# Speaker Diarization for Multiple Speaker datasets using a neural diarizer

Myat Aye Aye Aung
Faculty of Computer Science, Natural Language Processing Lab
*University of Computer Studies, Yangon*
Yangon, Myanmar
myatayeayeaung@ucsy.edu.mm

Win Pa Pa
Faculty of Computer Science, Natural Language Processing Lab
University of Computer Studies, Yangon
Yangon, Myanmar
winpapa@ucsy.edu.mm

Hay Mar Soe Naing
Faculty of Computer Science, Natural Language Processing Lab
University of Computer Studies, Yangon
Yangon, Myanmar
haymarsoenaing@ucsy.edu.mm

***Abstract*—Overlapped speech is a challenge in speaker diarization systems, especially when dealing with multiple speakers. Implementing overlapped-aware speech resolution can significantly enhance the performance of speaker diarization systems. This system comprises three Myanmar language speaker datasets, each featuring 2-speaker, 3-speaker, and 4-speaker scenarios. The conversations are unscripted and spontaneous, particularly in live interviews, social trend news coverage, Zoom meetings conducted during the COVID-19 epidemic, and keynote speeches, all occurring in natural settings with background noise, rather than controlled studio environments. These datasets featured instances of overlapped speech, arising from the natural conversations involving two or more speakers. To tackle the challenge of overlapped speech, neural model such as the neural diarizer was utilized in this experiment. In this study, the primary contribution involves showcasing multiple speakers by utilizing two multi-scale weight values, denoted as baseline parameters P1 and P2. Furthermore, a neural diarizer, which employs the Multiscale diarization decoder model, is employed to analyze natural conversations from three separate Myanmar speaker datasets. Parameter P2, with a base scale of 0.5 seconds, shift length of 0.25 seconds, and 5 scale weight values, outperforms the baseline parameter P1. Additionally, the overlapped-aware detection in distinct speaker datasets, involving three or more speakers, outperforms that of two speakers when using the multiscale diarization decoder model than cluster diarizer. The neural diarizer operates with three different settings: with and without overlap 0.25 sec collar and with overlap 0.0 sec collar. The system achieves 12.26% with and without overlap (0.25 sec) Diarization Error Rate (DER) in three-speaker scenario. In four-speaker scenario, the system achieves 5.54% with and without overlap (0.25 sec) DER.***

***Keywords— muti speakers, neural diarizer, multiscale diarization decoder***

## I. INTRODUCTION

The process of determining “who said when” in an input audio is known as speaker diarization. It plays a fundamental role in comprehending spoken language within multi-speaker conversations across diverse scenarios, encompassing everyday dialogues, doctor-patient interactions, meetings, lectures, and video content [1][2]. Diarization systems employ voice activity detection (VAD) to identify segments that are likely to belong to the same speaker [3]. Additionally, they are exploring end-to-end algorithms that can directly map input speech into a sequence of speaker labels for each time segment [4]. Traditionally, its development has primarily based on the clustering of speaker embeddings. However, such clustering-based methods have a number of problems. To begin with, they cannot be directly optimized to minimize diarization errors, as the clustering process falls under the category of unsupervised learning methods. Furthermore, they face difficulties when dealing with speaker overlaps, as clustering algorithms inherently assume a single speaker per segment. To address these issues, an end-to-end neural diarizer can effectively manage overlapped speech [5].

Modern speaker diarization systems, such as end-to-end approaches or those utilizing target-speaker voice activity detection, make use of frame-level speaker labels as part of their methodology. These systems typically rely on sequence models like Long Short-Term Memory (LSTM) networks or Transformer-style encoder-decoder architectures [5][6]. These methods offer the advantage of overlap-aware diarization, where in the neural network model can produce multiple speaker label outputs. Several studies have tackled multi-scale value speaker diarization, utilizing window lengths of (1.5, 1.25, 1.0, 0.75, 0.5) seconds and shift lengths of (0.75, 0.625, 0.5, 0.375, 0.1) for segmentation and constructing a multi-scale diarization decoder model [7].

In our experiments, applied two types of multi-scale speaker embeddings for segmentation, and designed a neural diarizer (multi-scale diarization decoder) to facilitate the following capabilities: handling overlap-aware diarization in multi-speaker scenarios with natural speech and enhancing diarization performance. This system discovered the significance of the parameters related to multi-scale values in speaker embeddings and observed that multi-scale diarization decoder (MSDD) models exhibit enhanced performance, particularly in scenarios involving three or more speakers.

## II. NEURAL DIARIZER

The term "neural diarizer" describes trainable neural modules responsible for estimating speaker labels based on provided audio features or inputs. In contrast, a clustering diarizer differs in that it is not a trainable module. The utilization of a neural diarizer is essential for achieving overlap-aware diarization, enhancing accuracy, and facilitating joint training with speaker embedding models using multi-speaker datasets, also known as diarization training datasets. In this study, the default evaluation settings for the neural diarizer encompass three different configurations: Firstly, with a collar of 0.25, and the default parameter "ignore_overlap" set to True, reflecting the standard setup for evaluating clustering diarizers. Secondly, with a collar of 0.25 and "ignore_overlap" set to False, where there is still a 0.25-second margin around boundaries that is not evaluated, but overlaps are assessed. Lastly, with a collar

of 0.0 and "ignore_overlap" set to False, indicating no collar is applied at all, and overlaps are considered during the evaluation process [7]. The system design is shown in the following Figure I.

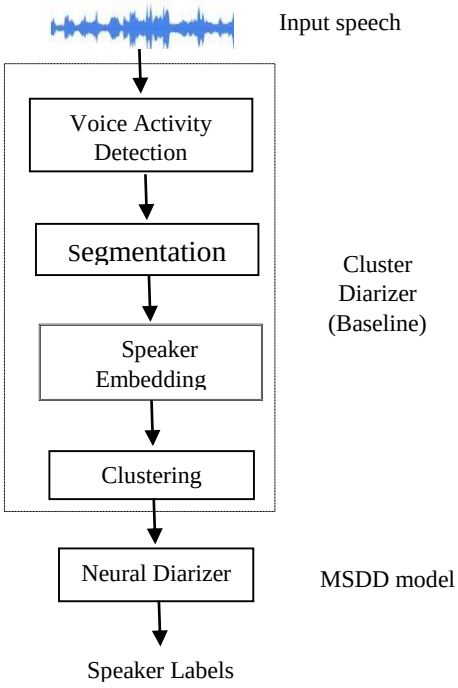


Figure I. System design of Neural Diarizer

### A. *Multi Scale Diarization Decoder (MSDD)*

The multi-scale approach is introduced to mitigate this trade-off by extracting speaker features from various segment lengths and then merging the results obtained from multiple scales. It is accomplished by implementing multi-scale segmentation and extracting speaker embeddings at each scale [7][8]. MSDD takes multiple speaker embedding vectors derived from various scales and subsequently calculates the appropriate scale weights. Speaker labels are then generated based on these estimated scale weights. Multi-scale audio segments are extracted from the input audio, and the corresponding speaker embedding vectors for these multi-scale audio segments are generated using the speaker embedding extractor (TitaNet) [9]. Subsequently, the extracted multi-scale embeddings are subjected to a clustering algorithm, which generates an initial clustering result used as input for the MSDD module. The MSDD module then employs cluster-average speaker embedding vectors to compare them with the input speaker embedding sequences. The estimation of scale weights is carried out at each step to determine the significance of each scale. Ultimately, the sequence model is trained to produce probabilities for speaker labels associated with each speaker [7][10].

A neural network model known as the multi-scale diarization decoder (MSDD) is trained to leverage the multi-scale approach by dynamically computing the weight of each scale [11]. MSDD utilizes the initial clustering results and compares the extracted speaker embeddings with the cluster-average speaker representation vectors. The estimated scale weights are used to adjust the cosine similarity values computed for each speaker and each scale. This adjustment involves calculating the context vector by applying the estimated scale weights to the cosine similarities computed between the cluster-average speaker embedding and the input speaker embeddings [12].

## III. DATASETS

This section offers three types of datasets for the Myanmar Language, including two-speaker, three-speaker, and four-speaker conversational datasets. All dataset encompasses a wide range of conditions and diverse circumstances. The primary reason for this diversity was the presence of interruptions in communication channels during conversations, environmental noises, frequent speaker changes, disfluencies, short speech segments, and various recording configurations at both ends.

This system have focused on simulating three types of datasets conversation scenarios by using interview, meeting, discussion conversations between two or more individuals in standard speaking conditions. While real-world scenarios frequently entail diarization and speaker recognition in multi-speaker conversations, this approach serves as a foundation for low-resource languages to establish fundamental groundwork for handling more complex situations. And collecting datasets with three or more speakers is more challenging compared to those with two speakers, primarily due to no enough data. Consequently, datasets containing three or four speakers typically have smaller sizes when compared to two-speaker counterparts.

### A. *Datasets Collection and Preparation*

The data has been gathered from publicly accessible sources, including Facebook pages, YouTube channels, and official Myanmar websites. The speaking style in these sources is spontaneous and free-form. The audio files comprise segments with non-overlapping speech, occasionally featuring instances of speech overlap in conversations involving two or more speakers. The dataset categories encompass live interviews and discussions covering a wide range of topics, including career, social issues, health, daily life, family matters, work-related discussions, beauty, diet for health, and panel discussions.

In dataset preparation, initially in original video format, were subsequently converted to the wave file format. Following this conversion, the audio files underwent segmentation using Audacity [1] . All audio files were standardized to a sample frequency of 16,000 Hz and a mono channel. The next step involved utilizing Praat[2] to generate a TextGrid file for each speaker segment from the segmented audio. Finally, these TextGrid files were converted into RTTM file format to create ground truth labels for each of the formatted audio files. Detailed information is presented in the following tables.

Table I. Detail Information of Multiple speaker dataset

| **No. of Speakers** | **Data Size** | **No. of Speaker** | | | **No. of Utterance** |
|---|---|---|---|---|---|
| | | ***Female*** | ***Male*** | ***Total*** | |
| Two-speaker | 3 hours | 55 | 75 | 130 | 700 |
| Three-speaker | 1 hour | 35 | 56 | 91 | 332 |
| Four-speaker | 30 mins | 6 | 16 | 22 | 20 |

## IV. EXPERIMENTAL RESULTS

This section outlines the experimental setup and presents the results. The datasets employed in these experiments are derived from the two or more speaker dataset in the Myanmar Language. In this experiments, various parameters for multi-

[1] https://www.audacityteam.org/

[2] https://www.fon.hum.uva.nl/praat/

scale segmentation values were utilized, denoted as P1 (baseline) and P2. Table II provides the values for scale weights, window length, and shift length.

Table II. Parameters of mult-scale weight, window length and shift length

| **Parameters** | | **Size 1** | **Size 2** | **Size 3** | **Size 4** | **Size 5** |
|---|---|---|---|---|---|---|
| P1 | Window Length | 1.5 | 1.25 | 1.0 | 0.75 | 0.5 |
| | Shift Length | 0.75 | 0.625 | 0.5 | 0.375 | 0.1 |
| | Multi-scale Weights | 1 | 1 | 1 | 1 | 1 |
| P2 | Window Length | 1.5 | 1.25 | 1.0 | 0.75 | 0.5 |
| | Shift Length | 0.75 | 0.625 | 0.5 | 0.375 | 0.25 |
| | Multi-scale Weights | 1 | 1 | 1 | 1 | 1 |

The proposed system demonstrates competitive performance, especially when evaluating it with two different parameters, particularly those associated with speaker embeddings, and the neural diarizer employing a multiscale diarization decoder model across three types of datasets. This system aims to enhance diarization performance in scenarios involving multiple speakers with overlapped speech. The system was executed using the Nemo Speaker Diarization inference tool from the Nemo toolkit[3]. The detailed results are showed, including the Diarization Error Rate (DER), in the following table III.

Table III. Experimental Results of Multiple speaker datasets

| ***Datasets and Parameters*** | | **Cluster Diarizer** | **Neural Diarizer** | | |
|---|---|---|---|---|---|
| | | ***0.25 sec*** | ***0.25 sec*** | ***0.25 sec*** | ***0.0 sec*** |
| | | ***Without overlap DER (%)*** | ***Without overlap DER (%)*** | ***With overlap DER (%)*** | ***With overlap DER (%)*** |
| Two-speaker | P1 | 13.88% | 13.64% | 13.64% | 17.75% |
| | P2 | 6.64% | **7.53%** | **7.53%** | 8.58% |
| Three-speaker | P1 | 22.98% | 22.76% | 22.76% | 24.88% |
| | P2 | 13.54% | **12.26%** | **12.26%** | 16.33% |
| Four-speaker | P1 | 11.38% | 9.87% | 9.87% | 14.37% |
| | P2 | 8.21% | **5.54%** | **5.54%** | 12.22% |

This study highlights the significance of window length and shift length values in the context of speaker embedding. In both P1 (baseline) and P2, which share identical 5-scale weight values, including the base scale of 0.5 seconds, along with shift lengths of 0.1 seconds and 0.25 seconds, the configuration featuring a base scale of 0.5 seconds and a shift length of 0.25 seconds consistently demonstrates superior performance in the proposed datasets. Moreover, when comparing the performance of the overlapped-aware pre-trained model (MSDD) in three and four-speaker datasets, it outperforms the cluster diarizer.

The experimental results suggest that setting the shift length to half of the window length typically leads to a lower diarization error rate across two parameters. Additionally, it was noted that neural diarizer pre-trained models outperform clustering diarizers, especially in scenarios involving multiple speakers, where conversations tend to have more overlap across various datasets in the Myanmar Language.

[3] https://github.com/NVIDIA/NeMo

## V. Conclusion

In this paper, proposed five scale parameters for multi-scale weight values in the context of three dataset for the Myanmar Language. Gathering datasets featuring three or more speakers proves to be more challenging than acquiring those involving just two speakers because of the limited availability of adequate data. These different datasets are subsequently utilized in the neural diarizer end-to-end model. The datasets consist of real-time live data and include instances of overlapped speech and background noise. This study highlights the importance of parameter values in speaker embeddings with respect to multiscale weight values. Furthermore, it was found that the multiscale diarization decoder neural model demonstrates superior performance, especially in scenarios involving three or more speakers in conversational data. For future research, increasing the dataset size in each category to include real-time conversations and implementing improved methods to address overlapped speech in natural discourse could significantly enhance the performance of Myanmar Speaker Diarization.

## References


1. Xavier Anguera, Chuck Wooters, and Javier Hernando, "Acoustic beamforming for speaker diarization of meetings," TASLP, vol. 15, no. 7, pp. 2011–2022, 2007.
2. Pierre Lison and Raveesh Meena, "Automatic turn segmentation for movie & tv subtitles," in SLT, 2016, pp. 245–25.
3. Gregory Gelly and Jean-Luc Gauvain, "Optimization of rnn-based speech activity detection," TASLP, vol. 26, no. 3, pp. 646–656, 2017.
4. Latane Bullock, Herve Bredin and Leibny Paola Garcia-Perera, "Overlap-aware Diarization: Resegmentation using Neural End-To-End Overlapped Speech Detection", arXiv:1910.11646v1 [eess.AS] 25 Oct 2019.
5. Tae Jin Parka, et al., "A Review of Speaker Diarization: Recent Advances with Deep Learning", Computer Speech & Language, Volume 72, March 2022.
6. Quan Wang, Carlton Downey, Li Wan1 Philip Andrew Mansfield and Ignacio Lopez Moreno, "Speaker Diarization with LSTM", arXiv:1710.10468v7 [eess.AS] 23 Jan 2022.
7. Tae Jin Park, Nithin Rao Koluguri, Jagadeesh Balam and Boris Ginsburg, "Multi-scale Speaker Diarization with Dynamic Scale Weighting", arXiv:2203.15974v1 [eess.AS] 30 Mar 2022.
8. Ruiqing Yin, Herve Bredin, Claude Barras, "Neural speech turn segmentation and affinity propagation for speaker diarization", Interspeech 2018 2-6 September 2018, Hyderabad, India.
9. Nithin Rao Koluguri, Taejin Park, Boris Ginsburg, "TITANET: NEURAL MODEL FOR SPEAKER REPRESENTATION WITH 1D DEPTH-WISE SEPARABLE CONVOLUTIONS AND GLOBAL CONTEXT", arXiv:2110.04410v1 [eess.AS] 8 Oct 2021.
10. Youngki Kwon, et al., "Multi-scale speaker embedding-based graph attention networks for speaker diarizaiton", ICASSP 2022 IEEE International Conference on Acoustics, Speech and Signal Processing.
11. Tae Jin Park, Manoj Kumar and Shrikanth Narayanan, "Multi-scale speaker diarization with neural affinity score fusion", ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
12. M. Senoussaoui, P. Kenny, T. Stafylakis, and P. Dumouchel, "A study of the cosine distance-based mean shift for telephone speech diarization," IEEE/ACM TASLP, vol. 22, no. 1, pp. 217–227, 2013.