# THE WHU-ALIBABA AUDIO-VISUAL SPEAKER DIARIZATION SYSTEM FOR THE MISP 2022 CHALLENGE

*Ming Cheng*[1,3], *Haoxu Wang*[1,3], *Ziteng Wang*[2], *Qiang Fu*[2], *Ming Li*[1,3†]

[1]School of Computer Science, Wuhan University, Wuhan, China
[2]Alibaba Group, China
[3]Data Science Research Center, Duke Kunshan University, Kunshan, China

## ABSTRACT

This paper describes the system developed by the WHU-Alibaba team for the Multi-modal Information Based Speech Processing (MISP) 2022 Challenge. We extend the Sequence-to-Sequence Target-Speaker Voice Activity Detection framework to detect multiple speakers' voice activities from audio-visual signals simultaneously. The final system achieves a diarization error rate (DER) of 8.82% on the evaluation set of the competition database, which ranks 1st in the speaker diarization track of the MISP 2022, ICASSP Signal Processing Grand Challenge.

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

## 1. INTRODUCTION

Speaker diarization is the process of detecting speakers' voice activities in conversational data. Many classical methods are proposed, including clustering-based, end-to-end neural diarization, target-speaker voice activity detection methods and following modifications [1]. Also, audio-visual methods are explored to take advantage of different modalities (e.g., AVSD [2]).

This paper extends the Seq2Seq-TSVAD [3] framework to an audio-visual system, which can handle audio and visual lip-motion information to detect multiple speakers' voice activities. Our proposed system obtains a DER of 8.82% on the competition evaluation set to win first place in the speaker diarization track of the MISP 2022, ICASSP Signal Processing Grand Challenge.

## 2. SYSTEM DESCRIPTION

Fig. 1 depicts the designed Audio-Visual Seq2Seq-TSVAD framework. The audio front-end model, Conformer, and Speaker-wise Decoder (SW-Decoder) are the same as the original work [3]. The differences are described as follows.

We introduce the ResNet18-3D as the visual front-end model to process each speaker's lip video. Based on its standard implementation in Pytorch, there are three modifications: the convolutional kernel size, stride and output channels of the first stem layer are set to 7, 2, and 32 without max pooling; Output channels of the residual blocks are set to $\{32, 64, 128, 256\}$; All temporal downsampling operations (pooling/stride) in residual blocks are removed. By adding the last spatial global average pooling layer, this front-end extractor finally transforms a sequence of lip images into a sequence of feature embeddings.

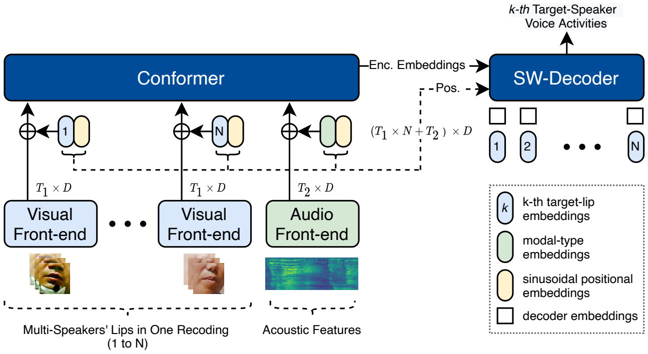

**Fig. 1**. The Framework of Audio-Visual Sequence-to-Sequence Target-Speaker Voice Activity Detection. For clarity, modality-dependent linear layers are omitted from the plot.

The inputs of TSVAD-based methods usually consist of acoustic features and speaker enrollments (e.g., x-vectors). The order of speaker enrollments determines the corresponding target-speaker voice activities. As there is no off-screen speaker in the competition database, we directly utilize lip videos as visual features as well as enrollments. A set of learnable embeddings represents the relative identities of input lips and related voice activities, namely target-lip embeddings. Meanwhile, learnable modality-type embeddings are initialized to differentiate encoded acoustic and visual features.

Target-lip embeddings, modality-type embeddings and sinusoidal positional embeddings are added to construct the new positional embeddings. After front-end models, two modality-dependent linear layers map the audio-visual features to the exact dimension of positional embeddings. Then, the Conformer module takes the sum of aligned feature sequences and positional embeddings as the inputs. On the decoder side, target-speaker embeddings in the original Seq2Seq-TSVAD model are replaced with the newly introduced target-lip embeddings. Decoder embeddings are set to zeros.

Let $N$ and $D$ denote the speaker number and feature dimension. $\mathbf{E_{lip}} = [\mathbf{e_1} \ldots \mathbf{e_N}]^{\mathrm{T}} \in \mathbb{R}^{\mathrm{N}\times\mathrm{D}}$ and $\mathbf{E_{mod}} \in \mathbb{R}^{1\times\mathrm{D}}$ denote the learnable target-lip and modality-type embeddings. Given a training sample, the outputs from audio and visual extractors can be denoted as $\mathbf{F^A} \in \mathbb{R}^{\mathrm{T_2}\times\mathrm{D}}$ and $\left\{\mathbf{F^V_n} \in \mathbb{R}^{\mathrm{T_1}\times\mathrm{D}} \mid 1 \leq n \leq N\right\}$, respectively. For each $\mathbf{F^V_n}$, the corresponding lip-embedding $\mathbf{e_n} \in \mathbb{R}^{1\times\mathrm{D}}$ is repeated to the length of $T_1$, then added with sinusoidal positional embeddings to become part of the final positional embeddings with the length of $T_1 \times N + T_2$. The rest dimension computations are the same as the original Seq2Seq-TSVAD.

† Corresponding Author, E-mail: ming.li369@dukekunshan.edu.cn

## 3. EXPERIMENTS

### 3.1. Data

The training data is from the MISP 2022 Challenge, including multi-channel audio and video data in near, middle, and far fields. We utilize the NARA-WPE [1] and SETK [2] toolkits for dereverberation and beamforming to extend available audio channels.

1) Basic Augmentation: Musan and RIRs corpora are applied as the audio augmentation. Furthermore, input videos undergo each item of the following procedures with a probability of 0.5: rotation with an angle range $[5, 20]$; horizontal flipping; cropping with the scale range $[0.8, 1]$; transformation of contrast, brightness, and saturation in the range $[-25, 25]$.

2) Extra Augmentation: When lip videos in a recording are not enough to fit the pre-set speaker number, each empty input has a probability of 0.5 to be padded by non-spoken lip videos randomly extracted from the training data. This way forces the model to distinguish valid and invalid inputs, namely Negative Sampling. In addition, we adopt the MixUp [4] method. Let $x$ and $y$ denote the lip video and voice activities that should be predicted, respectively. In each mixup, we randomly select the *i-th* and *j-th* speakers within a recording to construct the new training sample as follows:

$$\hat{x} = \lambda x_i + (1 - \lambda)\, x_j, \quad \hat{y} = \lambda y_i + (1 - \lambda)\, y_j, \tag{1}$$

where $\lambda \in [0, 1]$ is sampled from the $Beta\,(\alpha, \alpha)$ distribution. In practice, we set $\alpha = 0.2$ and perform the mixup between all within-recording speakers during model training.

### 3.2. Training

The model is set to the maximum speaker capacity of 6 and temporal resolution of 10 ms. The acoustic inputs take 80-dim log Mel-lterbank energies with a frame length of 25 ms and frameshift of 10 ms. The extraction of lip regions is the same as our previous work [5], and grayscale videos with a resolution of 88 × 88 and FPS of 25 are adopted as visual inputs. All training samples are split into 8-sec audio and video chunks with a stride of 4 seconds. Then, they are normalized with a mean of 0 and a standard deviation of 1.

The BCE loss and Adam optimizer are employed to train the neural network. First, we extract speakers' speaking and non-speaking audiovisual corpus from all available training data to perform online data simulation. The audio front-end module is frozen and initialized by its pre-trained speaker embedding model. With a learning rate of *1e-4* and a warmup of 2000 iterations, the whole model is trained by fully simulated data for 50 epochs until back-end convergence. Second, all model parameters are unfrozen to train around 50 epochs on the real far-field data without simulation. Finally, the model is fine-tuned around 10 epochs by decreasing the learning rate to *1e-5*.

### 3.3. Inference

We utilize far-field dereverberation audio and related videos as the test data. All test samples are split into 8-sec chunks with a 1-sec stride. Predicted results are stitched chunk by chunk. As a score-level fusion, multiple predictions from different audio channels and overlapped chunks are averaged at identical timestamps.

Lastly, we adopt the Oracle VAD provided by the competition to revise the diarization results as post-processing. The timestamps marked the VAD as active speech will directly assign a positive label to the speaker with the highest predicted score. Predictions at the timestamps marked as non-speech will be zeroed.

**Table 1**. DERs (%) of different systems on the MISP 2022 Database. The symbol + denotes the cumulative addition of the current method based on the preceding ones.

| System | Dev Set | Eval Set |
|---|---|---|
| Official Baseline [2] | 13.09 | 13.88 |
| Ours with Basic Augmentation | 9.07 | 11.01 |
| + Negative Sampling | 8.41 | 9.81 |
| + MixUp | 7.84 | 8.82 |

### 3.4. Evaluation

Table 1 illustrates the comparisons between our proposed system and the official baseline [2]. Ablation experiments show that the Negative Sampling and MixUp methods can effectively improve the primary system. The final system outperforms the original baseline significantly and obtains DERs of 7.84% and 8.82% on the development and evaluation sets of the MISP 2022 database, respectively.

## 4. CONCLUSIONS

This paper presents an Audio-Visual Seq2Seq-TSVAD framework for speaker diarization. By the ability of cross-speaker and cross-modal voice activity detection, our proposed method achieves the DER of 8.82% on the competition evaluation set, which ranks first place in the speaker diarization track of the MISP 2022 Challenge.

## 5. ACKNOWLEDGMENTS

This research is funded in part by the National Natural Science Foundation of China (62171207), Science and Technology Program of Guangzhou City (202007030011), Science and Technology Program of Suzhou City (SYC2022051), and Alibaba Air Fund. Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

## 6. REFERENCES

[1] Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J. Han, Shinji Watanabe, and Shrikanth Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Computer Speech & Language*, vol. 72, pp. 101317, 2022.

[2] Mao-Kui He, Jun Du, and Chin-Hui Lee, "End-to-End Audio-Visual Neural Speaker Diarization," in *Proc. Interspeech*, 2022, pp. 1461–1465.

[3] Ming Cheng, Weiqing Wang, Yucong Zhang, Xiaoyi Qin, and Ming Li, "Target-speaker voice activity detection via sequence-to-sequence prediction," *arXiv preprint arXiv:2210.16127*, 2022.

[4] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. ICLR*, 2018.

[5] Ming Cheng, Haoxu Wang, Yechen Wang, and Ming Li, "The dku audio-visual wake word spotting system for the 2021 misp challenge," in *Proc. ICASSP*, 2022, pp. 9256–9260.

[1] https://github.com/fgnt/nara_wpe

[2] https://github.com/funcwj/setk