A Synopsis of FAME 2024 Challenge: Associating Faces with Voices in Multilingual Environments

Muhammad Saad Saeed*
University of Engineering and Technology
Taxila, Pakistan
saad.saeed@uettaxila.edu.pk

Shah Nawaz*†
Institute of Computational Perception, Johannes Kepler University
Linz, Austria
shah.nawaz@jku.at

Marta Moscati*
Institute of Computational Perception, Johannes Kepler University
Linz, Austria
marta.moscati@jku.at

Rohan Kumar Das*
Fortemedia
Singapore, Singapore
rohankd@fortemedia.com

Muhammad Salman Tahir
University of Engineering and Technology
Taxila, Pakistan
18-cp-58@students.uettaxila.edu.pk

Muhammad Zaigham Zaheer
Mohamed bin Zayed University of Artificial Intelligence
Abu Dhabi, United Arab Emirates
zaigham.zaheer@mbzuai.ac.ae

Muhammad Irzam Liaqat
IMT School for Advanced Studies of Lucca
Lucca, Italy
irzam.liaqat@imtlucca.it

Muhammad Haris Khan
Mohamed bin Zayed University of Artificial Intelligence
Abu Dhabi, United Arab Emirates
muhammad.haris@mbzuai.ac.ae

Karthik Nandakumar
Mohamed bin Zayed University of Artificial Intelligence
Abu Dhabi, United Arab Emirates
karthik.nandakumar@mbzuai.ac.ae

Muhammad Haroon Yousaf
University of Engineering and Technology
Taxila, Pakistan
haroon.yousaf@uettaxila.edu.pk

Markus Schedl
Institute of Computational Perception, Johannes Kepler University
Human-centered AI Group, AI Lab, Linz Institute of Technology, Austria
Linz, Austria
markus.schedl@jku.at

Abstract

Over half of the world's population is bilingual and people often communicate under multilingual scenarios. The Face-Voice Association in Multilingual Environments (FAME) 2024 Challenge, held at ACM Multimedia 2024, focuses on establishing face-voice association to analyze the impact of multiple languages on the verification process. This report provides a brief summary of the challenge.

CCS Concepts

• **Computing methodologies** → **Biometrics**; **Machine learning**.

Keywords

Multimodal learning, Face-voice association

ACM Reference Format:
Muhammad Saad Saeed*, Shah Nawaz*†, Marta Moscati*, Rohan Kumar Das*, Muhammad Salman Tahir, Muhammad Zaigham Zaheer, Muhammad Irzam Liaqat, Muhammad Haris Khan, Karthik Nandakumar, Muhammad Haroon Yousaf, and Markus Schedl. 2024. A Synopsis of FAME 2024 Challenge: Associating Faces with Voices in Multilingual Environments. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3664647.3688978

*Equal contribution.
†Corresponding author.

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3688978

1 Introduction

Humans can associate the voices and faces of people because the neuro-cognitive pathways for both modalities share the same structure [4]. Nagrani et al. have leveraged deep learning methods to establish an association between voices and faces for cross-modal verification and matching tasks [6, 7]. Since then, the task has received notable research interest [1, 2, 8, 12]. As over half of the world population is bilingual [5], it is important to investigate the impact of language on face-voice (F-V) association. The Face-Voice Association in Multilingual Environments (FAME) Challenge 2024, as seen in Figure 1, aims to study this aspect through a cross-modal verification task.

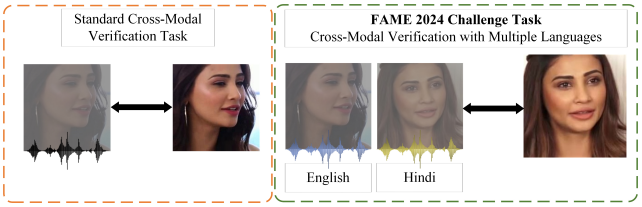

Figure 1: (Left) F-V association is established with a cross-modal verification task [6]. (Right) The FAME 2024 Challenge extends the task to analyze the impact of multiple languages.

2 The Challenge

The task of the challenge is cross-modal verification, where the goal is to verify whether the audio segment from multiple languages and a face image belong to the same identity. For this challenge, an "unseen" test set was curated specifically to evaluate solutions using these samples of MAV-Celeb [9], ensuring that the solutions proposed by the participants are evaluated on the task while mitigating the risk of exploiting biases. The challenge underwent two phases: the first phase using the development set, which lasted from April 15, 2024, to June 14, 2024, with 10 submissions allowed per day for each team with a maximum number of 100 submissions during that entire phase; and the second phase using the unseen evaluation set, lasting from June 15, 2024, to June 22, 2024, with only five submissions allowed in total.

Baseline Method & Starter Kit: The baseline approach named FOP consists of a two-branch network that takes as input the embeddings of face and voice. The embeddings for the first branch are obtained using a popular convolutional neural network pre-trained on a large-scale facial recognition dataset [10]. The embeddings of the other branch are extracted using an audio encoding network [14] trained using the *heard* language. The network utilizes complementary cues available in the embeddings of both modalities to form enriched fused embeddings and imposes orthogonal constraints on them for learning discriminative joint representation. More information is available in the prior work on the baseline [11] and the repository: https://github.com/mavceleb/mavceleb_baseline.

Evaluation Plan: Comprehensive details on data, baseline, metrics, submission portal, and rules are provided in [13].

3 Winning solutions

The results of the top 3 teams and their comparison to the challenge baseline are reported in Table 1.

HLT uses a data filtering approach to find the keynote speaker in the audio segment [3]. Afterward, it leverages a two-branch network consisting of modality-specific encoders and a fusion module to learn joint representation. It achieved the first rank in the FAME 2024 Challenge with an impressive overall EER of 19.9%.

Audio-visual leverages a pre-trained baseline (FOP_{frozen}) and the FOP_{update} in a dual-branch structure. Specifically, it freezes the parameters of the FOP_{frozen} during training, while the FOP_{update} is trained from scratch. It allows the FOP_{frozen} model to serve as a fixed embedding extractor, while the FOP_{update} dynamically complements the output results of the FOP_{frozen}. FOP_{update} is trained regularly using augmented face-voice pairs to enhance its generalization capability and robustness. It achieved the second rank in FAME 2024 Challenge with an excellent overall EER of 20.5%.

Table 1: Comparison of baseline method (FOP) and top 3 winning teams on FAME 2024 Challenge evaluation set.

		V1-EU		
Method	Configuration	English test (EER ↓)	Urdu test (EER ↓)	Overall score (EER ↓)
FOP [11]	English train	29.3	37.9	33.4
	Urdu train	40.4	25.8	
Xaiofei	English train	28.5	33.2	21.8
	Urdu train	28.6	20.9	
Audio-visual	English train	17.1	28.2	20.5
	Urdu train	18.3	18.4	
HLT	English train	21.8	27.3	**19.9**
	Urdu train	15.8	14.7	

Xaiofei leverages supervised cross-contrastive learning to establish associations between faces and voices. Subsequently, the joint representations are employed in a post-processing phase that incorporates the chaining-cluster re-score technique to address outliers prevalent in data. It achieved the third rank in the FAME 2024 Challenge with a notable overall EER of 27.8%.

References

[1] Guangyu Chen, Deyuan Zhang, Tao Liu, and Xiaoyong Du. 2023. Local-Global Contrast for Learning Voice-Face Representations. In *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 51–55.

[2] Shota Horiguchi, Naoyuki Kanda, and Kenji Nagamatsu. 2018. Face-voice matching using cross-modal embeddings. In *Proceedings of the 26th ACM international conference on Multimedia*. 1011–1019.

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

[4] Miyuki Kamachi, Harold Hill, Karen Lander, and Eric Vatikiotis-Bateson. 2003. Putting the face to the voice': Matching identity across modality. *Current Biology* 13, 19 (2003), 1709–1714.

[5] Jay Mathews. 2019. Half of the world is bilingual. What's our problem? www.washingtonpost.com/local/education/half-the-world-is-bilingual-whats-our-problem/2019/04/24/1c2b0cc2-6625-11e9-a1b6-b29b90efa879_story. [Online; accessed 10-June-2024].

[6] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. 2018. Learnable pins: Cross-modal embeddings for person identity. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 71–88.

[7] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. 2018. Seeing voices and hearing faces: Cross-modal biometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8427–8436.

[8] Shah Nawaz, Muhammad Kamran Janjua, Ignazio Gallo, Arif Mahmood, and Alessandro Calefati. 2019. Deep Latent Space Learning for Cross-modal Mapping of Audio and Visual Signals. In *2019 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 1–7.

[9] Shah Nawaz, Muhammad Saad Saeed, Pietro Morerio, Arif Mahmood, Ignazio Gallo, Muhammad Haroon Yousaf, and Alessio Del Bue. 2021. Cross-modal Speaker Verification and Recognition: A Multilingual Perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1682–1691.

[10] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep face recognition. (2015).

[11] Muhammad Saad Saeed, Muhammad Haris Khan, Shah Nawaz, Muhammad Haroon Yousaf, and Alessio Del Bue. 2022. Fusion and Orthogonal Projection for Improved Face-Voice Association. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7057–7061.

[12] Muhammad Saad Saeed, Shah Nawaz, Muhammad Haris Khan, Muhammad Zaigham Zaheer, Karthik Nandakumar, Muhammad Haroon Yousaf, and Arif Mahmood. 2023. Single-branch network for multimodal training. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[13] 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. 2024. Face-voice Association in Multilingual Environments (FAME) Challenge 2024 Evaluation Plan. *arXiv preprint arXiv:2404.09342* (2024).

[14] Weidi Xie, Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2019. Utterance-level aggregation for speaker recognition in the wild. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5791–5795.