Received: July 10, 2024. Revised: August 7, 2024.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

End-to-End Neural Diarization for Unknown Number of Speakers with Multi-Scale Decoder

Myat Aye Aye Aung[1]* **Win Pa Pa[1]** **Hay Mar Soe Naing[1]**

[1]*Natural Language Processing Lab, University of Computer Studies, Yangon, Myanmar*
* Corresponding author's Email: myatayeayeaung@ucsy.edu.mm

Abstract: Speaker diarization is crucial for enhancing speech communication across various domains, including broadcast news, meetings, conferences featuring multiple speakers. Nevertheless, real-time diarization applications face persistent challenges due to overlapping speech and varying acoustic conditions. To address these challenges, End-to-End Neural Diarization (EEND) has demonstrated superior performance compared to traditional clustering-based methods. Conventional neural techniques often rely on fixed datasets, which can hinder their ability to generalize across different speech patterns and real-world environments. Therefore, this research proposes an EEND model utilizing a Multi-Scale approach to compute optimal weights, essential for generating speaker labels across multiple scales. The Multi-Scale Diarization Decoder (MSDD) approach accommodates a flexible number of speakers, overlap-aware diarization, and integrates a pre-trained speaker embedding model. The investigation included different languages and datasets, such as the proposed Myanmar M-Diarization dataset and the English AMI meeting corpus. Notably, many benchmark multi-speaker datasets for speaker diarization include no more than 8 speakers per audio and have fixed-length speakers per audio. Hence, this study developed its own dataset featuring up to 15 speakers with flexible number of speakers. Furthermore, the study demonstrates language-independence, underscoring its efficacy across diverse linguistic contexts. Comparative analysis revealed that the proposed model outperformed clustering baseline methods (i-vectors and x-vectors) and single-scale EEND approaches in both languages regarding Diarization Error Rate (DER). Additionally, proposed M-Diarization dataset included audio of varying lengths and scenarios with an overlap ratio of 10%. The model was validated on the M-Diarization dataset, demonstrating its capability to handle flexible speaker counts and audio durations efficiently. This experiment marks the first implementation of an EEND with a Multi-Scale approach on a fixed-speaker English language corpus and the variable-speaker M-Diarization dataset. It achieved notable results: 44.63% for i-vectors, 47.38% for x-vectors, 19% for the EEND single-scale approach, and 4.37% for the EEND MSDD approach on overlap ratio 3.31% on the M-Diarization dataset. The experimental outcomes clearly indicate that the proposed method significantly enhances diarization performance, particularly in scenarios involving varying numbers of speakers and diverse audio conversation lengths.

Keywords: End-to-End Neural diarization, Overlap speech conversations, Speaker diarization, Multi-scale approach, diarization error rate.

1. Introduction

The primary function of speaker diarization is to accurately assign speech segments to specific speakers, enhancing communication fidelity in environments with dynamic interactions and varying speaker conditions [1, 2].

A standard diarization system follows a clustering-based methods (i-vectors and x-vectors) [3, 4]. Nevertheless, clustering-based methods that cannot be directly optimized to minimize diarization errors and have difficulty handling speaker overlaps [5].

To address these challenges, many researchers proposed EEND. It can effectively handle speaker overlaps during both training and inference by employing a multi-label classification framework [5]. However, they are prone to overfitting if the training data lacks diversity, leading to poor generalization on new, unseen data. Applying single-scale and multi-scale segmentation approaches to neural diarization can significantly impact the performance and robustness, particularly in scenarios with long audio recordings and multiple speakers. Single-scale segmentation is efficient but may struggle with varying speaker activity and rapid changes. In contrast, multi-scale segmentation, which analyses audio at multiple temporal resolutions, can better capture speaker turn changes and transitions, reducing errors caused by missed speaker turns or overlapping speech.

Therefore, the proposed model employs two types of multi-scale segmentation 5-scale and 6-scale on two different corpora: the AMI meeting corpus [13] and the M-Diarization dataset [10, 11]. The M-Diarization dataset represents a significant advancement in the field of Myanmar Language Speech Conversation, providing a valuable resource for research and development in the area, updated with natural, real-time speech. The dataset includes audio recordings with a variable number of speakers. A clustering-based approach using i-vectors and x-vectors serves as the baseline. Traditional end-to-end neural diarization with single segmentation is compared against the proposed multi-scale method. The emphasis is on handling overlapping speech and employing convolutional neural networks for weighting schemes. The proposed multi-scale end-to-end neural diarization method excels by managing overlapping speech, leveraging convolutional neural networks for precise weighting, and demonstrating superior performance across variable speaker numbers and diverse languages.

The contributions of this experiment, as highlighted below, are:

1. Enhancing the M-Diarization dataset with real-time multi-speaker conversations in the Myanmar language.
2. Focusing on speaker embedding features, the study examines i-vectors and x-vectors clustering-based methods as baselines, and end-to-end methods, compared to traditional single-scale and multi-scale approaches.
3. Proposing an end-to-end neural diarization method utilizing multi-scale segmentation with two different scales to manage a variable number of speakers.
4. Performing evaluations on two distinct language corpora: the AMI meeting corpus and the M-Diarization dataset.

The rest sections of the paper are structured as follows: Section 2 is provided a comprehensive review of existing research on i-vectors, x-vectors, and EEND. Section 3 is introduced the EEND approach is developed in this study, while Section 4 is presented the proposed methodology in detail. The M-Diarization dataset is described in Section 5. Section 6 is detailed the implementation of the proposed model, and Sections 7 and 8 is evaluated the proposed method compared to i-vectors, x-vectors, and single-scale EEND on two distinct datasets, respectively. Finally, Section 9 is offered a summary of the research findings. This paper is part of the ASEAN IVO 2023 project, 'Spoof Detection for Automatic Speaker Verification,' which aims to enhance the reliability of features and datasets for speaker verification in the Myanmar language.

2. Related works

In speaker diarization, extracting speaker embeddings is crucial for accurately clustering speech segments. This section reviews related work on speaker embedding extraction, highlighting key techniques that have advanced diarization performance.

Early approaches for speaker diarization utilized i-vectors to represent segmented speech, applying cosine similarity for scoring and clustering with K-means or spectral clustering. Experimental results with i-vector features often involved scenarios such as telephone conversations with two speakers in clean environments [3]. However, i-vectors are sensitive to channel and noise variability, which limits their robustness in diverse acoustic conditions.

To address these limitations, [4] introduced the x-vector system, which employs a Deep Neural Network (DNN) architecture to replace traditional i-vectors. The x-vector approach learns fixed-dimensional embeddings for acoustic segments of varying lengths and incorporates a scoring metric. Despite these advancements, both i-vectors and x-vectors still rely on clustering-based methodologies.

Building on this, [5] proposed the EEND model, which directly minimizes diarization errors and can handle overlapping speech. However, the single-scale approach used in EEND, which was trained with an overlap ratio of 5.8%, struggled with performance variations under different overlap conditions, particularly due to the significant discrepancy between training and test set overlap ratios. The highlighted the challenge of adapting end-to-end models to varying overlap ratios in practice.

In our work, we proposed a multi-scale approach for segmentation, speaker embedding extraction, and clustering in End-to-End Neural Diarization. This approach balances temporal resolution between short and long segments to enhance speaker diarization accuracy, particularly in real-time scenarios with unseen data. Short segments (0.5–1.0 seconds) enable rapid updates to speaker profiles during dynamic conversations, while longer segments (2–3 seconds) provide a comprehensive view of speaker characteristics over extended periods.

Our experiments demonstrate the effectiveness of multi-scale approach. We compared two techniques: the 5-scale method, which adapts well to short segments and facilitates rapid updates [16], and the 6-scale method, which excels in analysing longer segments. The 6-scale approach outperformed other methods in handling lengthy audio meetings from two different datasets. Specifically, the 6-scale method reduced diarization errors in long meetings compared to uniform segmentation approaches, and improved clustering accuracy on short segments. The empirical evidence supports the significant contribution of our multi-scale method, demonstrating its robustness and adaptability in diverse speech contexts.

3. End-To-End neural diarization

It utilizes a neural network that takes audio features as input and produces a unified representation of speech activities involving multiple speakers [5]. The model is trained to include non-speech segments and speaker overlaps, to minimize diarization errors. EEND method for speaker diarization task can be formulated as a multi-label classification problem, as follows [6], Let

$$X = (x_t \in \mathbb{R}^F | t = 1, \dots\dots, T) \tag{1}$$

$$Y = (y_t | t = 1, \dots\dots, T) \tag{2}$$

$$y_t = [y_{t,c} \in \{0,1\} | c = 1, \dots..C] \tag{3}$$

$$S = (s_t | t = 1, \dots., T) \tag{4}$$

$$S_t = [S_{t,n} \in \{0,1\} | n = 1, \dots.N] \tag{5}$$

In the EEND framework, we have the following formulas to compute the estimated output, *H:*

$$\begin{aligned} H &= (h_t \in \mathbb{R}^D | t = 1, \dots, T) \\ &= Encoder\ (X) \in \mathbb{R}^{D \times T} \end{aligned} \tag{6}$$

$$\begin{aligned} \hat{Y} &= (\hat{y}_t \in \mathbb{R}^C | t = 1, \dots., T) \\ &= \sigma\ (Linear\ (H) \epsilon\ \mathbb{R}^{C \times T}) \end{aligned} \tag{7}$$

Then the PIT loss in written as follows:

$$L^R = \frac{1}{TC} \min_{\emptyset \in perm(c)} \sum_t BCE(1_t^{\emptyset}, \hat{y}_t) \tag{8}$$

Table 1. Nomenclature section of the paper

Notation	Definition
X	The sequence of input features
x_t	The feature vector at time step *t*, with x_t being a *F*-dimensional vector
T	The total number of time steps in the sequence
$\mathbb{R}^F$	The *F*-dimensional feature space
Y	The sequence of ground-truth speaker labels for the input features.
y_t	The ground-truth speaker label vector at time step *t*
$y_{t,c}$	Binary value indicating the presence (1) or absence (0) of speaker *c* at time step *t*
C	The total number of speakers in the current recording.
S	The sequence of absolute ground-truth speaker labels for the input features.
s_t	The absolute ground-truth speaker label vector at time step *t*
$S_{t,n}$	Binary value indicating the presence (1) or absence (0) of speaker *n* at time step *t*
H	The sequence of feature vectors obtained from the Encoder layer
h_t	*D*-dimensional vector at time step *t*
$\mathbb{R}^D$	The *D*-dimensional feature space
$\hat{Y}$	The sequence of predicted speaker labels
$\hat{y}_t$	*C*-dimensional vector representing the probabilities for each speaker at time step t
σ	The sigmoid function applied element-wise to the output of the linear layer
$perm(c)$	The set of all possible permutations of 1,…,C
BEC(-,-)	Binary Cross Entropy function, measuring the difference between the predicted probabilities and the ground-truth labels.
$1_t^{\emptyset}$	The ground-truth label vector for the *ϕ-th* permutation at time step *t*
$v\ (t)$	The binary mask from the oracle VAD indicating speech activity.
$x_s\ (t)$	The masked audio signal at scale *s*, where only speech segments are retained.
F_s	The features extracted from the masked audio signal at scale *s*
$\alpha_s\ (t)$	The weight associated with scale *s* at time step *t*.
$E_s\ [t]$	The multi-scale embedding vector at time step *t* for scale *s*
C_k	The cluster-average embedding vector for the *k-th* cluster

The nomenclature section of the paper is shown in Table 1.

4. Proposed methodology

This section covers the theoretical concepts associated with the proposed EEND multi-scale approach. MSDD combines information from these varying scales, enabling the model to capture detailed speaker characteristics even in brief utterances.

By employing multi-scale inputs, MSDD significantly improves diarization performance, particularly in dynamic conversational contexts characterized by frequent speaker transitions. Our methodology integrates three neural network models: Multilingual Marblenet for Voice Activity Detection (VAD) [7], Titanet Large for generating Speaker Embeddings [8], and Diarization Multi-scale Clustering [8-9, 15-16]. Our approach incorporates two diarization methods on MSDD: Cluster Diarizer and Neural Diarizer. Cluster Diarizer is typically assessed with a default collar of 0.25 seconds, excluding overlap. Neural Diarizer, on the other hand, uses a MSDD combined with Oracle VAD to improve speaker diarization accuracy. The default configuration utilizes 5 embeddings with time scales ranging from 0.5 to 1.5 seconds. In less complex scenarios, a single embedding with longer durations, such as 1.0 or 1.5 seconds, may suffice. However, for unknown number of speakers, we employ 6 embeddings with time scales extending from 0.5 to 3.0 seconds. The overview of proposed method is shown in Fig. 1.

4.1 Oracle voice activity detection (VAD)

The proposed system employs a multi-scale approach in Oracle VAD evaluation to enhance speaker diarization accuracy. By utilizing ground-truth VAD timestamps at various temporal resolutions, it effectively captures speech activity and speaker changes across multiple time scales [7].

Given an audio signal *x(t),* the oracle VAD provides a binary mask *v(t)* indicating speech activity (1 for speech, 0 for non-speech). The audio signal is segmented into multiple scales *S*, where each scale *s* $\in S$ represents a different temporal resolution. Only segments marked as speech by the oracle VAD are processed:

$$x_s\,(t) = x\,(t).\,v\,(t) \tag{9}$$

For each scale *s*, features F_s is extracted from the masked audio signal:

$$F_s = FeatureExtractor\,(x_s\,(t)) \tag{10}$$

4.2 Speaker embeddings (TitaNet)

TitaNet, a cutting-edge speaker embedding extractor, generates speaker embeddings from extracted features [8]. For each scale and segment, TitaNet processes the features to produce a speaker embedding vector $E_s[i]$ that encapsulates the speaker's characteristics. The process is repeated across all scales and segments, resulting in a collection of speaker embeddings that capture speaker information at various temporal resolutions. These multi-scale embeddings provide a rich and comprehensive representation of the speaker's identity, essential for accurate diarization.

The multi-scale embeddings are then combined to form a single, integrated embedding for each segment. It can be achieved using methods such as concatenation, averaging, or attention mechanisms. The combined embedding is fed into a decoder, which assigns speaker labels to each segment by computing the probability of segment *i* belonging to a particular speaker through a *softmax* function.

4.3 Multi-scale diarization decoder (MSDD)

Following the extraction of multi-scale embeddings, these embeddings are processed by a clustering algorithm to provide initial clustering results for the MSDD module. The MSDD module uses cluster-average speaker embedding vectors to compare with input speaker embedding sequences [9]. For each step, the importance of each scale is weighed through estimated scale weights.

Ultimately, a neural network model called the MSDD is designed to leverage the multi-scale approach by dynamically calculating the weight of each scale. The MSDD utilizes the initial clustering results to compare the extracted speaker embeddings with the cluster-average speaker representation vectors.

Crucially, the weight of each scale at each time step is determined using a scale weighting mechanism, where the scale weights are derived from 1-D convolutional neural networks (CNNs) applied to the multi-scale speaker embedding inputs and the cluster average embeddings. The calculated scale weights are then applied to cosine similarity values computed for each speaker and each scale. The scale weights $\alpha_s\,[t]$ at time step t are calculated using 1-D CNNs applied to the multi-scale embeddings and cluster-average embeddings:

$$\alpha_s\,(t) = CNN_s\,(E_s\,[t], C_k) \tag{11}$$

Received: July 10, 2024. Revised: August 7, 2024.

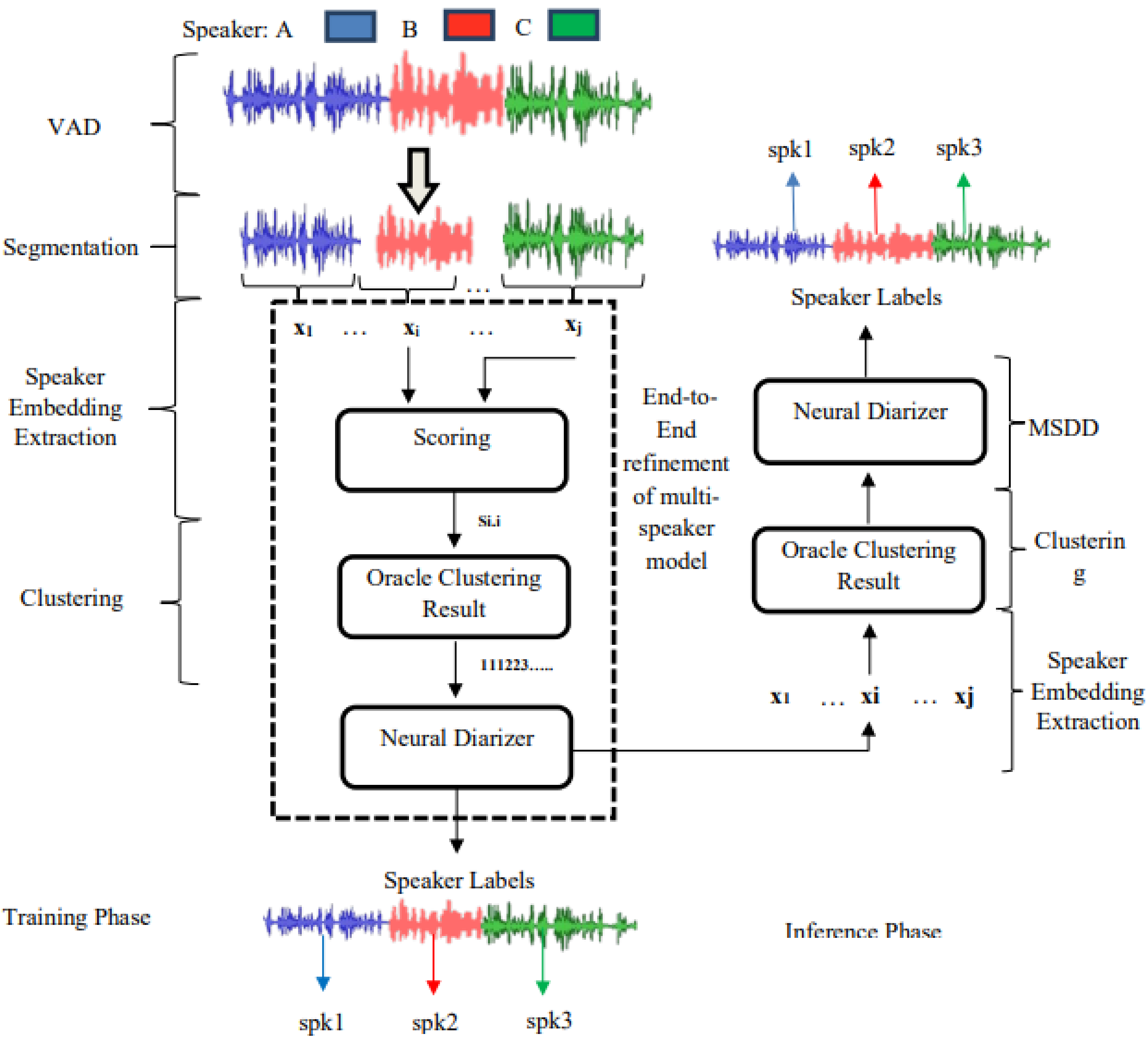

Figure. 1 Framework of EEND Multi-Scale Approach

where C_k is cluster-average embedding for cluster k and CNN_s represents the 1-D CNN applied at scale s. The process is illustrated by applying the estimated scale weights on the cosine similarity between the cluster-average speaker embedding and the input speaker embeddings. Aside from the CNN-based weighting scheme, a conv_scale_weight approach utilizes 1-D CNN filters to calculate scale weights. Finally, each context vector for each step is fed into a multi-layer LSTM model that generates per-speaker existence probabilities.

5. M-Diarization dataset

The creation of the M-Diarization dataset marked a significant milestone in the field of speech technology and natural language processing by introducing the first-ever speech conversation dataset for the Myanmar language. Intended as a benchmark, the dataset begins with two-speaker conversations as a baseline [10] and later expands to include multi-speaker interactions in real-time scenarios [11]. This research utilized the updated M-Diarization dataset, maintaining the original settings while adding audio files from real-time conversations, such as recorded Facebook Live session discussions, YouTube stream discussions, and UCSY seminar meetings.

Updating the M-Diarization dataset introduces the first Myanmar multi-speaker dataset and enhances real-world diarization research with recordings from live events and seminars. Moreover, the dataset features unlike other datasets with fixed-length audio segments.

5.1 Dataset collection procedure

The process started by acquiring the original video files and converting them into wave file format.

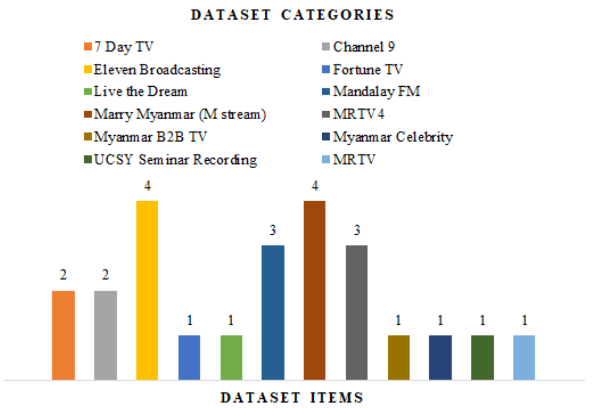

Figure. 2 Details statistics of dataset categories and items

This conversion ensured a consistent 16 kHz sample frequency and mono channel configuration. Next, textGrid files were generated from the wave files using the Praat toolkit [12]. And then applied to convert the textGrid files into RTTM files. These RTTM files are essential for speaker diarization and related speech processing tasks. Fig. 2 displays distinct categories within the dataset.

5.2 Dataset size

The dataset comprises 41 hours of audio recordings with a total of 443 speakers, including 200 female speakers and 243 male speakers. The dataset also includes a total of 1677 utterances. Further details can be found in Table 2 below.

In Fig. 3 illustrates the variation in the number of speakers in the M-Diarization dataset.

This dataset focuses on real-world scenarios often involve diarization and speaker recognition in multi-speaker conversations, this approach lays the groundwork to establish as a fundamental basis for tackling more complex situations.

Table 2. Details statistics of dataset size

Dataset Size	No. of Speakers			No. of Utterances
	Female	***Male***	***Total***	
41 hours	200	243	443	1677

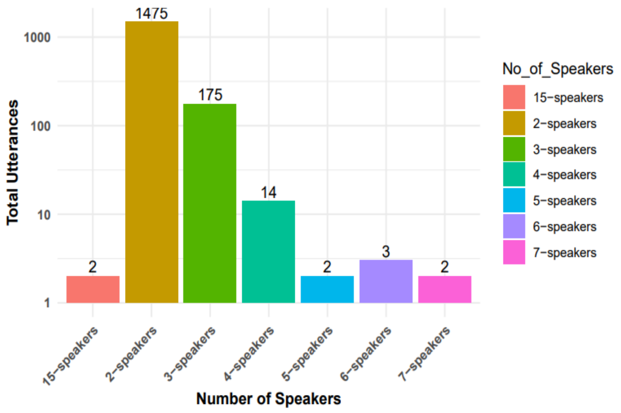

Figure. 3 Varying numbers of speakers

Table 3. Details overlap ratio for M-Diarization dataset

Dataset	Overlap Ratio (%)
Train	10
Development	0.01
Testset1	0.03
Testset2	0
Testset3	3.31

5.3 Overlap ratio

The metric is essential for accurately segmenting and identifying speakers in multi-speaker scenarios, with natural overlaps. In this experiment, the overlap

$$O = \frac{T_{total}}{T_{overlap}} \times 100\ \% \tag{12}$$

where $T_{overlap}$ is the total duration of overlapping speech, T_{total} is the total duration of the segment. The overlap ratio details for the M-Diarization dataset are shown in Table 3.

6. Main text

This section covers the AMI meeting corpus, focusing on overlap ratio calculation and utilizing i-vectors and x-vectors as fundamental components.

It provides a comprehensive analysis of training methodologies for end-to-end neural networks, addressing both single-scale and multi-scale frameworks, and includes a rigorous evaluation of performance metrics.

6.1 AMI meeting corpus

AMI meeting corpus [13] included with a 19.4% overlap ratio. This experiment applies the AMI meeting corpus to evaluate a proposed method across different languages.

The corpus, comprising 100 hours of multi-modal meeting recordings, is preferred for developing speaker diarization systems due to its diverse scenarios in academic and professional domains. This study specifically utilizes 10 hours of close-talking meeting recordings from AMI, each featuring fixed configurations of four speakers.

6.2 I-vectors and X-vectors implementation

I-vectors condense speaker characteristics into a low-dimensional space using Gaussian Mixture Models (GMMs) for feature extraction.

Table 4. Implementations of I-vectors and X-vectors

Aspect	I-vector	X-vector
Model Basis	GMM-UBM	DNN
Dimensionality	128	512
Speaker Discrimination	Low robust	More robust
Feature Extraction	GMM features	DNN features

In contrast, x-vectors utilize deep neural networks (DNNs) to provide more robust feature representations, capturing complex patterns in speech. The implementations are shown in the following Table 4.

6.3 EEND (Single scale approach)

The implementation begins with SincNet [14] for detailed audio feature extraction, followed by two stacked bidirectional LSTM layers, each with 128 units in both directions, as part of the EEND single-scale approach.

Temporal pooling is intentionally omitted to preserve temporal dynamics and long-term dependencies crucial for tasks like audio stream analysis. The subsequent two feed-forward layers, each with 128 units and tanh activation, transform and abstract learned features. The final classification layer, consisting of two units with softmax activation, optimizes categorical classification tasks based on preserved temporal context. The experiment utilized a sliding window with a size of 5 seconds and a shift of 0.5 seconds [15]. For segmentation, a collar tolerance duration of 0.00 seconds required segments to precisely match start or end times to be considered consecutive or overlapping. The pre-trained pipeline employed segmentation (threshold 0.4442), clustering (threshold 0.7154, method: Centroid), and specified minimum duration off (0.5817). The fine-tuned pipeline adjusted segmentation (threshold 0.5895) and clustering (threshold 0.6422) parameters to enhance performance.

6.4 EEND (Multi-scale approach)

In this experiment, Multilingual Marblenet for VAD, Titanet Large for generating speaker embeddings, and Diarization MSDD [8-9, 15-16] neural networks are used. The standard setup incorporates five embeddings ranging from 0.5 to 1.5 seconds. In simpler scenarios, a single embedding with a longer duration, such as 1.0 or 1.5 seconds, may be adequate. However, in more complex situations involving unknown number of speakers, the use of multi-scale embeddings significantly improves diarization precision.

Table 5. Training parameters for EEND multi-scale

Model	Parameter Name	Value
General	Input sample rate	16 000
	Batch size	16
VAD	Window length	0.8
	Shift length	0.04
	Pad onset	0.1
	Pad offset	-0.05
Speaker embedding (5 scales)	Window length	[1.5,1.25,1.0,0.75,0.5]
	Shift length	[0.75,0.625,0.5,0.375,0.25]
Speaker embedding (6 scales)	Window length	[3.0,2.5,2.0,1.5,1.0,0.5]
	Shift length	[1.5,1.25,1.0,0.75,0.5,0.25]

In our experiments, we employed six embeddings, extending from 0.5 to 3.0 seconds. Detailed training parameters are shown in Table 5.

In our work, encompasses two main diarization methods: Cluster Diarizer and Neural Diarizer.

6.4.1. Cluster diarizer

Cluster Diarizer is typically evaluated using a default collar of 0.25 seconds without overlap. The evaluation setting focuses on assessing non-overlapping segments around boundaries, ensuring accurate segmentation without considering overlaps. The method effectively identifies and separates speakers based on distinct embeddings, enhancing speaker diarization in various audio processing applications [8].

6.4.2. Neural diarizer

The Neural Diarizer employs a MSDD coupled with Oracle VAD to enhance speaker diarization accuracy [8]. A critical parameter, sigmoid_threshold, determines the sensitivity to speech overlaps: lower values increase sensitivity, potentially reducing false alarms and missed detections but also risking over-segmentation. The default sigmoid_threshold is set to 0.7 for telephonic models; setting it to 1.0 results in detecting no overlap speech, suitable for tasks requiring precise segmentation. The flexible approach adapts to different speech scenarios, optimizing diarization performance based on specified threshold settings.

6.5 Evaluation metric

In this research, Diarization Error Rate (DER) is a key metric for evaluating speaker diarization systems by measuring segmentation and speaker labelling errors within an audio dataset [17]. It is typically calculated using the following equation:

$$DER = \frac{100\times(Misses+FA+Overlaps)}{Total\ Speech\ Segments} \tag{13}$$

where *Misses* represent segments where a speaker change is missed. *False Alarms* denote segments incorrectly labelled as speaker changes. *Speaker Overlaps* indicate segments where more than one speaker is detected simultaneously. *Total Speech Segments* refer to the total duration of speech segments in the evaluation dataset.

7. Experimental results and analysis on AMI corpus

This section provides a comparative analysis of baseline and proposed methods using the AMI Meeting Corpus [13]. It included with a 19.4% overlap ratio. This experiment applies the AMI meeting corpus to evaluate a proposed method across different languages.

The corpus, comprising 100 hours of multi-modal meeting recordings, is preferred for developing speaker diarization systems due to its comprehensive coverage of diverse scenarios in academic and professional domains. This study specifically utilizes 10 hours of close-talking meeting recordings from AMI, each featuring fixed configurations of four speakers.

7.1 Experimental setup

The corpus is characterized by a 19.4% overlap ratio. The training set comprises 28 audio files, totalling 8 hours and 46 minutes, featuring 112 speakers. The development set includes 3 files with a duration of 56 minutes, involving 12 speakers. Similarly, the test set contains 3 files, also spanning 56 minutes, with 12 speakers. Each audio file consistently features 4 speakers.

7.2 Experimental results

The experimental results compare our 5-scale and 6-scale cluster diarizer models, serving as a baseline, and the EEND single-scale approach models. Fig. 4 presents these comparative results, focusing on DER percentages. This comparison provides a clear highlight of DER percentages across various methods and pipelines in speaker diarization, demonstrating the effectiveness of our approach as applied to the English AMI meeting corpus.

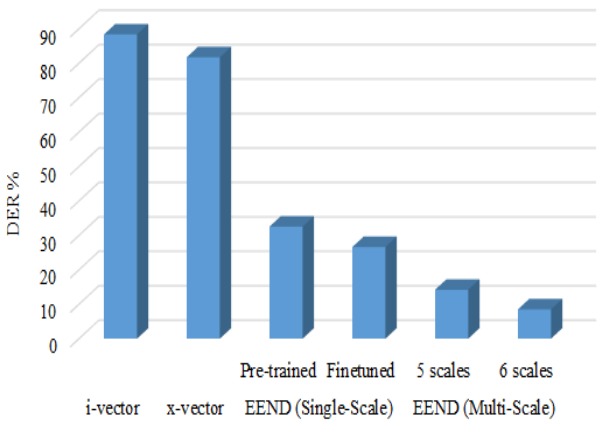

Figure. 4 Comparative results for English AMI corpus

Firstly, traditional methods such as i-vector and x-vector exhibit higher error rates of 88.32% and 81.69%, respectively. These methods rely on handcrafted features and statistical modelling, which can lead to higher error rates compared to end-to-end approaches.

Secondly, the EEND model operating at a single scale demonstrates improvements over traditional methods. The pre-trained configuration achieves a DER of 32.5%, while the Finetuned version further reduces this to 26.6%. The underscores the effectiveness of neural networks in learning discriminative features directly from data, thereby enhancing diarization accuracy.

Additionally, the proposed EEND model implemented with multiple scales using Cluster Diarizer, specifically using a collar of 0.25 seconds without overlap, shows significant advancements in diarization performance. The configuration using 5 scales achieves a DER of 14.18%, and scaling up to 6 scales further reduces the DER to 8.5%. The proposed multi-scale representations of 6 scales enables the model to capture temporal and hierarchical dependencies more effectively, leading to superior speaker segmentation and clustering.

Furthermore, Table 6 provides detailed results for the EEND multi-scale Neural Diarizer, showcasing its performance in both the 5-scale and 6-scale configurations.

In the context of the Neural Diarizer using EEND with 6 scales and a threshold of 0.7, the results demonstrate superior performance compared to other configurations. Specifically, at this threshold and scale setup, DER is notably lower, achieving 13.07%.

The results indicate that the model effectively balances the trade-off between sensitivity and specificity in speaker segmentation, leading to more

Table 6. The performance of multi-scale EEND on AMI corpus

Proposed Method	Multi Scales	DER %					
Neural Diarizer		Threshold: 0.7			Threshold: 1.0		
EEND (Multi-Scale)		collar 0.25 sec, without overlap	collar 0.25 sec, with overlap	collar 0.0 sec, with overlap	collar 0.25 sec, without overlap	collar 0.25 sec, with overlap	collar 0.0 sec, with overlap
EEND (Multi Scale)	5 scales	13.87	18	24.21	12.31	16.62	23.16
	6 scales	**13.07**	**17.29**	**23.38**	**12**	**16.29**	**23**

accurate identification and clustering of speakers within audio recordings.

Comparatively, at 6 scales with a threshold of 1.0, the DER increases to 17.29%, indicating a less precise segmentation compared to the 0.7 threshold setting. Similarly, variations in collar settings (0.25 sec with or without overlap) and scale adjustments show fluctuations in DER, but the 6 scales with a threshold of 0.7 consistently stands out with the lowest error rate among the presented options.

In our experiments, the findings highlight the efficacy of configuring the Neural Diarizer with 6 scales and a threshold of 0.7 for achieving enhanced accuracy in speaker diarization tasks.

7.3 Analysis and discussion

Building on the results of our experiments, the findings underscore significant advancements in speaker diarization achieved through the EEND model with multi-scale configurations. The notable reduction in DER to as low as 8.5% with 6 scales using Cluster Diarizer demonstrates the capability to effectively capture complex temporal and hierarchical dependencies inherent in audio data. The improvement over traditional methods like i-vector and x-vector, which exhibited much higher error rates due to their reliance on manually crafted features, highlights the transformative impact of neural network-driven approaches in speaker diarization tasks.

Additionally, the sensitivity analysis across different thresholds and collar settings within the Neural Diarizer further emphasizes the optimal performance of the EEND model at 6 scales with a threshold of 0.7, ensuring robust speaker segmentation and clustering. These results not only validate the efficacy of multi-scale representations but also pave the way for enhanced applications in real-world scenarios.

8. Experimental results and analysis on M-Diarization dataset

This section offers a comparative evaluation of the baseline and proposed techniques in multi-speaker environments utilizing the Myanmar M-Diarization dataset.

8.1 Experimental setup

The training set comprises 1673 audio files, totalling 41 hours, and includes recordings from 443 speakers. The development set consists of a single 30-minute file involving 4 speakers, designed to assess multi-speaker interactions. Testset1 features a 25-minute file with 2 speakers. Testset2 includes a 13-minute Zoom interview with unbalanced dialogue. Testset3 comprises a 31-minute recording of a UCSY seminar, notable for its high overlap.

8.2 Experimental results

A comparative analysis of the proposed models against baseline clustering methods and the EEND single-scale approach was conducted, evaluating various speaker configurations within the M-Diarization dataset. The results of this comparative study are illustrated in Fig. 5. This experiment provides detailed insights into DER percentages for both state-of-the-art methods and our proposed models, showcasing the effectiveness of our approach across different speaker configurations, as outlined in our research survey.

Fig. 5 provides a comprehensive overview of speaker diarization methods across different test sets, highlighting their performance under various conditions. In the development set, i-vector achieving 1.72%, indicating effective speaker separation. In contrast, x-vector shows a significantly higher error rate of 24.67%, suggesting less precise performance in this controlled setting. The EEND models, both single - scale and multi - scale, show substantial

Table 7. The performance of multi-scale EEND on M-Diarization dataset

Proposed Method	Multi Scales	DER %					
Neural Diarizer		Threshold: 0.7			Threshold: 1.0		
EEND (Multi-Scale)		collar 0.25 sec, without overlap	collar 0.25 sec, with overlap	collar 0.0 sec, with overlap	collar 0.25 sec, without overlap	collar 0.25 sec, with overlap	collar 0.0 sec, with overlap
Dev	5 scales	0.07	0.07	0.12	0.05	0.05	0.1
	6 scales	**0.04**	**0.04**	**0.1**	**0.04**	**0.04**	**0.1**
Testset1	5 scales	2.64	2.64	2.74	2.7	2.7	2.79
	6 scales	**2.62**	**2.62**	**2.4**	**2.62**	**2.62**	**2.4**
Testset2	5 scales	0.05	0.05	0.12	0.03	0.03	0.07
	6 scales	**0**	**0**	**0.08**	**0**	**0**	**0.08**
Testset3	5 scales	9.9	11.42	14.16	9.16	10.87	13.65
	6 scales	**2.45**	**4.37**	**6.72**	**2.45**	**4.37**	**6.72**

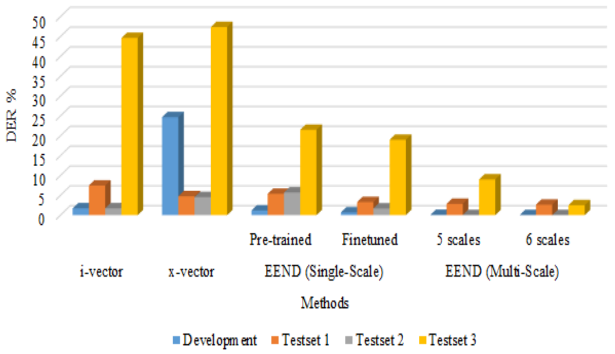

Figure. 5 Comparative results for Myanmar M-Diarization dataset

improvements through pre-training and fine-tuning, with multi-scale variants achieving remarkably low error rates as fine as 0.03%.

Moving to the test sets, each presents unique challenges. Testset1, which exhibits less speaker overlap, poses a challenge to traditional methods like i-vector and x-vector, resulting in higher error rates of 7.47% and 4.75%, respectively. This scenario highlights the difficulty in accurately distinguishing speakers when overlap is minimal. Testset2, characterized by an unbalanced speaker distribution, challenges diarization methods differently. Despite the imbalance, EEND models perform robustly, particularly at multiple scales, showcasing their ability to handle varied speaker distributions effectively with error rates as low as 0.03%.

Testset3 represents the most complex scenario with high speaker overlap, where traditional methods like i-vector and x-vector struggle significantly, showing error rates of 44.63% and 47.38%, respectively. In contrast, EEND models demonstrate their superiority in handling such challenges, achieving error rates ranging from 2.5% in 6 scales. The capability underscores the advantage of EEND multi-scale approach in capturing and separating speakers amidst overlapping speech patterns, making it particularly effective for real-world applications where speaker interactions are intricate and varied. Overall, the findings underscore the importance of adaptive diarization methods like EEND in addressing the diverse complexities encountered across different speaker diarization tasks.

Table 7 presents the results of the proposed Neural Diarizer (5-scale and 6-scale), detailing its performance across various evaluation sets and conditions. The table compares the multi-scale Neural Diarizer effectiveness at two thresholds (0.7 and 1.0) and different collar settings (0.25 seconds and 0.0 seconds for overlap handling). The analysis underscores the 6-scale configuration consistent and superior performance across diverse scenarios.

In the development set, for instance, the 6 scales setup achieves a remarkably low DER of 0.04 with a 0.25-second collar, whether handling overlap or not. This indicates precise and reliable speaker segmentation, showcasing the effectiveness in controlled environments.

The test sets, the 6 scales configuration maintains strong performance across different challenges. In Testset1, it sustains a DER of 2.62% under both collar settings, demonstrating stable performance in a dataset with moderate speaker overlap. Testset2 showcases the capability further, achieving a perfect

DER of 0% with a 0.25-second collar without overlap, highlighting its robustness in scenarios where speakers are distinctly separated. Even in Testset3, known for high speaker overlap, the 6 scales configuration manages a competitive DER of 2.45% with a 0.25-second collar, indicating effective speaker separation despite challenging conditions. The experiment results show the significant advantage when employing the 6 scales configuration in the Neural Diarizer model.

8.3 Analysis and discussion

The results highlight the proposed EEND multi-scale models outperform, especially in challenging contexts characterized by high speaker overlap. While i-vector and x-vector demonstrate effectiveness in controlled settings with low speaker overlap, as seen in the development dataset, their performance falters significantly in environments like Testset3, where speaker interactions are complex and overlap is prevalent. The limitation is evident in their markedly higher error rates compared to EEND models, particularly those employing multiple scales. The EEND models consistently outperform traditional methods, showcasing superior adaptability and accuracy across all evaluation sets. The 6 scales configuration within the EEND framework stands out for its ability to achieve remarkably low error rates even in the face of challenging speaker overlap scenarios, underscoring its efficacy in accurately capturing and separating speakers amidst varying speech patterns.

9. Conclusion and future works

In this study, we conducted a thorough evaluation of traditional clustering-based methods, such as i-vectors and x-vectors, alongside EEND methods, focusing on their capacity to handle speaker overlap and distribution. We applied these techniques to two datasets: the Myanmar M-Diarization dataset and the English AMI meeting corpus, both featuring diverse languages and speaker configurations. According to the comparative results that while traditional methods like i-vectors and x-vectors performed well in controlled environments, they struggled with complex scenarios involving significant speaker overlap. And the single-scale EEND approach outperformed conventional methods. However, the single-scale EEND approach faces limitations when dealing with an unknown number of speakers. In contrast, EEND models, particularly those employing multi-scale approaches (5-scale and 6-scale), demonstrated enhanced adaptability and precision. Additionally, our study introduced the first M-Diarization dataset and implemented a flexible speaker count system, validating the performance of our proposed model with high-overlap test sets and the AMI meeting corpus. The 6-scale EEND configuration notably outperformed traditional methods in terms of DER in high-overlap conditions. These findings highlight the effectiveness of multi-scale approaches in improving diarization accuracy. Our work underscores the significant contribution of EEND multi-scale models in advancing speaker separation in challenging environments. Future research should explore the application of EEND models in real-time online diarization systems to further assess their effectiveness in dynamic scenarios.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

The first author led the conceptualization, methodology design, and software development, and conducted the formal analysis and investigation. Additionally, managed resources, curated the data, and prepared the original draft of the manuscript, while also handling visualization tasks. The second authors provided overall supervision, contributed to the validation of results, and was involved in the review and editing of the manuscript. Furthermore, they managed project administration. All authors have reviewed and approved the final version of the manuscript.

Acknowledgments

The authors express gratitude to colleagues at the Natural Language Processing Lab (NLP) of the University of Computer Studies, Yangon (UCSY), Myanmar, for their invaluable assistance in dataset preparation and technical support. Moreover, this work is the output of the ASEAN IVO (http://www.nict.go.jp/en/asean_ivo/index.html) project, "Spoof Detection for Automatic Speaker Verification", and financially supported by NICT (http://www.nict.go.jp/en/index.html).

References

[1] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker Diarization Using Deep Neural Network Embeddings", In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017.

[2] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland and O. Vinyals, "Speaker Diarization: A Review of Recent Research", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 2, pp. 356-370, 2012.

[3] G. Sell and D. Garcia-Romero, "Speaker diarization with plda i-vector scoring and unsupervised calibration", In: *Proc. of 2014 IEEE Spoken Language Technology Workshop (SLT)*, South Lake Tahoe, NV, USA, pp. 413-417, 2014.

[4] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey and S. Khudanpur, "Speaker Recognition for Multi-speaker Conversations Using X-vectors", In: *Proc. of ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, pp. 5796-5800, 2019.

[5] Y. Fujita1, N. Kanda, S. Horiguchi1, K. Nagamatsu1 and S. Watanabe, "End-to-End Neural Speaker Diarization with Permutation-Free Objectives", In: *Proc. of INTERSPEECH 2019*, Graz, Austria, 2019.

[6] C. Wang, J. Li, X. Fang, J. Kang, Y. Li, "End-to-End Neural Speaker Diarization with Absolute Speaker Loss", In: *Proc. of INTERSPEECH 2023*, 3577-3581, 2023.

[7] P. Vecchiotti, G. Pepe, E. Principi, S. Squartini, "Detection of activity and position of speakers by using deep neural networks and acoustic data augmentation, "Expert Systems with Applications", *Expert Systems with Applications*, Vol. 134, Pages 53-65, 2019.

[8] N. R. Koluguri, T. Park and B. Ginsburg, "TitaNet: Neural Model for Speaker Representation with 1D Depth-Wise Separable Convolutions and Global Context", In: *Proc. of ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, pp. 8102-8106, 2022.

[9] T. J. Park, N. R. Koluguri, J. Balam and B. Ginsburg, "Multi-scale Speaker Diarization with Dynamic Scale Weighting", In: *Proc. of Interspeech 2022* 18-22, Incheon, Korea, 2022.

[10] M. A. A. Aung, W. Pa Pa and H. M. Soe Naing, "M-Diarization: A Myanmar Speaker Diarization using Multi-scale dynamic weights", In: *Proc. of 2023 26th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, Delhi, India, pp. 1-5, 2023.

[11] M. A. A. Aung, W. P. Pa and H. M. S. Naing, "End-to-End Sequence Labeling for Myanmar Speaker Diarization", In: *Proc. of 2024 IEEE Conference on Computer Applications (ICCA)*, Yangon, Myanmar, pp. 1-6, 2024.

[12] J. M. Garrido, "A Praat-Based Tool for the Automatic Detection and Annotation of Prosodic Boundaries in Speech Corpora", In: *Proc. of TRASP'2013*, Aix-en-Provence, France, 2013.

[13] J. Carletta, S. Ashby, S. Bourban, M. J. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, L. McCowan, W. Post, D. Reidsma, P. Wellner, "The AMI Meeting Corpus: A Pre-announcement", *In: Renals, S., Bengio, S. (eds) Machine Learning for Multimodal Interaction. MLMI 2005. Lecture Notes in Computer Science*, Vol.3869, 2005.

[14] M. Ravanelli and Y. Bengio, "Speaker Recognition from Raw Waveform with SincNet", In: *Proc. of 2018 IEEE Spoken Language Technology Workshop (SLT)*, Athens, Greece, pp. 1021-1028, 2018.

[15] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, M. P. Gill, "Pyannote.Audio: Neural Building Blocks for Speaker Diarization", In: *Proc. of ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, pp. 7124-7128, 2020.

[16] T. J. Park, M. Kumar and S. Narayanan, "MULTI-SCALE SPEAKER DIARIZATION WITH NEURAL AFFINITY SCORE FUSION", *arXiv: 2011.10527v1 [eess.AS]*, 2020.

[17] O. Galibert, "Methodologies for the evaluation of Speaker Diarization and Automatic Speech Recognition in the presence of overlapping speech", In: *Proc. of Interspeech.2013-303*, 2013.