POLY-SIM: POLYglot Speaker Identification with Missing Modality Grand Challenge 2026

Muhammad Saad Saeed[1†], Marta Moscati[2†], Marina Zanoni[2,3], Mubashir Noman[4], Rohan Kumar Das[5], Monorama Swain[2], Yufang Hou[6], Elisabeth André[7], Khalid Mahmood Malik[1], Markus Schedl[2,8] Shah Nawaz[2]
[1]University of Michigan-Flint, USA, [2]Institute of Computational Perception, Johannes Kepler University Linz, Austria, [3]Sapienza University of Rome, Italy [4]Mohamed bin Zayed University of Artificial Intelligence, [5]Fortemedia Singapore, Singapore, [6]IT:U Interdisciplinary Transformation University Austria, [7]University of Augsburg, Germany, [8]Human-centered AI Group, AI Lab, Linz Institute of Technology, Austria
mavceleb@gmail.com

A. *A brief description to explain why the challenge problem is important and relevant to the multimedia research community, industry, and society over (at least) the next 3–5 years.*

1) Multimodal learning tasks under missing modalities, such as identifying a speaker when visual cues are unavailable, or when language conditions differ, reflect **real-world scenarios** faced by modern multimedia systems. Addressing this problem is therefore critical for building robust, flexible, and fair multimedia systems.
2) For the **research** community, this challenge pushes advances in representation learning, cross-modal alignment, domain adaptation, and generalization under distribution shift. For **industry**, it directly impacts the reliability of biometric systems, media analytics, human–computer interaction, and security applications deployed in real-world scenarios.
3) This challenge is a **continuation** of the previous FAME 2024 [1] and 2026 [2] Grand Challenges, which focused on understanding and analyzing the impact of language on face-voice association. Building on this foundation, the current challenge addresses a critical and increasingly realistic setting in which multimedia models must operate across languages while coping with missing modalities.

B. *A description of a specific set of research tasks or sub-tasks to tackle the challenge problem in the long run.*

To learn how to associate the face and voice of a same speaker, the model is trained on paired face images and segments of speech in one language (e.g., English). At inference time, the face modality is missing and the available speech segment is in a different language (e.g., German). This simulates both (i) the missing visual modality and (ii) the cross-lingual scenario, see Fig 1. Let $\mathcal{D}_{\text{train}} = \{(F_i^f, V_i^{a,\ell_{\text{en}}}, y_i)\}_{i=1}^{N}$ represent the training dataset consisting of N samples, where

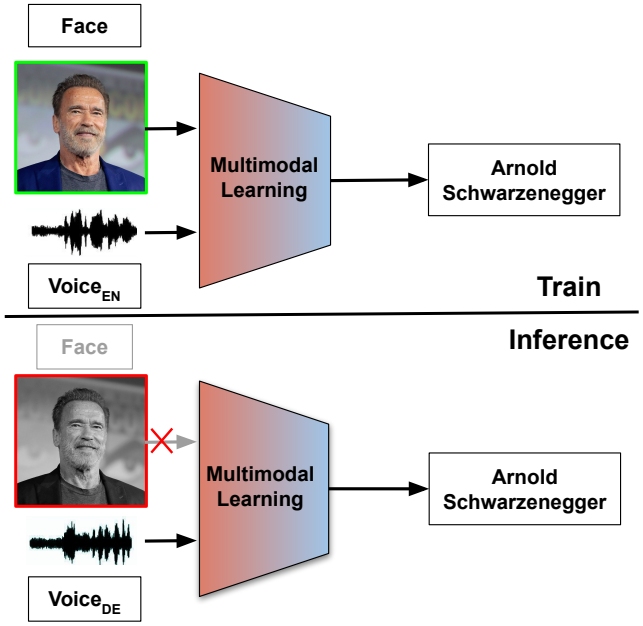

Fig. 1: POLY-SIM: POLYglot Speaker Identification with Missing Modality. The model is trained on paired face images and audio segments in a specific language (e.g., English). At inference, the face modality is missing and the input consists of audio segments in another language (e.g., German) only.

ℓ_{en} is a label indicating the training language, while F_i^f and $V_i^{a,\ell_{\text{en}}}$ denote the face and audio modalities, respectively. Each instance is associated with a class label $y_i \in \mathcal{Y}_{i=1}^{S}$, where S denotes the number of speakers. The modality-specific embeddings are defined as $x_i^f = \phi_f(F_i^f)$ and $x_i^{a,\ell_{\text{en}}} = \phi_a(V_i^{a,\ell_{\text{en}}})$, where $\phi_f(\cdot)$ and $\phi_a(\cdot)$ denote the face and audio encoders. At inference time, the face modality is missing and only the audio modality is available in a different language $\ell \in \mathcal{L}_{\text{test}} = \{\ell_{\text{de}}, \ell_{\text{ur}}, \ell_{\text{hi}}\}$, corresponding to German, Urdu, and Hindi. The test set is thus given by $\mathcal{D}_{\text{test}} = \{(V_j^{a,\ell})\}_{j=1}^{M}$ and analogously to the training time, the audio embedding at inference time is computed as $x_j^{a,\ell} = \phi_a(V_j^{a,\ell})$. The task's

[†]Equal Contribution.

goal is to predict the speaker label y_j based on the audio embedding in language $\ell \neq \ell^{\text{en}}$, and using a model f trained on paired audio–visual English data:

$$\hat{y}_j = f(x_j^{a,\ell}),$$

despite the *missing-modality* and the *language shift* in the audio modality.

C. *An outline of current state-of-the-art techniques and why this Grand Challenge would help accelerate research in this important area.*

Missing modalities. Multimodal learning models have demonstrated superior performance compared to unimodal methods [3]–[6]. However, such methods are primarily developed to work under modality-complete setting [7]. In contrast, real-world data is often incomplete due to factors such as sensor failure, and missing streams. These situations therefore constitute a significant challenges for existing multimodal methods [8]. In other words, due to their reliance on complementary information from multiple and diverse modalities, most existing multimodal learning methods exhibit a notable performance deterioration when evaluated under missing-modality scenarios [9], [10]. As multimodal learning are increasingly deployed in real-world scenarios, there is growing interest in investigating their vulnerability to missing data and enhancing their robustness to these situations [11], [12]. For example, Ma et al. [7] demonstrated that multimodal transformer models are not robust to missing data, and proposed the use of multi-task optimization to improve model robustness. More recently, Lee et al. [11] introduced missing-modality-aware prompts which are integrated into multimodal transformers to address missing data. Despite extensive research on missing-modality, existing strategies largely overlook cross-lingual settings. We bridge this gap by studying missing modalities in cross-lingual multimodal learning, a novel setting that more closely reflects real-world deployment.

Cross-lingual. Several studies [13]–[16] have shown that the performance of machine learning models on tasks related to voice and language, such as speaker verification, are strongly influenced by whether the evaluation language was present in the training set and, if not, by the similarity between the training and the evaluation languages. Similar trends have also been observed for models pre-trained on a high-resource language such English, and evaluated on lower-resource target languages. Although the pre-training provides useful representations, models experience a performance deterioration on target languages, and the deterioration is less substantial if the training and evaluation languages are related. To reveal these effects, several researchers shifted the focus from in-domain accuracy, to robustness under train-evaluation language mismatch. The interest in the topic is also reflected by the curation of dedicated benchmark datasets and the proposal of conference challenges. In speaker verification, challenges such as the VoxSRC [17] and the Short-Duration Speaker Verification Challenge [18] highlighted the impact of domain and language mismatch on model performance. These insights motivated language-aware training, calibration, and specific techniques to compensate cross-lingual score shifts [18]. Baali et al. [19] provide a substantial contribution to the topic by evaluating state-of-the-art models on a newly developed, open-set speaker verification benchmark dataset; the evaluation jointly considers language and age mismatches, as well as code compression, and explicitly targets realistic operating conditions as opposed to clean, monolingual settings. In the multimodal face-voice association domain, the FAME challenge [1], [20] and the MAV-Celeb [21] dataset play a similar role, since the training and test languages differ, and since models are also evaluated on missing-modality scenarios. Recent systems developed within this strand of research [22] demonstrated that even strong multimodal baselines remain sensitive to language mismatch and modality imbalance. Taken together, these works demonstrate how missing modalities, as well as imbalanced modalities between train and test data, can severely degrade the performance of face–voice association models, even those relying on the use of state-of-the-art pre-trained encoders for feature extraction. In this challenge we provide a substantial contribution to multimodal speaker identification, aimed at reducing the gap between research and real-world applications. To do so, we focus on the two real-world scenarios that we identified as crucial for model performance deterioration: the cross-lingual and the missing-modality scenarios. We design a simple, transparent framework that allows researchers to explicitly measure the impact of missing modalities on the performance on the task of multilingual face-voice matching.

D. *Link to sites containing relevant datasets to be used for objective training and evaluation of the Grand Challenge tasks. Documentation on the datasets should be provided or made accessible..*

We base our Grand Challenge on the MAV-Celeb dataset, which allows studying the impact of languages on face-voice association formulated as cross-modal verification task [21]. The dataset consists of audio-visual samples obtained from YouTube videos of speakers appearing in interviews, talk shows, and television debates. Most importantly, each speaker is bilingual, and we select the dataset subset in which each speaker appears in videos while speaking English and Urdu. We adapted the dataset for the task of multimodal speaker identification under missing-modality and cross-lingual scenarios. Specifically, for each speaker S and language ℓ we collect all available videos into $\mathcal{X}_{i,\ell} = \{x_{i,\ell}^{(1)}, \ldots, x_{i,\ell}^{(m_{i,\ell})}\}$ and split them into three mutually disjoint subsets, $\mathcal{X}_{i,\ell}^{\text{train}}$ (training), $\mathcal{X}_{i,\ell}^{\text{val}}$ (validation), and $\mathcal{X}_{i,\ell}^{\text{test}}$ (test), such that $|\mathcal{X}_{i,\ell}^{\text{val}}| = 1$, $|\mathcal{X}_{i,\ell}^{\text{test}}| = 1$, and $|\mathcal{X}_{i,\ell}^{\text{train}}| = m_{i,\ell} - 2$, with $\mathcal{X}_{i,\ell}^{\text{train}} \cap \mathcal{X}_{i,\ell}^{\text{va}} = \mathcal{X}_{i,\ell}^{\text{train}} \cap \mathcal{X}_{i,\ell}^{\text{te}} = \mathcal{X}_{i,\ell}^{\text{val}} \cap \mathcal{X}_{i,\ell}^{\text{te}} = \emptyset$, ensuring that the same video never appears in more than one split. Table I provides detailed statistics of the dataset, while Figure 2 presents audio-visual samples from the newly collected dataset split. We have provided dataset, embeddings, evaluation and baseline method[1] on the challenge

[1] https://github.com/msaadsaeed/polysim

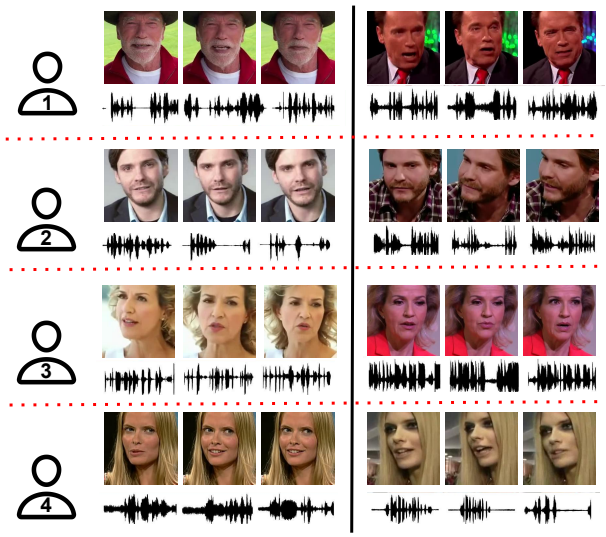

Fig. 2: Audio-visual samples randomly selected from the MAV-Celeb [20], [21], [23]. The visual data contains different variations such as pose, lighting condition, and motion. The left block shows data of speakers speaking English. The right block shows data of the same speakers speaking German language.

Lang. Pair	Lang.	Total Videos (Tr./Val./Test)	Audio-visual Samples (Tr./Val./Test)
Eng–Urdu	Eng	262 / 70 / 70	4039 / 1290 / 1521
	Urdu	415 / 70 / 70	9304 / 1779 / 1623
Eng–Hindi	Eng	478 / 84 / 84	9736 / 1386 / 1457
	Hindi	316 / 84 / 84	5739 / 1314 / 1083
Eng–German	Eng	84 / 36 / 36	679 / 249 / 232
	German	84 / 36 / 36	683 / 461 / 342

TABLE I: MAV-Celeb dataset statistics for English–Urdu, English–Hindi, and English–German language pairs.

website[2], which hence provides the complete experimental setup required by participants, allowing them to focus on method development and ensuring reproducibility and fair evaluation.

E. A description of rigorously defined objective criteria and/or procedures on how the submissions will be evaluated or judged.

We follow the protocol below to investigate the robustness of multimodal networks under missing-modality and cross-lingual scenarios.

P1. **In-language multimodal:** Training and testing on the same language with both modalities available.
P2. **Missing-modality:** Testing with only the audio modality while the face modality is missing.
P3. **Cross-lingual and multimodal:** Training on one language and testing on another with both modalities available.
P4. **Cross-lingual and missing-modality:** Cross-lingual testing under missing-modality.

We will use the top-1 accuracy (Acc.) metric to evaluate the performance of multimodal networks on P1, P2, P3, and P4 protocol. The challenge will be hosted on CodaBench[3].

F. Contact information of at least two organizers who will be responsible for organizing, publicizing, reviewing, and judging the Grand Challenge submissions as described in the proposal.

The five primary organizers of the challenge are:

1) Marta Moscati, PhD Student, Johannes Kepler University Linz, Austria; Applied Scientist, Albatross AI
2) Muhammad Saad Saeed, Master Student, University of Michigan-Flint, USA
3) Rohan Kumar Das, R&D Manager, Fortemedia Singapore, Singapore
4) Mubashir Noman, Postdoc, Mohamed bin Zayed University of Artificial Intelligence, UAE
5) Shah Nawaz, Assistant Professor, Johannes Kepler University Linz, Austria
6) Markus Schedl, Professor, Johannes Kepler University Linz, Austria

G. Acknowledgments.

This research was funded in whole or in part by the Austrian Science Fund (FWF): Cluster of Excellence *Bilateral Artificial Intelligence* (https://doi.org/10.55776/COE12) and the doc.funds.connect project *Human-Centered Artificial Intelligence* (https://doi.org/10.55776/DFH23).

REFERENCES

[1] 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.

[2] Marta Moscati, Ahmed Abdullah, Muhammad Saad Saeed, Shah Nawaz, Rohan Kumar Das, Muhammad Zaigham Zaheer, Junaid Mir, Muhammad Haroon Yousaf, Khalid Mahmood Malik, and Markus Schedl, "Linking faces and voices across languages: Insights from the fame 2026 challenge," *arXiv preprint arXiv:2512.20376*, 2025.

[3] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González, "Gated multimodal units for information fusion," *arXiv preprint arXiv:1702.01992*, 2017.

[4] Ignazio Gallo, Alessandro Calefati, and Shah Nawaz, "Multimodal classification fusion in real-world scenarios," in *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*. IEEE, 2017, vol. 5, pp. 36–41.

[5] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine, "The hateful memes challenge: Detecting hate speech in multimodal memes," *Advances in neural information processing systems*, vol. 33, pp. 2611–2624, 2020.

[6] Muhammad Saad Saeed, Shah Nawaz, Muhammad Zaigham Zaheer, Muhammad Haris Khan, Karthik Nandakumar, Muhammad Haroon Yousaf, Hassan Sajjad, Tom De Schepper, and Markus Schedl, "Modality invariant multimodal learning to handle missing modalities: A single-branch approach," *arXiv preprint arXiv:2408.07445*, 2024.

[2] https://mmosc.github.io/fame2027.github.io

[3] https://www.codabench.org/

[7] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng, "Are multimodal transformers robust to missing modality?," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18177–18186.

[8] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng, "Smil: Multimodal learning with severely missing modality," in *Proceedings of the AAAI conference on artificial intelligence*, 2021, vol. 35, pp. 2302–2310.

[9] Ronghao Lin and Haifeng Hu, "Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1686–1702, 2023.

[10] Christian Ganhör, Marta Moscati, Anna Hausberger, Shah Nawaz, and Markus Schedl, "A multimodal single-branch embedding network for recommendation in cold-start and missing modality scenarios," in *Proceedings of the 18th ACM Conference on Recommender Systems*, 2024, pp. 380–390.

[11] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee, "Multimodal prompting with missing modalities for visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14943–14952.

[12] Muhammad Irzam Liaqat, Qaiser Abbas, Shah Nawaz, Zaigham Zaheer, Marta Moscati, Yufang Hou, Muhammad Haris Khan, Salman Khan, Elisabeth Andre, and Markus Schedl, "Multimodal learning under imperfect data conditions: A survey," *Authorea Preprints*, 2025.

[13] Joyce Nakatumba-Nabende, Sulaiman Kagumire, Caroline Kantono, and Peter Nabende, "A systematic literature review on bias evaluation and mitigation in automatic speech recognition models for low-resource african languages," *ACM Comput. Surv.*, vol. 58, no. 4, Oct. 2025.

[14] Jenthe Thienpondt, Brecht Desplanques, and Kris Demuynck, "Cross-lingual speaker verification with domain-balanced hard prototype mining and language-dependent score normalization," 10 2020.

[15] Jenthe Thienpondt, Brecht Desplanques, and Kris Demuynck, "Tackling the score shift in cross-lingual speaker verification by exploiting language information," *Computer Speech and Language*, 2025.

[16] N. Reuter et al., "On the influence of language similarity in non-target cross-lingual speaker verification trials," in *Proc. Interspeech*, 2025.

[17] Jaesung Huh, Joon Son Chung, Arsha Nagrani, Andrew Brown, Jeeweon Jung, Daniel Garcia-Romero, and Andrew Zisserman, "The voxceleb speaker recognition challenge: A retrospective," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[18] Hossein Zeinali, Kong Aik Lee, Jahangir Alam, and Lukas Burget, "Short-duration speaker verification (sdsv) challenge 2021: the challenge evaluation plan," 2021.

[19] Massa Baali, Sarthak Bisht, Francisco Teixeira, Kateryna Shapovalenko, Rita Singh, and Bhiksha Raj, "Sveritas: Benchmark for robust speaker verification under diverse conditions," 2025.

[20] Marta Moscati, Ahmed Abdullah, Muhammad Saad Saeed, Shah Nawaz, Rohan Kumar Das, Muhammad Zaigham Zaheer, Junaid Mir, Muhammad Haroon Yousaf, Khalid Mahmood Malik, and Markus Schedl, "Linking faces and voices across languages: Insights from the fame 2026 challenge," 2025.

[21] Shah Nawaz, Muhammad Saad Saeed, Pietro Morerio, 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.

[22] Christopher Simic, Korbinian Riedhammer, and Tobias Bocklet, "Shared multi-modal embedding space for face-voice association," 2025.

[23] 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," 2025.