Utilizing Self-Supervised Embeddings for Improving Audio-Visual Speaker Diarization at EGO4D Challenge 2023

Chin-Jou Li*, WenZhe Ren*, Chia-Wei Chen*
National Taiwan University, Taiwan
b09902035@csie.ntu.edu.tw

Ernie Chu*, Tzu-hsuan Huang*, Jen-Cheng Hou*, Jun-Cheng Chen, Yu Tsao
Academia Sinica, Taiwan
yu.tsao@citi.sinica.edu.tw

Abstract

This technical report demonstrates an approach for the audio-visual speaker diarization (AVD) task at EGO4D Challenge 2023. The approach is based on the baseline system of the challenge, where several building blocks are replaced with more advanced ones. The model is improved in three-fold. Firstly, we adopt a better face detection and tracking pipeline to improve visual front-end processing. Secondly, we use an AV-HuBERT based model for active speaker detection (ASD). It utilizes multi-modal self-supervised embeddings (SSE) and improves ASD at accuracy of frame-level prediction from 79% to 84%. Lastly, a pre-trained HuBERT model is used for generating audio embeddings to boost speaker matching, and yields a gain of +3% at diarization error rate. Our model is shown to perform better than the baseline on the validaton set, confirming the effectiveness of utilizing SSE in the AVD task.

1. Introduction

The EGO4D Challenge [1] is a contest dedicated for egocentric video analysis, such as audio-visual speaker diarization (AVD). The AVD task focuses on tackling the problem of 'who spoke when' in a given video. Inspired by the recent success of self-supervised learning (SSL) in speech processing [8], and audio-visual applications [10]. We wonder if self-superivsed embeddings (SSE) from the SSL models can benefit the AVD task as well. Specifically, we investigate SSE from the HuBERT [8] and AV-HuBERT [10] models, which have been shown to be helpful for downstream tasks such as speaker recognition [11], speech enhancement and separation [5].

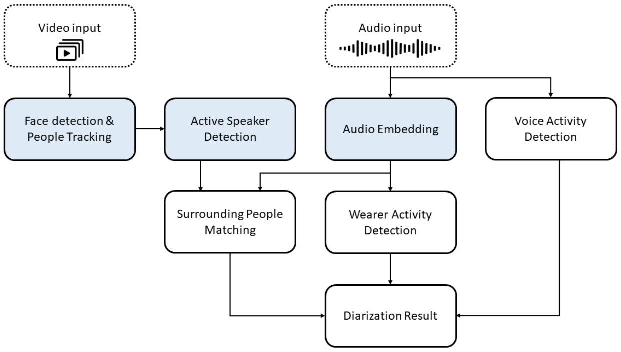


Figure 1. Overview of the model. The model follows the pipeline of the baseline system. The blocks in blue denote they are replaced with more advanced ones.

Our proposed approach, as shown in Fig. 1, is based on the baseline system. Some building blocks are replaced by more effective ones. For example, we use an improved pipeline for face detection and tracking to enhance the visual front-end processing. In addition, we adopt audio-visual SSE from AV-HuBERT for active speaker detection (ASD), and audio-only SSE from HuBERT for boosting the performance of speaker matching. Other models used in building blocks such as voice activity detection are kept the same. The following sections will elaborate our used methods and the experimental results.

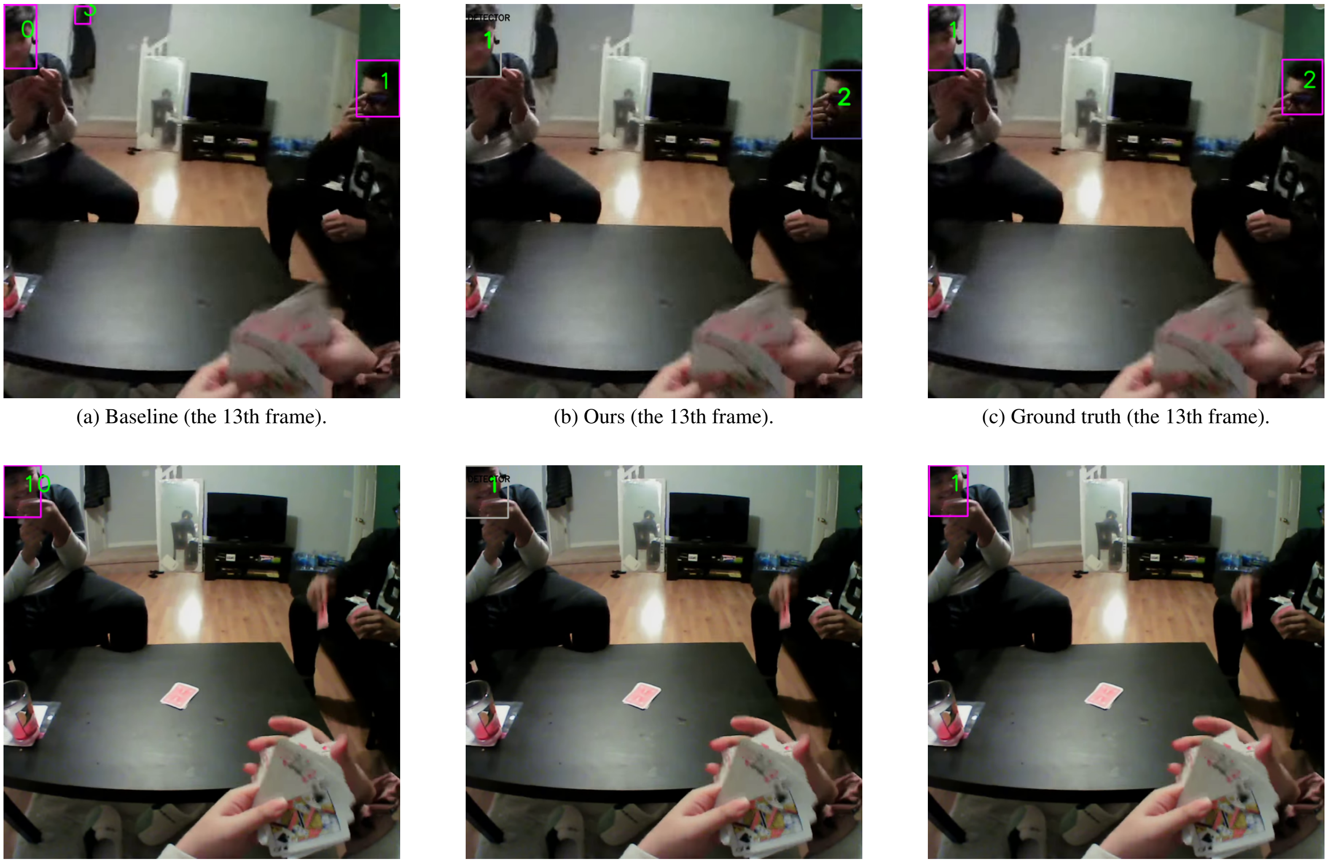
2. Methodology and implementation

2.1. Improved face detection and tracking

2.1.1 Short-term tacking

During the tracking procedure, we utilize a face detector [6] to identify and locate all the faces within a video clip. This

*These authors contributed equally to this work



(a) Baseline (the 13th frame).

(b) Ours (the 13th frame).

(c) Ground truth (the 13th frame).

(d) Baseline (the 140th frame).

(e) Ours (the 140th frame).

(f) Ground truth (the 140th frame).

Figure 2. This is the comparison between the baseline model, our method, and the ground truth. Sample frames are from the clip: $0fe736da$–$48f1$–4961–$bd7a$–$b75ce13c91b4$. Upon observation, it is evident that in the 13th frame, the baseline model detect an object that is not a human face. In the 140th frame, the baseline model incorrectly assigns the different IDs to the same people.

enables us to gather the bounding box coordinates and facial landmarks associated with each face. Subsequently, we generate the tracklet by sorting [4] these detected faces. Regardless of whether the faces are identical, each tracklet is assigned a new bounding box ID. Additionally, we store various information during the tracking process, such as frame ID, bounding box ID, bounding box coordinates and facial landmarks.

2.1.2 Long-term tacking

Long term tracking aims to merge the short-term tracklets associated with the same person. We begin by cropping all the faces within a clip, using the bounding box coordinates and facial landmarks obtained from the short-term tracking, and organizing them based on their bounding box IDs for storage. These cropped faces are then sent to a face recognition model [7] to extract their feature representations. Utilizing the cosine similarity metric, we calculate the similarity between each pair of faces based on their extracted feature. For example, Assuming we have 100 bounding boxes with different IDs. In this case, we would get a similarity map of size 100x100, capturing the pairwise similarities between all the faces. Through the application of a threshold, we determine whether faces with different box IDs correspond to the same individuals, if the similarity surpasses the threshold and there is no matching ID within the same frame, we proceed to replace the bounding box ID accordingly. The above method is implemented with an open-source tool [2] .

2.2. AV-HuBERT based ASD model

For audio-visual ASD, we adopt the model in [5], and make minimum modifications to fit the model for the ASD task. Hence our AV-HuBERT based ASD model consists of an AV-HuBERT module as an upstream processor, followed by a bidirectional long short-term memory (BLSTM) network and a classification head. The model specifications other than the binary classification layer follow the one in [5]. We fine-tune the AV-HuBERT from a pre-trained

Method	Accuracy
TalkNet [12]	0.79
AV-HuBERT based model (Ours)	0.84

Table 1. Fame-level accuracy comparison between the baseline ASD model (TalkNet) and our AV-HuBERT based model on the validation set of the EGO4D dataset.

Method	DER
Baseline	0.86
HuBERT + PCA (Ours)	0.83

Table 2. DER comparison of our audio embedding extraction method with the baseline model on the validation set of the EGO4D dataset.

Method	DER
Baseline	0.80
Ours	0.79

Table 3. DER comparison of our full model with the baseline model on the validation set of the EGO4D dataset.

checkpoint on LRS3 [3]. The learning objective is to minimize the cross-entropy loss.

2.3. Audio embeddings with the HuBERT model

To produce effective audio embeddings for speaker matching, we use a HuBERT model pre-traind on Librispeech [9] to generate SSE at a dimension of 768. Principal component analysis (PCA) is then utilized to reduce the dimensionality to 150. The HuBERT model is not fine-tuned or optimized on the EGO4D dataset. Zero-shot inference is performed.

3. Experimental results

Table 1 and Table 2 show ablation studies of how our ASD and audio embedding extraction blocks outperform the baselines. In Table 1, we can observe that at frame-level prediction for ASD, our method can obtain a gain of +5% on the validation set of the EGO4D dataset. In Table 2, it shows our audio embedding extraction approach can outperform the baselines in terms of diarization error rate (DER) in the validation set. These results confirm the effectiveness of utilizing the SSE. Lastly, by integrating all the three advanced blocks, as shown in Table 3, our model can outperform the baseline model in DER on the validation set.

4. Conclusion

In this report, we propose three advanced building blocks to improve the baseline model of the AVD task at the EGO4D Challenge 2023. They include an improved face tracking pipeline, AV-HuBERT based ASD model, and HuBERT based audio embedding extraction. In particular, our visual front-end processing can outperform the baseline, providing more robust results of face detection and tracking. In addition, we show that by introducing SSE from pre-trained SSL models, both ASD and quality of audio embeddings can have a significant improvement. Lastly, by integrating all the proposed blocks, our method can outperform the baseline model on the validation set, confirming the effectiveness of our model design choices.

References

[1] EGO4D challenge 2023. https://ego4d-data.org/docs/challenge/. 1

[2] facexlib. https://github.com/xinntao/facexlib. 2

[3] T. Afouras, J.-S. Chung, and A. Zisserman. LRS3-TED: a large-scale dataset for visual speech recognition. In *arXiv preprint arXiv:1809.00496*, 2018. 3

[4] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, 2016. 2

[5] I-Chun Chern, Kuo-Hsuan Hung, Yi-Ting Chen, Tassadaq Hussain, Mandar Gogate, Amir Hussain, Yu Tsao, and Jen-Cheng Hou. Audio-visual speech enhancement and separation by leveraging multi-modal self-supervised embeddings. *arXiv preprint arXiv:2210.17456*. 1, 2

[6] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-shot multi-level face localisation in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020. 1

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[8] W.-N. Hsu et al. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. 1

[9] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Apr. 2015. 3

[10] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*. 1

[11] Bowen Shi, Abdelrahman Mohamed, and Wei-Ning Hsu. Learning lip-based audio-visual speaker embeddings with AV-HuBERT. In *Interspeech 2022*, 2022. 1

[12] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 3