{k,t} = c \mid \mathbf{o}_{k,t}) = s \right\}$$
$$x_{k,s} = \text{Concatenate}\left(\{x_{k,t}\}_{t \in T_{k,s}}\right)$$
$$\mathbf{e}_{k,s} = \text{SpeakerEncoder}(x_{k,s})$$

3. Agglomerative Hierarchical Clustering

We perform agglomerative hierarchical clustering on the embeddings using average linkage and cosine affinity, merging until two clusters remain to assign consistent speaker labels across the recording.

B. Data Preparation

1. Vietnamese Data Simulation for Training

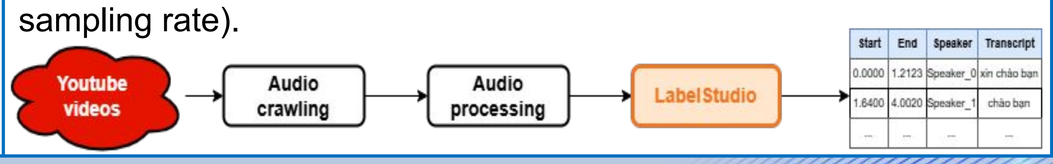
Algorithm 1 simulate_data

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
```

2. Vietnamese Data Annotation for Testing: ViYT-Diar

Includes 100 single-channel audio recordings (~60s each, 16kHz sampling rate).



RESULT

- Achieves **<3% DER** on Vietnamese conversations
- Outperforms existing baselines by large margins.
- Stable performance across chunk sizes → practical for real-world deployments

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**

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