HLT SYSTEM FOR FAME CHALLENGE 2024

Ruijie Tao[1], *Shi Zhan*[2], *Yidi Jiang*[1], *Duc-Tuan Truong*[3], *Eng-Siong Chng*[3], *Massimo Alioto*[1] *and Haizhou Li*[1,2,4]

[1]National University of Singapore, Singapore [2]The Chinese University of Hong Kong, Shenzhen, China
[3]Nanyang Technological University, Singapore [4] Shenzhen Research Institute of Big Data, Shenzhen, China

ABSTRACT

The 2024 Face-voice Association in Multilingual Environments (FAME) Challenge focuses on learning face-voice associations. The general relationship between the face and voice is deeply explored in this challenge. Meanwhile, language effect, especially the cross-language evaluation, is studied in this process. Our team explored the importance of data quality, framework design and training strategy to achieve a robust cross-modal system. The experimental results achieved the first rank in 2024 FAME with the overall EER that equals 19.9 %.

Index Terms— cross-modal speaker recognition, multilingual

1. INTRODUCTION

Supervised speaker recognition aims to verify the identity of a speaker based on their voice characteristics. Deep learning-based speaker representation systems, with the help of large-scale datasets, have become mainstream methods [1–3].

Recent research has found that this verification pipeline can also be extended to the cross-modal condition [4]. Specifically, for one static face and one heard voice, the human brain can detect whether they come from the same person. This judgement is based on the general cross-modal relationship, such as gender, nationality and age. Meanwhile, facial appearance can also be linked to the characters' voices. That can be used as the instruction for downstream speech tasks, such as cross-modal speech separation.

In the FAME Challenge 2024 [5], the main focus is to explore and enhance the integration of face and voice embeddings for robust speaker recognition in multilingual contexts. The challenge extends beyond traditional methods, which often rely solely on cross-validation within a single language, by examining the effects of multiple languages on the verification process.

In this report, we study the cross-modal speaker recognition system from three main aspects: data preparation and quality, audio and visual pre-training strategy, and framework architecture. Our system achieves 19.9% EER for the final evaluation, which gets the first rank in the challenge.

2. PROPOSED FRAMEWORK

In this section, we describe the proposed audio-visual learning method for multilingual speaker recognition. We attempted many aspects, while most of them cannot boost the system, so in this report, we only highlight the successful parts of our system.

2.1. Keynote speaker frontend

Ideally, each speech file should contain only one talking person. However, some speech files in the given dataset contain conversations from two or more speakers. That can damage the system for training and predicting. In other words, filtering the dataset and guaranteeing only one speaker in each speech file is important.

We attempted some speaker diarization methods but could not help due to the imbalanced speaker distribution. Here, we propose an assumption: the person who talks the most is the interested person, whom we call the 'keynote speaker'. Motivated by our previous target speech diarization work [6], we involve such a keynote speaker frontend for data filtering.

The keynote speaker frontend is a transformer-based structure; the input is the long conversation, and the output is the combination of all the speeches from the keynote speaker (who talks the most). We find this frontend can significantly boost the dataset and remove the incorrect speech segments or non-speech segments.

2.2. Framework design

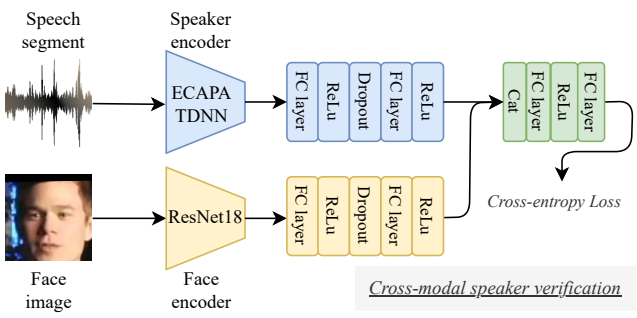

Fig. 1: Our framework for cross-modal speaker verification task in FAME challenge.

As shown in Figure 1, our framework contains the speaker encoder, face encoder and the fusion block. The fusion block begins by altering the dimension of both speaker and face embeddings to the same value, incorporating ReLU activation and the dropout rate of 0.5 to avoid over-fitting. These embeddings are then concatenated. Following this, the fully connected layers process the combined data, resulting in a unified embedding with a dimension of 256. The training loss is the simple binary cross-entropy. We also tried or combined some other loss functions such as contrasting loss, supervised contrasting learning [7] and symmetric loss [8]. But they cannot perform well.

2.3. Training strategy

Cross-modal speaker recognition aims to learn the general relationship between the human voice and face, so it is important to understand the facial appearance and speech representation. Based on that, we propose a three-stage training strategy for boosting performance by leveraging the power of pre-training. Note that we have strictly met the unseen-language requirements in these experiments.

2.3.1. Single-modal pre-train

For speech modality, we employ the ECAPA-TDNN model [9] as the speaker encoder to learn the voice character. This is a popular, efficient model for learning speaker embedding through classification learning. For the 'seen' evaluation condition, this speaker model is pre-trained on the entire VoxCeleb2 dataset [10]. For the 'unseen' evaluation condition, it is pre-trained on VoxCeleb2 without the corresponding language (English or Urdu) to meet the requirements.

For face modality, we use the ResNet-18 model as the face encoder trained on the Glink360K database [11]. Benefit from this large-scale face recognition dataset, this face encoder can extract distinctive facial features.

2.3.2. Multi-modal pre-train

After we obtain the robust speaker and face encoder, we perform the second-stage pre-training on the VoxCeleb1 [12] dataset plus the given FAME dataset. For the seen condition, the entire VoxCeleb1 and FAME datasets are used. We skip this step for 'unseen English' condition since non-English videos are very limited in VoxCeleb1. In 'unseen Urdu' condition, we use VoxCeleb1-without-Urdu plus FAME-without-Urdu for pre-training.

2.3.3. FAME dataset fine-tune

Finally, we fine-tune the pre-trained system on FAME dataset. This step shares the same pipeline with multi-modal pre-train, it uses FAME, FAME-no-English or FAME-no-Urdu for corresponding seen-or-unseen requirements.

3. EXPERIMENTAL SETUP

The 2024 FAME Challenge provides the MAV-Celeb corpus, which includes 154 celebrities speaking three languages: English, Hindi, and Urdu. For each video in the dataset, we extract the complete audio to capture the speaker's voice and apply face detection to identify prominent faces corresponding to the speakers. To construct our data samples, we pair a voice and a face from the same individual to create positive samples, while negative samples pair elements from different speakers. To monitor the performance, we split 6 speakers from the training set to generate the validation dataset. We notice that the model is very easy to over-fit, so the number of training epochs is very small in all experiments. We found that a low learning rate can boost the system, so we set it as 1e-4. Training and testing share a similar pipeline to predict the final binary classification label. System performance is reported in terms of equal error rate (EER).

4. RESULTS

Our best results can be found in Table 1. The lowest EER has achieved 14.7% for Urdu, which proves the huge potential of the cross-modal speaker verification technique. Meanwhile, the heard results perform better than the unheard condition. That proves the importance of language in this task. Also, we found that the results for Urdu were much better than those for English.

Table 1: Evaluation results (EER) on FAME dataset, V1.

	Eng. test	Urdu. test	Overall
Eng. train	21.8	15.8	19.9
Urdu train	27.3	14.7	

5. SUMMARY OF FAILURE METHODS

We have attempted many methods in this challenge, however most of them cannot work in our experiments or need further verification:

- Data clean: VAD; clustering-based diarization; evaluation with split segments
- Data augmentation: audio augmentation; face augmentation; face alignment
- Loss: contrastive learning, supervised contrastive learning, threshold loss
- Architecture: BLSTM fusion, heavy fusion layer
- Pre-training: large speech model (WavLM, ResNet152, CAM++); large face model (ResNet100)
- Method: gender detection, score fusion

6. REFERENCES

[1] Duc-Tuan Truong, Ruijie Tao, Jia Qi Yip, Kong Aik Lee, and Eng Siong Chng, "Emphasized non-target speaker knowledge in knowledge distillation for automatic speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10336–10340.

[2] Duc-Tuan Truong, Tran The Anh, and Chng Eng Siong, "Exploring speaker age estimation on different self-supervised learning models," in *IEEE APSIPA ASC*, 2022, pp. 1950–1955.

[3] Tianchi Liu, Lin Zhang, Rohan Kumar Das, Yi Ma, Ruijie Tao, and Haizhou Li, "How do neural spoofing countermeasures detect partially spoofed audio?," *arXiv preprint arXiv:2406.02483*, 2024.

[4] Ruijie Tao, Rohan Kumar Das, and Haizhou Li, "Audio-visual speaker recognition with a cross-modal discriminative network," in *Interspeech 2020*, 2020.

[5] Muhammad Saad Saeed, Shah Nawaz, Muhammad Salman Tahir, Rohan Kumar Das, Muhammad Zaigham Zaheer, Marta Moscati, Markus Schedl, Muhammad Haris Khan, Karthik Nandakumar, and Muhammad Haroon Yousaf, "Face-voice association in multilingual environments (fame) challenge 2024 evaluation plan," *arXiv preprint arXiv:2404.09342*, 2024.

[6] Yidi Jiang, Zhengyang Chen, Ruijie Tao, Liqun Deng, Yanmin Qian, and Haizhou Li, "Prompt-driven target speech diarization," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11086–11090.

[7] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan, "Supervised contrastive learning," *arXiv preprint arXiv:2004.11362*, 2020.

[8] Nontawat Charoenphakdee, Jongyeong Lee, and Masashi Sugiyama, "On symmetric losses for learning from corrupted labels," in *International Conference on Machine Learning*. PMLR, 2019, pp. 961–970.

[9] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech*, 2020, pp. 3830–3834.

[10] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep speaker recognition," in *Interspeech*, 2018, pp. 1086–1090.

[11] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou, "Retinaface: Single-stage dense face localisation in the wild," *arXiv preprint arXiv:1905.00641*, 2019.

[12] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Interspeech*, 2017, pp. 2616–2620.