ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096327

# MULTI-SPEAKER END-TO-END MULTI-MODAL SPEAKER DIARIZATION SYSTEM FOR THE MISP 2022 CHALLENGE

*Tao Liu, Zhengyang Chen, Yanmin Qian*, Kai Yu**

MoE Key Lab of Artificial Intelligence, AI Institute, X-LANCE Lab, Shanghai Jiao Tong University

## ABSTRACT

This paper presents the design and implementation of our system for Track 1 of the Multi-modal Information based Speech Processing (MISP) 2022 Challenge. We design an end-to-end transformer-based multi-talker system. The transformer backbone is well-suited to capture long-term features, which is crucial for multi-modal speaker diarization in cases where temporal modalities are missing. Besides, we employ several loss functions and image data augmentation techniques to prevent over-fitting during training. Moreover, to further improve the system's performance, we incorporate Inter-channel Phase Difference (IPD) to model the location features and pre-train an ECAPA-TDNN-based model to extract speaker embedding features. Our system achieved a diarization error rate (DER) of 10.82% on the evaluation set, which earned us second place in the audio-visual speaker diarization task of the MISP 2022 challenge.

***Index Terms***— MISP Challenge, Audio-visual, Speaker Diarization

## 1. INTRODUCTION

Speaker diarization, solving the problem of 'who spoke when,' is a combined task of speaker identification and voice activity detection. Speaker diarization has various application scenarios, such as a meeting or telephone recording. The MISP challenge[1, 2] releases a multi-modal dataset whose scenario is a casual home chatting in the living room with two to six speakers. The dataset collects audio-visual data from far, middle, and near fields, which provides a valuable benchmark for the community.

Target-Speaker Voice Activity Detection (TS-VAD) [3] shows excellent performance in multiple speaker diarization challenges. TS-VAD takes speaker embedding, like i-vector, as input to get a personalized VAD output. This architecture shows superior performance in the overlapped speech condition. The speaker embedding is initialed with a fair result and improved by an iterative process. Inspired by end-to-end neural diarization (EEND) [4] and active speaker detection (ASD) [5], the speaker in context is a crucial cue in the multi-talker condition, which can suppress non-talkers, especially when there is only one talker. Based on speaker context, multi-talker TS-VAD takes all speakers as input to train the network jointly.

In our system, we also take similar architecture to the multi-talker TS-VAD. The main differences are as follows. First, we adopt transformer-based architecture to grasp long-term features for single-speaker feature extraction. Second, several losses are designed to regularize the training process, and angular loss is added to encourage the embedding to have a small intra-speaker and large inter-speaker distance. Third, we add Inter-channel Phase Difference (IPD) to model the spatial information and leverage an ECAPA-TDDNN-based [6] model as the speaker embedding module. Finally, our system achieves a 22% relative improvement compared to the baseline. Extensive experiment results are listed in Section 2.5.

*Yanmin Qian and Kai Yu are the corresponding authors.

## 2. METHODS

The system architecture shown in Figure 1 consists of a single-speaker encoder and a multi-speaker decoder. The single-speaker encoder combines different features for each speaker, while the multi-speaker decoder models the speaking relationships among speakers.

### 2.1. Single-speaker Encoder

**Location Encoder** Based on the phase part of STFT, we calculate Inter-channel Phase Difference (IPD) as the spatial feature. Specifically, we calculate the difference between (0,3), (1,4), and (2,5), where the number is the microphone number, and there are six microphones in total.

**Audio Encoder** We extract a filter bank with 80-dimensional features from the first microphone as the audio features by using Hamming windowed frame of 25ms with a shift of 10ms. To match the visual frame sampling rate, audio features are down-sampled from 100Hz to 25Hz through Conv1D.

**Speaker Embedding Encoder** We pre-train the ECAPA-TDNN [6] model on VoxCeleb and CNCeleb with 10,365 speakers. Then we fine-tune the model on the MISP training set with 910 speakers. In the feature extraction module, we take a 192-dimensional feature from the last FC layer as the speaker embedding feature. The first initial embedding is extracted from the visual VAD, and the embedding is updated iteratively until the result converges.

**Visual Encoder** We adopt ResNet34 and TCN as our optical encoder. We do not initial the encoder with any pre-trained parameters. We directly use the face RoI result by the MISP dataset, and the lacking face is filled with zero matrices. Apart from flipping, cropping, or rotation, we design a data augmentation called 'random delete,' which randomly sets 10% to 60% frames to zero matrices to simulate the missing visual modality. In our experiment, we find that this technique can prevent the model from over-fitting.

**Feature Fusion Encoder** We use two encoder blocks with 256 attention units and four heads. Before being fed into the transformer layers, all features will be projected a unified dimension: 128, and four features will be concatenated to 512. IPD features and speaker embedding features will be repeated T times along the time axis. Additionally, we have incorporated a skip connection to link the visual and fused features, as we consider the visual cue is essential.

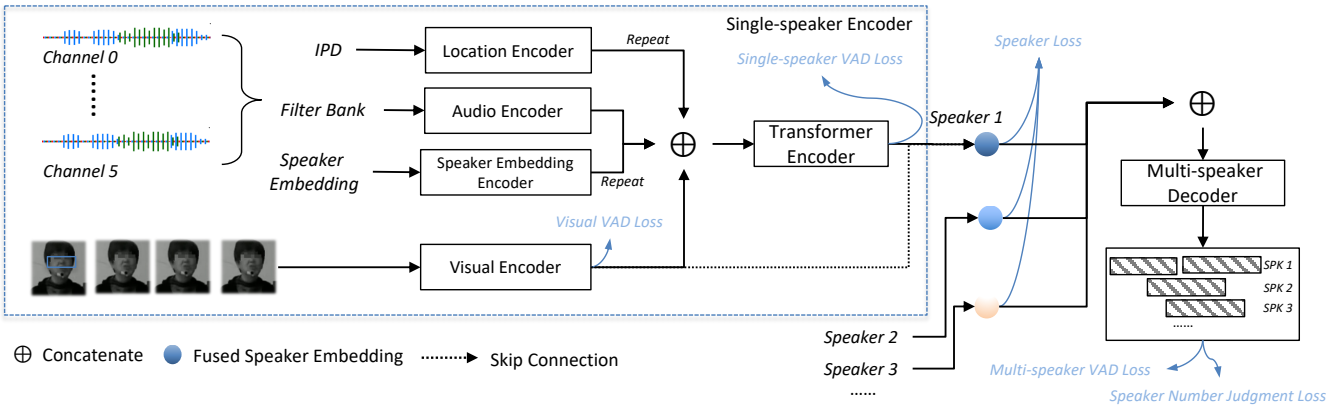


**Fig. 1**. The overview of our multi-speaker end-to-end multi-modal speaker diarization system.

### 2.2. Multi-speaker Decoder

This decoder is a two-layer BLSTM, which learns the speaker context cues by predicting the VAD result and the speaker number.

### 2.3. Losses

We use five losses: visual VAD loss, single-speaker VAD loss, multi-speaker VAD loss, speaker loss, and speaker number judgment loss. We calculate the time-weighted speaker embedding and optimize the embedding via an angular loss: AAM-Softmax. Apart from the speaker loss, other losses are BCE losses that are same to the baseline. The loss weights are 0.01, 0.1, 1, 0.01, and 0.01, respectively.

### 2.4. Post-processing Methods

**Logit Average.** We generate multiple results by a two-second-step sliding window and average the logit to smooth the result.

**Dual Threshold.** To expand the segment boundaries, we adopt a dual-threshold strategy commonly used in VAD to adjust the segment boundary. The low and high thresholds are set to 0.45 and 0.7.

### 2.5. Results

As shown in Table 1, the experimental results demonstrate the significance of the components and techniques used in the system on its performance. For threshold strategy, our final system uses a dual threshold in the post-processing module, while systems with other setups use a single threshold of 0.65. The evaluation metric is the diarization error rate (DER), which is the sum of false alarm (FA), miss detection (MS), and speaker confusion (SC).

The final system achieves a DER of 10.82%, which is in line with the leaderboard. Removing the post-processing module, which includes the logit average and dual threshold, decreases performance by 0.54%. The use of ECAPA-TDNN for speaker feature extraction is found to be more effective than using an i-vector, as a replacement results in a decrease in performance by 0.67%. The 0.82% decrease in performance demonstrates the importance of speaker context cues after removing the multi-speaker decoder.

Updating the i-vector in the TS-VAD and including the IPD feature is also shown to be essential for performance, as not updating the i-vector and using a fixed speaker embedding resulted in a decrease of 0.85%, and removing the IPD feature results in a decline of 1.09%. The speaker embedding module (i-vector) is also found to be necessary, as removing it resulted in a decrease in performance by 0.76%. The 'random delete' augmentation is shown to prevent over-fitting, as removing it decreased performance by 0.84%.

**Table 1**. The false alarm (FA), miss detection (MS), speaker confusion (SC) and diarization error rate (DER) in the evaluation set.

| Method | FA [%] | MS [%] | SC [%] | DER [%] |
|---|---|---|---|---|
| **Our System** | 4.44 | 4.28 | 2.10 | **10.82** |
| - Post-processing | 3.88 | 4.84 | 2.64 | 11.36 |
| - ECAPA-TDNN [6] | 3.96 | 5.02 | 3.05 | 12.03 |
| - Multi-speaker | 4.12 | 5.33 | 3.40 | 12.85 |
| - TS-VAD [3] | 6.50 | 4.20 | 3.01 | 13.70 |
| - IPD | 4.94 | 5.77 | 4.07 | 14.79 |
| - I-vector | 4.92 | 5.86 | 4.77 | 15.55 |
| - Random delete | 5.97 | 5.55 | 4.87 | 16.39 |

## 3. ACKNOWLEDGEMENTS

This work was supported by State Key Laboratory of Media Convergence Production Technology and Systems Project (No. SKLMCPTS2020003), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and National Natural Science Foundation of China (Grant No. 92048205).

## 4. REFERENCES

[1] H. Chen *et al.*, "The first multimodal information based speech processing (misp) challenge: Data, tasks, baselines and results," in *Proc. ICASSP 2022*.

[2] Z. Wang *et al.*, "The multimodal information based speech processing (misp) 2022 challenge: Audio-visual diarization and recognition," 2023. [Online]. Available: https://arxiv.org/abs/2303.06326

[3] I. Medennikov *et al.*, "Target-Speaker Voice Activity Detection: A Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario," in *Proc. Interspeech 2020*.

[4] M.-K. He *et al.*, "End-to-End Audio-Visual Neural Speaker Diarization," in *Proc. Interspeech 2022*.

[5] J. Roth *et al.*, "Ava active speaker: An audio-visual dataset for active speaker detection," in *Proc. ICASSP 2020*.

[6] B. Desplanques *et al.*, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech 2020*.