

LINKING FACES AND VOICES ACROSS LANGUAGES: INSIGHTS FROM THE FAME 2026 CHALLENGE

Marta Moscatt^{1†}, Ahmed Abdullah^{2†}, Muhammad Saad Saeed^{3†}, Shah Nawaz^{1†}, Rohan Kumar Das^{4†}, Muhammad Zaigham Zaheer⁵, Junaid Mir⁶, Muhammad Haroon Yousaf⁶, Khalid Mahmood Malik³, Markus Schedl^{1,7}

¹Johannes Kepler University Linz, Austria, ²National University of Computer and Emerging Sciences, Pakistan

³University of Michigan, USA, ⁴Fortemedia Singapore, Singapore

⁵Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates

⁶University of Engineering and Technology Taxila, Pakistan,

⁷Human-centered AI Group, AI Lab, Linz Institute of Technology, Austria

mavceleb@gmail.com

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) 2026 Challenge, held at ICASSP 2026, focuses on developing methods for face-voice association that are effective when the language at test-time is different than the training one. This report provides a brief summary of the challenge.

Index Terms— Multimodal learning, Face-voice association, Cross-modal verification, Cross-modal matching

1. INTRODUCTION

Humans can associate the voices and faces of people because the neuro-cognitive pathways for both modalities share the same structure [1]. Inspired by this insight, Nagrani et al. have leveraged deep learning methods to establish an association between voices and faces for cross-modal speaker verification and matching tasks [2, 3]. Since then, the task of correctly associating the face and voice of a speaker has received notable research interest [4–10]. As over half of the world population is bilingual, it is important to investigate the impact of language on face-voice association. As displayed in Fig. 1, the Face-Voice Association in Multilingual Environments (FAME) 2026 Challenge aims to promote the development of cross-modal verification algorithms that are effective for multilingual speakers.

2. THE GRAND CHALLENGE: FAME 2026

Overview. The task of the grand challenge is cross-modal speaker verification, where the goal is to verify whether the

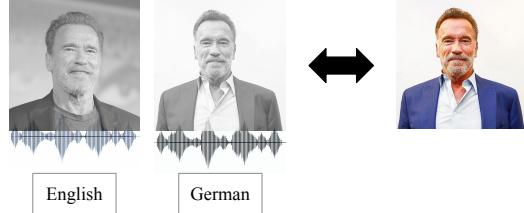


Fig. 1: The FAME 2026 Challenge overview: cross-modal verification to analyze the impact of multiple languages.

audio segment and a face image belong to the same speaker. In our challenge each speaker is represented while speaking more than one language. In particular, using the samples of MAV-Celeb [11, 12], we curated *unheard* test sets, in which the language spoken by the speakers was never represented in the training data. This allowed us to evaluate whether the proposed solutions were subject to a performance deterioration under a change of language.

Baseline Method & Starter Kit. The baseline approach is named FOP and 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 convolutional neural network pre-trained on a large-scale facial recognition dataset [13]. The embeddings of the other branch are extracted using an audio encoding network [14] trained using the *heard* language. The network combines the complementary information from both modalities in a fused embedding and imposes orthogonal constraints on the two modalities to learn joint representations that represent distinct speakers in different ways. More information is available in the prior work on the baseline [15] and the repository: https://github.com/mavceleb/mavceleb_baseline.

Dataset. Following the first edition of the FAME Grand Challenge hosted at ACM Multimedia 2024 [12, 16], we updated

[†]Equal Contribution.

the dataset including German as additional language, as well as bilingual speakers. We curated a new dataset split consisting of 58 English-German bilingual speakers. The split provides language annotations, which allow to analyze the impact of multiple languages on face-voice association. Analogously to the previous splits, the samples are obtained from YouTube videos and consist of celebrities appearing in interviews, talk shows, and television debates [11]. The visual data spans a vast range of setups, including different poses, motion blurs, background clutters, video qualities, occlusions, and lighting conditions. Moreover, since the videos originate from real-world situations, they reproduce the same challenges that are encountered when deploying face-voice association tools in real-world scenarios, such as noise, background chatter or music, overlapping voices, and compression artifacts. These aspects render the dataset both challenging for existing algorithms, and useful for developing algorithms that can have an impact on real applications.

Evaluation Plan. Comprehensive details on baseline, metric, submission portal, and rules are provided in [17].

Challenge Timeline and Results. The challenge consisted of two phases. In the first phase (August 15, 2025, to November 20, 2025) participants’ approaches were evaluated on the development set; each team was allowed to submit 10 submissions per day, with a maximum number of 100 submissions during the entire phase. In the second phase (November 05, 2025 to November 20, 2025) participants’ approaches were evaluated on the unseen evaluation set; each team was allowed for a maximum of 10 submissions over the whole phase. Fig. 2a shows the daily submission activity across all teams during the Progress and Evaluation phases, indicating distinct spikes that correspond to periods of increased engagement. Fig. 2b further breaks down these contributions by team, highlighting variations in participation intensity across the two phases. Finally, Fig. 2c shows the team-wise performance during the Evaluation phase, expressed as mean Equal Error Rate (EER) with standard deviation error bars. Finally, Table 1 presents the EER for the top-5 teams in comparison to the baseline method [15]. The participating teams demonstrated a notable improvement over the baseline method, with the leading team, Simicch [18], achieving an EER of 23.99, a considerable reduction from the baseline EER of 41.57. This substantial performance gap underscores the success of the

Rank	Team	EER (%)
-	Baseline [15]	41.57
1	Simicch [18]	23.99
2	Areffarhadi [19]	24.73
3	Alpha_code [20]	33.11
4	LTINI [21]	33.18
5	Punkmale [22]	33.51

Table 1: Performance comparison of top 5 teams with baseline in EER (%).

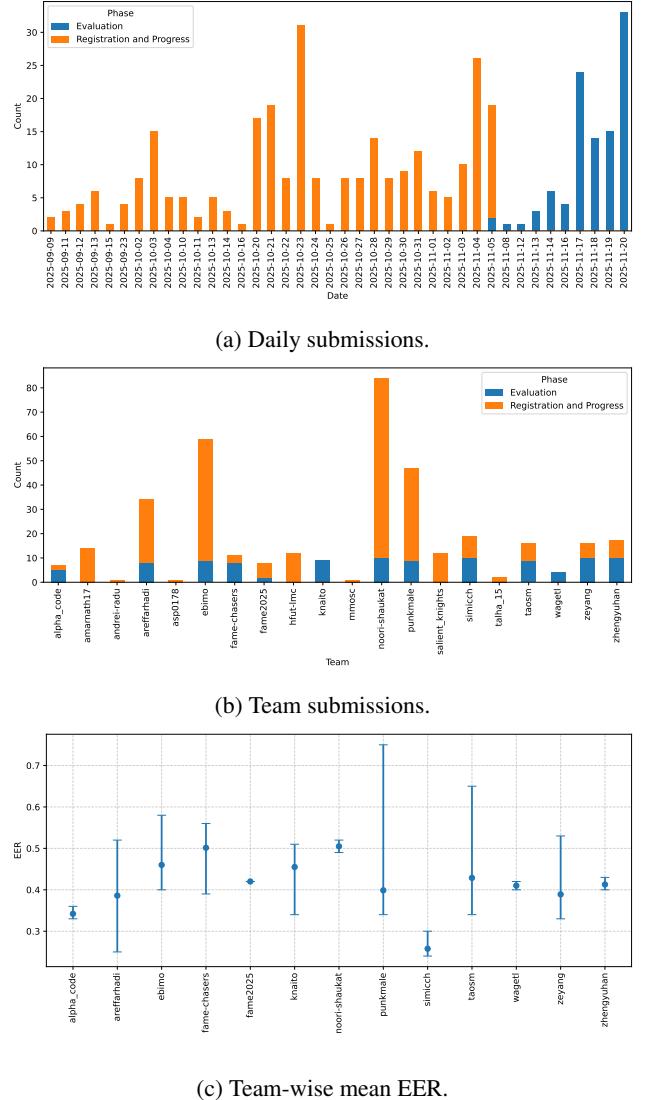


Fig. 2: (a) Overview of participant activity across grand challenge phases, showing the temporal distribution of submissions. (b) Aggregation of team submissions, emphasizing differences in total participation and the phase allocation between Registration/Progress and Evaluation phase. (c) Evaluation phase performance is summarized by team-wise mean EER, with error bars indicating standard deviation.

challenge, which led to the development of effective methodologies that advanced the state-of-the-art for face-voice association in multilingual environments.

3. ACKNOWLEDGMENTS

This research was funded in whole or in part by the Austrian Science Fund (FWF): <https://doi.org/10.55776/COE12> as part of the Cluster of Excellence Bilateral Artificial Intelligence.

4. REFERENCES

- [1] Miyuki Kamachi, Harold Hill, Karen Lander, and Eric Vatikiotis-Bateson, “Putting the face to the voice”: Matching identity across modality,” *Current Biology*, vol. 13, no. 19, 2003.
- [2] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman, “Seeing voices and hearing faces: Cross-modal biometric matching,” in *CVPR*, 2018.
- [3] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman, “Learnable pins: Cross-modal embeddings for person identity,” in *ECCV*, 2018.
- [4] Shota Horiguchi, Naoyuki Kanda, and Kenji Nagamatsu, “Face-voice matching using cross-modal embeddings,” in *ACM Multimedia*, 2018.
- [5] Shah Nawaz, Muhammad Kamran Janjua, Ignazio Gallo, Arif Mahmood, and Alessandro Calefati, “Deep latent space learning for cross-modal mapping of audio and visual signals,” in *DICTA*, 2019.
- [6] Ruijie Tao, Rohan Kumar Das, and Haizhou Li, “Audio-Visual Speaker Recognition with a Cross-Modal Discriminative Network,” in *Interspeech*, 2020.
- [7] Guangyu Chen, Deyuan Zhang, Tao Liu, and Xiaoyong Du, “Local-global contrast for learning voice-face representations,” in *ICIP*, 2023.
- [8] Muhammad Saad Saeed, Shah Nawaz, Muhammad Haris Khan, Muhammad Zaigham Zaheer, Karthik Nandakumar, Muhammad Haroon Yousaf, and Arif Mahmood, “Single-branch network for multimodal training,” in *ICASSP*, 2023.
- [9] Saqlain Hussain Shah, Muhammad Saad Saeed, Shah Nawaz, and Muhammad Haroon Yousaf, “Speaker recognition in realistic scenario using multimodal data,” in *ICAI*, 2023.
- [10] Abdul Hannan, Muhammad Arslan Manzoor, Shah Nawaz, Muhammad Irzam Liaqat, Markus Schedl, and Mubashir Noman, “PAEFF: Precise alignment and enhanced gated feature fusion for face-voice association,” in *Interspeech*, 2025.
- [11] Shah Nawaz, Muhammad Saad Saeed, Pietro Moretto, Arif Mahmood, Ignazio Gallo, Muhammad Haroon Yousaf, and Alessio Del Bue, “Cross-modal speaker verification and recognition: A multilingual perspective,” in *CVPR Workshops*, 2021.
- [12] Muhammad Saad Saeed, Shah Nawaz, Marta Moscati, Rohan Kumar Das, Muhammad Salman Tahir, Muhammad Zaigham Zaheer, Muhammad Irzam Liaqat, Muhammad Haroon Yousaf, Shah Nawaz, Muhammad Saad Saeed, Pietro Moretto, Arif Mahmood, Ignazio Gallo, Muhammad Haroon Yousaf, and Alessio Del Bue, “Cross-modal speaker verification and recognition: A multilingual perspective,” in *CVPR Workshops*, 2021.
- [13] Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman, “Deep face recognition,” in *BMVC*, 2015.
- [14] Weidi Xie, Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, “Utterance-level aggregation for speaker recognition in the wild,” in *ICASSP*, 2019.
- [15] Muhammad Saad Saeed, Muhammad Haris Khan, Shah Nawaz, Muhammad Haroon Yousaf, and Alessio Del Bue, “Fusion and orthogonal projection for improved face-voice association,” in *ICASSP*, 2022.
- [16] 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.
- [17] Marta Moscati, Ahmed Abdullah, Muhammad Saad Saeed, Shah Nawaz, Rohan Kumar Das, Muhammad Zaigham Zaheer, Junaid Mir, Muhammad Haroon Yousaf, Khalid Malik, and Markus Schedl, “Face-voice association in multilingual environments (FAME) 2026 challenge evaluation plan,” *arXiv preprint arXiv:2508.04592*, 2025.
- [18] Christopher Simic, Korbinian Riedhammer, and Tobias Bocklet, “Shared multi-modal embedding space for face-voice association,” *arXiv preprint arXiv:2512.04814*, 2025.
- [19] Aref Farhadipour, Teodora Vukovic, and Volker Dellwo, “Towards language-independent face-voice association with multimodal foundation models,” *arXiv preprint arXiv:2512.02759*, 2025.
- [20] Abdul Hannan, Furqan Malik, Hina Jabbar, Syed Suleiman Sadiq, and Mubashir Noman, “RFOP: Rethinking fusion and orthogonal projection for face-voice association,” *arXiv preprint arXiv:2512.02860*, 2025.
- [21] Zeyang Zhang, Katsuhiko Naito, and Hajer Dahmani, “Contrastive gated fusion for multilingual speaker verification,” Dec. 2025.
- [22] Zhihua Fang, Shumei Tao, Junxu Wang, and Liang He, “XM-ALIGN: Unified cross-modal embedding alignment for face-voice association,” *arXiv preprint arXiv:2512.06757*, 2025.