

# Face-voice Association in Multilingual Environments (FAME)

## 2026 Challenge Evaluation Plan

Marta Moscati<sup>1†</sup>, Ahmed Abdullah<sup>2†</sup>, Muhammad Saad Saeed<sup>3†</sup>, Shah Nawaz<sup>1†</sup>, Rohan Kumar Das<sup>4†</sup>, Muhammad Zaigham Zaheer<sup>5</sup>, Junaid Mir<sup>6</sup>, Muhammad Haroon Yousaf<sup>6</sup>, Khalid Malik<sup>3</sup>, Markus Schedl<sup>6,7</sup>  
<sup>1</sup>Johannes Kepler University Linz, Austria, <sup>2</sup>National University of Computer and Emerging Sciences, Pakistan  
<sup>3</sup>University of Michigan, USA, <sup>4</sup>Fortimedia Singapore, Singapore  
<sup>5</sup>Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates  
<sup>6</sup>University of Engineering and Technology Taxila, Pakistan,  
<sup>7</sup>Human-centered AI Group, AI Lab, Linz Institute of Technology, Austria  
mavceleb@gmail.com

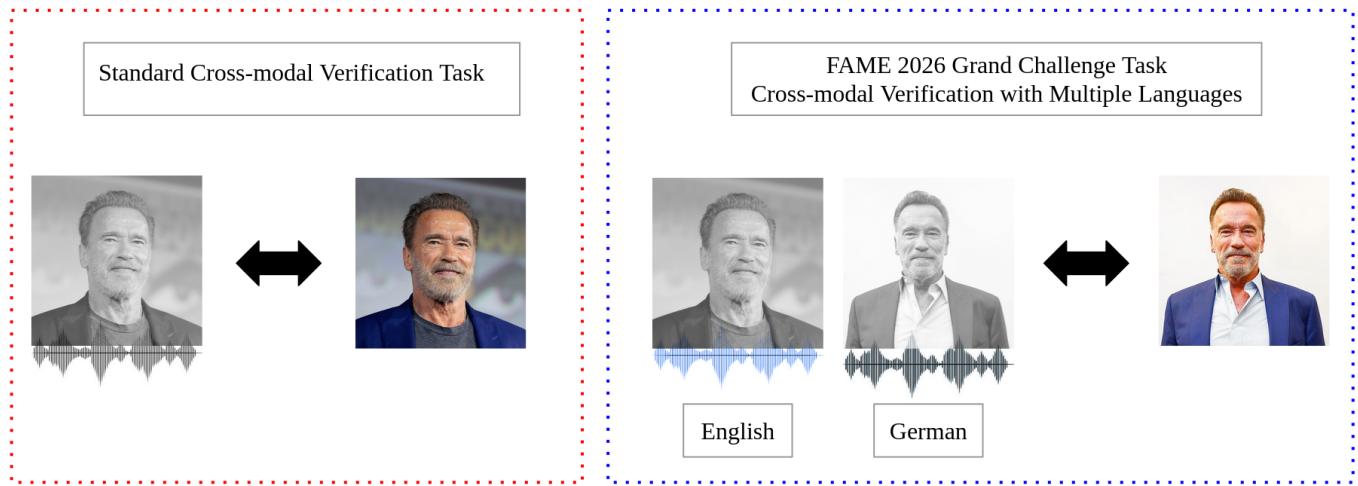


Fig. 1: (Left) Face-Voice association is established with a cross-modal verification task [1]. (Right) The FAME 2026 Challenge extends the task to analyze the impact of multiple languages.

**Abstract**—The advancements of technology have led to the use of multimodal systems in various real-world applications. Among them, audio-visual systems are among the most widely used multimodal systems. In the recent years, associating face and voice of a person has gained attention due to the presence of unique correlation between them. The Face-voice Association in Multilingual Environments (FAME) 2026 Challenge focuses on exploring face-voice association under the unique condition of a multilingual scenario. This condition is inspired from the fact that half of the world’s population is bilingual and most often people communicate under multilingual scenarios. The challenge uses a dataset named Multilingual Audio-Visual (MAV-Celeb) for exploring face-voice association in multilingual environments. This report provides the details of the challenge, dataset, baseline models, and task details for the FAME Challenge.

### I. INTRODUCTION

The face and voice of a person have unique characteristics and they are often used as biometric measures for person authentication, either as a unimodal or multimodal input [2], [3]. The strong correlation that humans establish in their perception of faces and voices of different people has inspired the development of tools for automated face-voice association

[4]–[8]. Though previous works have established association between faces and voices, none of these approaches investigated the effect of multiple languages on this task. Since half of the world’s population is bilingual and we are often communicating in multilingual scenarios, it is important to explore the influence of language on associating faces with voices. The goal of the FAME challenge, as also summarized in Fig. 1, is to analyze the impact of multilingual scenarios on face-voice association tasks.

### II. GRAND CHALLENGE OBJECTIVES

The goal of the Face-voice Association in Multilingual Environments (FAME) challenge, as also summarized in Fig. 1, is to analyze the impact of multilingual scenarios on face-voice association tasks. More concretely, the aim of the challenge is to promote research and development on the following aspects of face-voice association:

- Explore the influence of language on the performance of face-voice association algorithms
- Promote the development novel face-voice association algorithms that are effective in multilingual scenarios

<sup>†</sup>Equal Contribution.

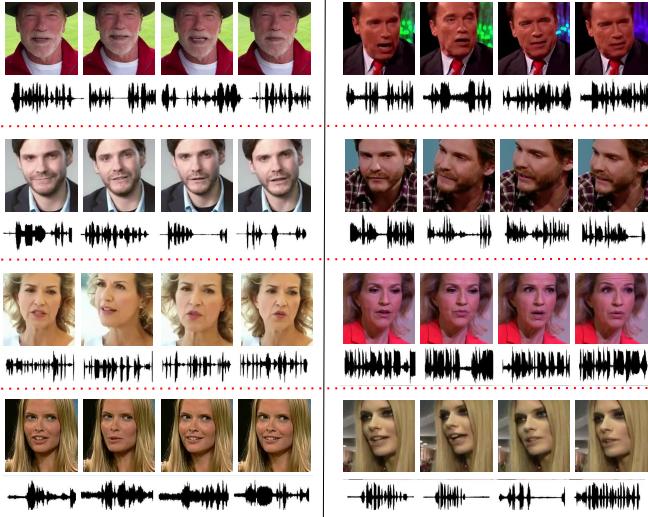


Fig. 2: Audio-visual samples selected from the MAV-Celeb dataset. The visual data contains different variations such as pose, lighting condition, and motion. The left block shows data of celebrities speaking English. The right block shows data of the same celebrities speaking German.

TABLE I: Summary of MAV-Celeb dataset.

Dataset	English-Urdu	English-German
Languages	E/U/EU	E/G/EG
# of celebrities	70	58
# of male celebrities	43	40
# of female celebrities	27	18
# of videos	560/406/966	212/216/428
# of hours	59/32/91	8.2/6.6/14.8
# of utterances	11835/6550/18385	2043/1769/3812
Avg # of videos per celebrity	8/6/14	3.7/3.7/7.4
Avg # of utterances per celebrity	169/94/263	35/31/66
Avg length of utterances (in s)	17.9/17.8/17.8	14.5/13.5/14.0

### III. GRAND CHALLENGE DESCRIPTION

**Dataset.** Following the prior FAME 2024 Grand Challenge hosted at ACM Multimedia [9], [10], we updated the dataset including an additional language as well as bilingual speakers, curating a new dataset split consisting of 58 English-German speakers. The split provides language annotations, which allow to analyze the impact of multiple languages on face-voice association. As for 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. Fig. 2 presents audio-visual samples from the newly collected dataset split, while Table I provides detailed statistics of the dataset.

The training and validation sets will be shared to the participants on the first day of the progress phase to allow the teams to develop their systems. In order to allow participants to benchmark their results, in the progress phase we will also release a pretrained model reaching competitive performance on the task. The progress phase will feature a leaderboard to keep participants up-to-date on the progress made by other teams on the validation set. The evaluation set (without ground-truth labels) will be shared on the first day of the evaluation phase. This will allow participants to test the generalization capability of the models developed and trained during the progress phase.

**Baseline Method & Starter Kit.** To allow participants to benchmark their results, we will release a pretrained instance of a competitive multimodal method for face-voice association task [12]. The model consists of a two-branch network that takes as input the embeddings of faces and voices. The embeddings to be used as input to the face-encoding branch are obtained using a popular convolutional neural network pre-trained on a large-scale facial recognition dataset [13]. The embeddings to be used as input to the voice-encoding branch are obtained using an audio encoding network for speaker recognition [14] trained using the language available in the training set (i.e., the *heard* language). The multimodal model further combines the face and voice embeddings, and is optimized by means of a loss function that imposes orthogonality constraints on the multimodal embeddings of different speakers (see Fig. 3). We refer the readers to Saeed et al. [12] and to the repository of the dataset<sup>1</sup> for more information on prior work on the baseline.

**Baseline results.** Table II provides the baseline face-voice association results with the impact of multiple languages on both splits of MAV-Celeb. The FAME challenge 2026 encourages participants to explore novel ideas to improve performance on heard and unheard languages.

**Challenge Setup.** The dataset consists of videos of several speakers. Each video is in one language only. However, each speaker is multilingual and appears in videos of at least two distinct languages. Each dataset set, i.e. the development set and the evaluation, is divided into train and test splits following the so-called *unseen-unheard* configuration [1]: the train and test splits consist of disjoint speakers speaking the same language [1]. Fig. 1 shows the evaluation protocol at train and test time. The evaluation is carried out on the task of cross-modal verification and on both a *heard* language, i.e., available during training, and an *unheard* language, i.e., *not* available during training. We refer to the dataset splits as V1-EU (64–6 train-test English-Urdu speakers), and V3-EG (50–8 train-test English-German speakers). V1-EU will be provided for the progress phase. The previous version of the dataset also includes V1-EU (64–6 train-test English-Urdu speakers). Although we will not use V2-EH as dataset to track teams' progress on the competition leaderboard, we will also provide V2-EH to allow teams to test their solutions on a separate dataset. The final evaluation will be carried out on the test

<sup>1</sup>[https://github.com/mavceleb/mavceleb\\_baseline](https://github.com/mavceleb/mavceleb_baseline)

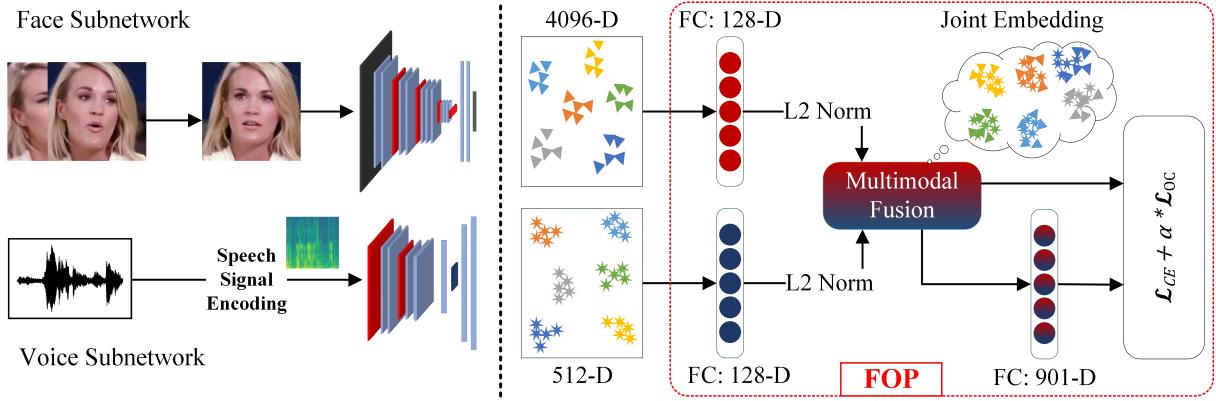


Fig. 3: Overall architecture of the baseline method. Face and voice embeddings are extracted by utilizing vision and audio encoders, respectively. Extracted features are then fed to linear layers to obtain the projected features of dimension  $D$ . Afterwards, embeddings are fused by using a gated feature fusion module. The fused features are fed to the logits layer. The model parameters are optimized by means of a linear combination of cross-entropy ( $L_{CE}$ ) and orthogonal constraints ( $L_{OC}$ ) losses.

set of the newly-developed V3-EG split. Alongside the audios (.wav) and images (.jpg), the datasets will include the .txt files of the face-voice pairs, for which each line has the following format:

- ysuvkz41 voices/English/00000.wav faces/English/00000.jpg
- tog3zj45 voices/English/00001.wav faces/English/00001.jpg
- ...
- ky5xfj1d voices/English/00002.wav faces/English/00002.jpg
- yx4nfa35 voices/English/01062.wav faces/English/01062.jpg

The first entry of the line (e.g. ysuvkz41) represents the id of the pair. The remaining two represent the local path of the corresponding audio (e.g. voices/English/00000.wav) and of a visual frame extracted from the video (e.g. faces/English/00000.jpg). This .txt file is associated with the ground-truth .txt file, containing the id of the pair, and a 1 or a 0, depending on whether the face and voice correspond to the same speaker or not.

The dataset is publicly available and provided alongside pre-extracted features representing the audios and images as encoded with state-of-the-art pre-trained architectures. We release all information related to the challenge on the challenge website.<sup>2</sup>

**Evaluation Metric.** As commonly done for tasks of face-voice association, we will evaluate the performance of participants' submissions with equal error rate (EER) as metric. The EER is the value of the false acceptance rate (FAR) and false rejection rate (FRR) for which both the errors are equal; As for FAR and FRR, a low value of EER indicates a good performance of the system. We expect participants to submit a .txt file containing output scores for every pair in the test

set, indicating the system's confidence that the face and voice are matching, or in other words, that they belong to the same person. As we will indicate in the challenge description, for a same .txt submission file, a face-voice test pair having a higher score than another face-voice test pair will be interpreted as the model having a higher confidence that the first pair is *matching*, i.e., that the face and voice correspond to the same person, as compared to the second pair. As in previous editions of the challenge, we choose this evaluation setup for several reasons. First, EER does not require the confidence scores of different models to span the same ranges in order to compare the models' performance. Second, EER does not require the use of a fixed threshold, as opposed to other metrics such as precision. Furthermore, in real-world applications, system developers may optimize the threshold on the confidence score to convert it to a binary value that determines the model's prediction on whether the face and voice belong to the same or to different person(s), depending on their specific needs. With a high threshold, the FAR is expected to be low, while the FRR is expected to be high. In summary, to evaluate the performance of different systems, EER is more suitable than threshold-dependent metrics such as accuracy, since it is independent of the confidence score ranges and of the threshold.

**Submission Format.** The grand challenge will be implemented using CodaBench<sup>3</sup>. Participants are expected to compute and submit text files including the id and confidence scores in the following format:

- ysuvkz41 0.9988
- tog3zj45 0.1146
- ...
- ky5xfj1d 0.6514
- yx4nfa35 1.5321
- bowsaf5e 1.6578

<sup>2</sup><https://mavceleb.github.io/dataset/competition.html>

<sup>3</sup><https://www.codabench.org/competitions/9467/>

TABLE II: Performance of baseline on cross-modal verification, between faces and voices across multiple language and on various test configurations of the MAV-Celeb dataset. Results indicate EER (in %), with lower values indicating a better performance. FOP [12] is used as baseline method.

V1-EU			
Configuration	Eng. test	Urdu test	Overall Score
Eng. train	29.3	37.9	33.4
Urdu train	40.4	25.8	
V3-EG			
	Eng. test	German test	Overall Score
Eng. train	34.5	43.7	40.2
German train	43.2	39.6	

Participants are expected to submit a zipped directory containing the submission files named as:

- sub\_score\_English\_heard.txt
- sub\_score\_English\_unheard.txt
- sub\_score\_Urdu\_heard.txt
- sub\_score\_Urdu\_unheard.txt

In the progress phase, each team will be allowed to submit a maximum of 100 submissions, with a maximum 10 per day. In the evaluation phase, the number of total submission will be limited to 5. The overall score will be computed as:

$$\text{Overall Score} = (\text{Sum of all EERs})/4 \quad (1)$$

**Rules for System Development.** Since the FAME 2026 challenge aims at analyzing whether face-voice association capabilities translate across languages, we will enforce the following rules for participation:

- A pretrained model on heard language is allowed.
- A pretrained model on unheard language is *not* allowed. The evaluation follows unheard-unseen and completely *unheard* protocol. Each celebrity in the split has audio information in two languages; the model will be trained on one language (e.g., German) and then tested on both the heard language (e.g., German) and completely *unheard* language (e.g., English).
- The participants are required to submit a 2-page system description following the ICASSP template to the challenge organizers. Teams without system description will be disqualified from the challenge. Teams describing a setup that violates one of the above rules will be disqualified.
- The participants are required to submit a link to a working version of their setup, e.g., on a platform for open-source development such as GitHub. Teams without code submission or with a setup that violates one of the above rules will be disqualified.

#### IV. REGISTRATION PROCESS

The following Google Form will be used to allow participants to register their teams to the challenge. [Registration form](#).

#### V. TENTATIVE TIMELINE

- Registration Period: 15 Aug. – 30 Sep. 2025
- Progress Phase: 1 Sep. – 31 Oct. 2025
- Evaluation Phase: 1 Nov. – 15 Nov. 2025
- Challenge Results: 16 Nov. 2025
- Submission of System Descriptions: 25 Nov. 2025
- Challenge Paper Submission: 07 Dec. 2025

#### VI. 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](#).

#### REFERENCES

- [1] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman, “Learnable pins: Cross-modal embeddings for person identity,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 71–88.
- [2] Anil K Jain, Arun Ross, and Salil Prabhakar, “An introduction to biometric recognition,” *IEEE Transactions on circuits and systems for video technology*, vol. 14, no. 1, pp. 4–20, 2004.
- [3] Arun Ross and Anil K Jain, “Multimodal biometrics: An overview,” in *2004 12th European signal processing conference*. IEEE, 2004, pp. 1221–1224.
- [4] 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 *2019 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2019, pp. 1–7.
- [5] Ruijie Tao, Rohan Kumar Das, and Haizhou Li, “Audio-Visual Speaker Recognition with a Cross-Modal Discriminative Network,” in *Proc. Interspeech 2020*, 2020, pp. 2242–2246.
- [6] Aihua Zheng, Menglan Hu, Bo Jiang, Yan Huang, Yan Yan, and Bin Luo, “Adversarial-metric learning for audio-visual cross-modal matching,” *IEEE Transactions on Multimedia*, vol. 24, pp. 338–351, 2021.
- [7] Saqlain Hussain Shah, Muhammad Saad Saeed, Shah Nawaz, and Muhammad Haroon Yousaf, “Speaker recognition in realistic scenario using multimodal data,” in *2023 3rd International Conference on Artificial Intelligence (ICAI)*. IEEE, 2023, pp. 209–213.
- [8] 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 *Proc. Interspeech 2025*.
- [9] 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.
- [10] 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, et al., “A synopsis of fame 2024 challenge: Associating faces with voices in multilingual environments,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 11333–11334.
- [11] Shah Nawaz, Muhammad Saad Saeed, Pietro Mororio, Arif Mahmood, Ignazio Gallo, Muhammad Haroon Yousaf, and Alessio Del Bue, “Cross-modal speaker verification and recognition: A multilingual perspective,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1682–1691.
- [12] 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 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7057–7061.
- [13] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman, “Deep face recognition,” 2015.
- [14] Weidi Xie, Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, “Utterance-level aggregation for speaker recognition in the wild,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5791–5795.