Cross-modal Speaker Verification and Recognition: A Multilingual Perspective

Shah Nawaz[1]*, Muhammad Saad Saeed[4]*, Pietro Morerio[1], Arif Mahmood [5], Ignazio Gallo[3], Muhammad Haroon Yousaf[4], Alessio Del Bue[1,2]

[1]Pattern Analysis & Computer Vision (PAVIS) - Istituto Italiano di Tecnologia (IIT),
[2]Visual Geometry & Modelling (VGM) - Istituto Italiano di Tecnologia (IIT),
[3]University of Insubria,
[4]University of Engineering and Technology Taxila, [5]Information Technology University

{shah.nawaz,pietro.morerio,alessio.delbue}@iit.it, arif.mahmood@itu.edu.pk,
{haroon.yousaf,saad.saeed}@uettaxila.edu.pk,ignazio.gallo@uninsubria.it

Abstract

Recent years have seen a surge in finding association between faces and voices within a cross-modal biometric application along with speaker recognition. Inspired from this, we introduce a challenging task in establishing association between faces and voices across multiple languages spoken by the same set of persons. The aim of this paper is to answer two closely related questions: "Is face-voice association language independent?" and "Can a speaker be recognized irrespective of the spoken language?". These two questions are important to understand effectiveness and to boost development of multilingual biometric systems. To answer these, we collected a Multilingual Audio-Visual dataset, containing human speech clips of 154 identities with 3 language annotations extracted from various videos uploaded online. Extensive experiments on the two splits of the proposed dataset have been performed to investigate and answer these novel research questions that clearly point out the relevance of the multilingual problem.

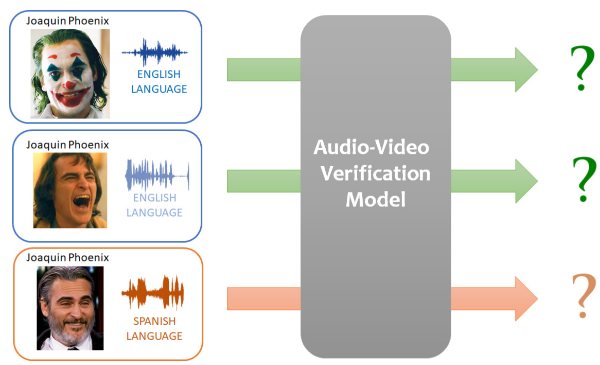

Figure 1. Multimodal data may provide enriched understanding to improve verification performance. Joaquin can wear make-up that makes visual identification challenging but voice can still bring enough cues to verify identity. In this work, we are interested to understand the effect of multilingual input when processed by audio-visual verification model (Q1) or just using the audio input (Q2). Joaquin is a perfect English-Spanish bilingual, would the system still be able to verify Joaquin when speaking Spanish even if the system was trained with English audio only?

1. Introduction

Half of the world population is bilingual with people often switching between their first and second language while communicating [28]. Therefore it is essential to investigate the effect of multiple languages on computer vision and machine learning tasks. As introduced in Fig. 1, this paper probes two closely related questions, which deal with the recent introduction of cross-modal biometric matching tasks in the wild:

Q1. Is face-voice association language independent?
Q2. Can a speaker be recognised irrespective of the spoken language?

Regarding the first question, a strong correlation has been recently found between face and voice of a person which has attracted significant research interest [18, 25, 31, 32, 35, 49, 50]. Though previous works have established an association between faces and voices, however none of these approaches investigate the effect of multiple languages on this task. In addition, existing datasets containing audio-visual information, *VoxCeleb* [33, 20, 34], *FVCeleb* [18], *FVMatching* [25] do not provide language level annotation. Therefore, we cannot deploy these datasets to analyse the effect of multiple languages on as-

*Equal contribution

sociation between faces and voices.

Thus, in order to answer both questions, we create a new Multilingual Audio-Visual *MAV-Celeb* dataset comprising of video and audio recordings with a large number of celebrities speaking more than one language in the wild. The proposed dataset paves the way to analyze the impact of multiple languages on association between faces and voices. Then, we propose a cross-modal verification approach to answer Q1 by analyzing the effect of multiple languages on face-voice association. In addition, the audio part of the dataset supplies samples of 3 languages with annotations which serves as a foundation to answer Q2 with speaker recognition baselines.
To summarise the paper main contributions are listed as follows:

- We first propose a cross-modal verification approach to analyze the effect of multiple languages on face-voice association;
- Likewise, we perform an analysis that highlights the very same problem of multilingualism for speaker recognition;
- We propose the *MAV-Celeb* dataset, containing $2,182$ language-annotated human speech clips with $41,674$ utterances of 154 celebrities, extracted from videos uploaded online.

The rest of the paper is structured as follows: Section 2 explores the related literature on the two introduced questions along with existing datasets. While Section 3 introduces the nature of proposed dataset followed by experimental evidence to answer both questions in Section 4 and 5. Finally, conclusion is presented in Section 6.

2. Related Work

We summarize previous work relevant to the two questions raised in the introduction. Q1 falls under cross-modal verification topic while Q2 deals with speaker recognition task.

2.1. Cross-modal Verification Between Faces and Voices

Last decade has witnessed an increasing use of multimodal data in challenging Computer Vision tasks including visual question and answering [2, 3], image captioning [22, 47], classification [17, 24], cross-modal retrieval [36, 48] and multimodal named entity recognition [5, 55]. Typically, multimodal applications are built on image and text information, however recent years have seen an increased interest to leverage audio-visual information [19, 37, 45, 52]. Previous works [1, 7] capitalize on natural synchronization between audio and visual information to learn rich audio representation via cross-modal distillation. More recently, Nagrani et al. [32] leveraged audio and visual information to establish an association between faces and voices in a cross-modal biometric matching. Furthermore, recent works [25, 31] introduced joint embedding to establish correspondences between faces and voices. These methods extract audio and face embedding to minimize the distance between embeddings of similar speakers while maximizing the distance among embeddings from different speakers. The framework used speaker identity information to eliminate the need of pairwise or triplet supervision [31, 32]. Wen et al. [50] presents a disjoint mapping network to learn a shared representation for audio and visual information by mapping them individually to common covariates (gender, nationality, identity). Similarly, Nawaz et al. [35] extracted audio and visual information with a single stream network to learn a shared deep latent space representation.

Our goal is similar to previous works [25, 31, 32, 50, 36], however, we investigate a novel problem: To understand if the association between faces and voices is language independent.

2.2. Speaker Recognition

Speaker recognition dates back to 1960s when Sandra et al. [40] laid the groundwork for speaker recognition systems attempting to find a similarity measure between two speech signals by using filter banks and digital spectrograms. In the following we provide a brief overview of speaker recognition methods as clustered in two main classes: Traditional and deep learning methods.
Traditional Methods – For a long time, low dimensional short-term representation of audio input has been basis for speaker recognition tasks e.g. Mel Frequency Cepstrum Coefficients (MFCC) and Linear Predictive Coding (LPC) based features. These features are extracted using short overlapping segments of audio samples. Reynolds et al. [42] introduced speaker verification method based on Gaussian Mixture Models using MFCCs. Differently, Joint Factor Analysis (JFA) models speaker and channel subspace separately [23]. Najim et al. [14] introduced i-vectors which combines both JFA and Support Vector Machines (SVM). Other works employed JFA as a feature extractor in order to train a SVM classifier. Furthermore, traditional methods have also been applied to analyze the effect of multiple languages on speaker recognition tasks [6, 27, 30]. Though, traditional methods showed reasonable performance on speaker recognition task, however these methods suffer performance degradation in real-world scenarios.
Deep Learning Methods – Neural Networks have provided more efficient methods of speaker recognition. Therefore, the community has experienced a shift from hand-crafted features to deep neural networks. Ellis et al. [15] introduced a system in which a Gaussian Mixture Model is trained from

Dataset	Condition	Free	Language annotations
The Mixer Corpus [13]	Telephone, Microphone	✗	✓
Vermobil [8]	Telephone, Microphone	✗	✓
Call My Net Corpus [21]	Telephone	✓	✓
Common Voice [4]	Microphone	✓	✓
SITW [29]	Multimedia	✓	✗
VoxCeleb [33, 20, 34]	Multimedia	✓	✗
MAV-Celeb (Proposed)	Multimedia	✓	✓

Table 1. Comparison of the proposed dataset with existing datasets.

Dataset	EU	EH
Languages	U/E/EU	H/E/EH
# of Celebrities	70	84
# of male celebrities	43	56
# of female celebrities	27	28
# of videos	560/406/966	546/668/1214
# of hours	59/32/91	48/60/109
# of utterances	11835/6550/18385	9974/13313/23287
Avg # of videos per celebrity	8/6/14	6/8/14
Avg # of utterances per celebrity	169/94/263	119/158/277
Avg length of utterances(s)	17.9/17.8/17.8	17.4/16.5/16.9

Table 2. Dataset statistics. The dataset is divided into 2 splits (EU, EH) containing audio samples from 3 languages, English(E), Hindi(H) and Urdu (U).

embedding of hidden layers of a neural network. Salman et al. [43] proposed a deep neural network which learn from speaker-specific characteristics from MFCC features for segmentation and clustering of speakers. Chen et.al. [11] used a Siamese feed forward neural network which can discriminatively compare two voices based on MFCC features. Lei et al. [26] introduced a deep neural model with i-vectors as input features for the task of automatic speaker recognition. Nagrani et al. [34] proposed adapted convolutional neural network (VGG-Vox) which aggregate frame-level feature vectors to obtain a fixed length utterance-level embedding. More recently, Xie et al. [53] improved this frame-level aggregation with NetVLAD or GhostVLAD layer. This paper has similarities with the previous work i.e. speaker identification and verification, however the objective is different: We evaluate and provide an answer about the effect of multiple languages on speaker identification and verification strategies in the wild. To this end we propose a dataset instrumental for answering such questions.

2.3. Related Datasets

There are various existing datasets for multilingual speaker recognition task but they are not instrumental to answer Q1/Q2 due to at least one of the following reasons: i) they are obtained in constrained environment [13]; ii) they are manually annotated so limited in size; iii) not freely available [8]; iv) not audio-visual [13, 21] v) missing language annotations [33, 20, 34]. A comparison of these dataset with our proposed MAV-Celeb dataset is provided in Table 1.

3. Dataset Description

Multilingual Audio-Visual *MAV-Celeb* dataset provide data of 154 celebrities in 3 languages (English, Hindi, Urdu). These three languages have been selected because of several factors: i) They represent approximately 1.4 Billion bilingual/trilingual people; ii) The population is highly proficient in two or more languages; iii) There is a relevant corpus of different media that can be extracted from available online repositories (e.g. YouTube). The collected videos cover a wide range of unconstrained, challenging multi-speaker environment including political debates, press conferences, outdoor interviews, quiet studio interviews, drama and movie clips.

It is also interesting to note that the visual data spans over a vast range of variations including poses, motion blur, background clutter, video quality, occlusions and lighting conditions. In addition, videos are degraded with real-world noise like background chatter, music, overlapping speech and compression artifacts. Fig. 2 shows some audio-visual samples while Table 2 shows statistics of the dataset. The dataset contains 2 splits English–Urdu (EU) and English–Hindi (EH) to analyze performance measure across multiple

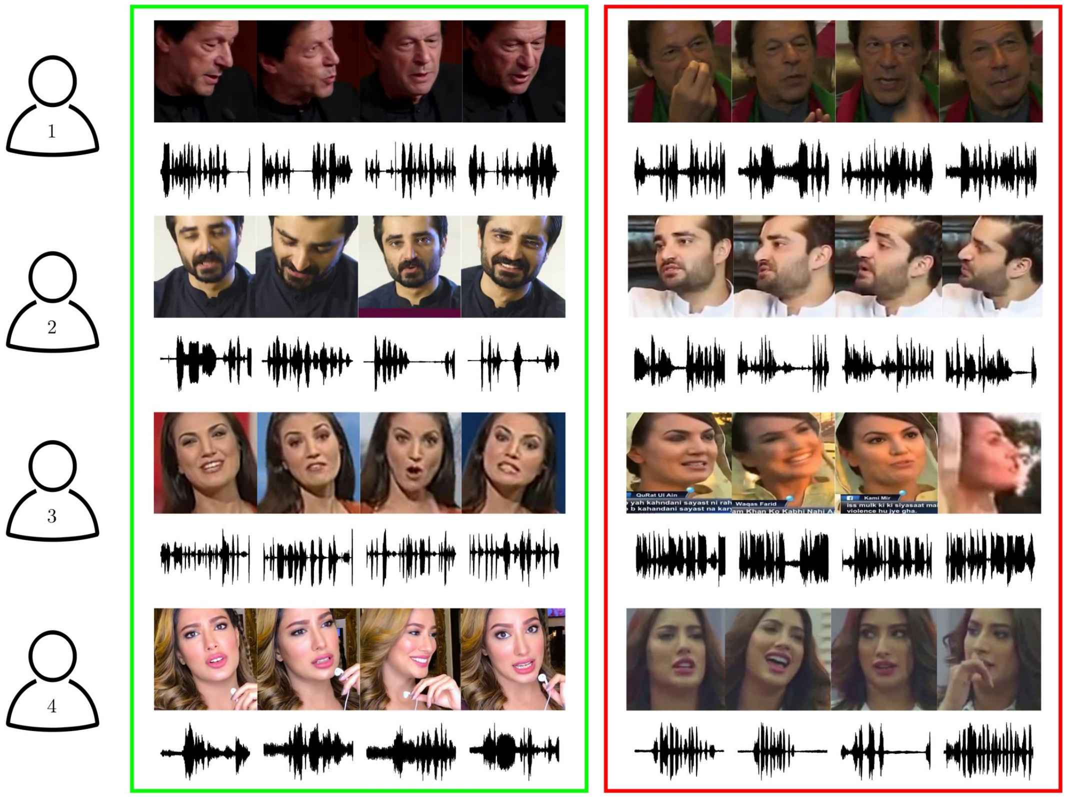

Figure 2. Audio-visual samples selected from proposed dataset. The visual data contains various variations such as pose, lighting condition and motion. The **green** block contains information of celebrities speaking English and the **red** block presents data of the same celebrity in Urdu.

languages. The pipeline followed in creating the dataset is discussed in Appendix A.

4. Face-voice Association

We introduced a cross-modal verification approach to analyze face-voice association across multiple languages using *MAV-Celeb* dataset in order to answer the question:

Q1. Is face-voice association language independent?

For example, consider a model trained with faces and voice samples of one language. At inference time, the model is evaluated with faces and audio samples of same language and a completely *unheard* language. This experimental setup provides a foundation to analyze association between faces and voices across languages to answer Q1 with a cross-modal verification method. Therefore, we extract face and voice embedding from two subnetworks trained on VGGFace [38] and voice samples from *MAV-Celeb* dataset respectively. Previous works showed that the faces and voices subnetworks can be trained jointly to bridge the gap between the two [25, 31]. However, we built a shallow architecture on top of face and voice embedding to reduce the gap between them to establish baseline on *MAV-Celeb* dataset. The approach is inspired from the previous work on images and text [48]. The details of these sub networks and shallow architecture are as follow:

Face Subnetwork – The face subnetwork is implemented using the VGG-Face CNN descriptor [38]. The input to the face subnetwork is an RGB image, cropped from the source frame to include only the face region and resized to 256×256. The final fully connected layer of the network produce embedding for every face input.

Voice Subnetwork – Nagrani et al. [34] introduced VGG-Vox network to process audio information. The network is trained with 'softmax' loss function in a typical classification scenario. In the current work, we configure the network to produce embedding from the *fc7* layer.

Cross-modal Verification – Finally, we learn a face-voice association for cross-modal verification approach using a two stream neural network with two layers of nonlinearities on top of the face and voice embedding. Fig. 3 shows the Two-Branch shallow architecture along with the pre-trained subnetworks. The shallow architecture consists of two branches, each composed of fully connected layer with weight matrices $A1$, $V1$ and $A2$, $V2$. In addition, layers are separated by Rectified ($ReLU$) followed by $L2$ normalization.

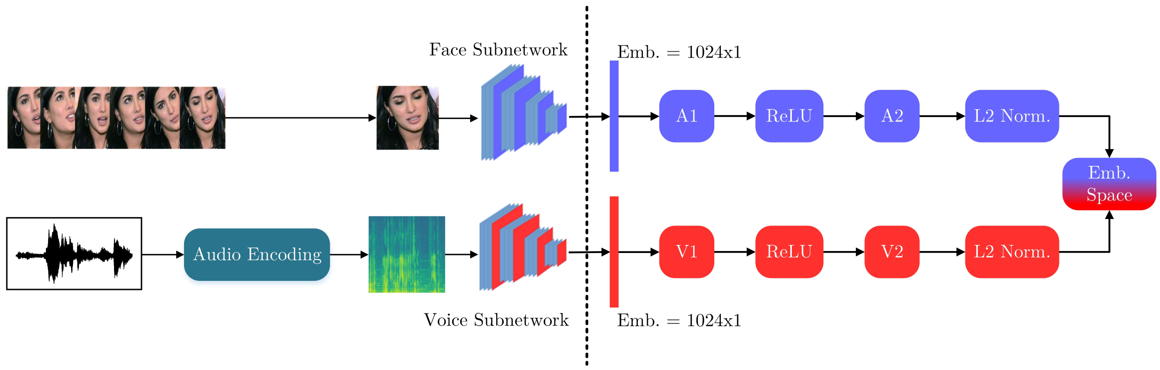

Figure 3. Cross-modal verification network configuration. The left side represents audio and face sub networks trained separately. Afterwards, audio and face embedding is extracted and fed to train a shallow architecture represented on the right side.

Loss Function – Given a training face f_i, let Y_i^+ and Y_i^- represent sets of positive and negative voice samples respectively. We impose the distance between f_i and each positive voice sample y_j to be smaller than the distance between f_i and each negative voice sample y_k with margin m:

$$d\left(f_i, y_j\right) + m < d\left(f_i, y_k\right) \quad \forall y_j \in Y_i^+, \forall y_k \in Y_i^-. \quad (1)$$

Eq. (1) is modified for a voice $y_{i'}$:

$$d\left(f_{j'}, y_{i'}\right) + m < d\left(f_{k'}, y_{i'}\right) \quad \forall f_{j'} \in X_{i'}^+, \forall f_{k'} \in X_{i'}^-, \quad (2)$$

where $X_{i'}^+$ and $X_{i'}^-$ represents the sets of positive and negative face for $y_{i'}$.

Finally, constraints are converted to the training objective using hinge loss. The resulting loss function is given by:

$$\begin{aligned} L(X,Y) = & \sum_{i,j,k} \max\left[0, m + d\left(f_i, y_j\right) - d\left(f_i, y_k\right)\right] \\ & + \lambda_1 \sum_{i',j',k'} \max\left[0, m + d\left(f_{j'}, y_{i'}\right) - d\left(f_{k'}, y_{i'}\right)\right] \\ & + \lambda_2 \sum_{i,j,k} \max\left[0, m + d\left(f_i, x_j\right) - d\left(f_i, x_k\right)\right] \\ & + \lambda_3 \sum_{i',j',k'} \max\left[0, m + d\left(y_{i'}, y_{j'}\right) - d\left(y_{i'}, y_{k'}\right)\right]. \end{aligned} \quad (3)$$

The shallow architecture configured with the loss function produce joint embedding of face and voice to study face-voice association across multiple languages using the proposed dataset. The hyperparameter λ_1 is fixed to 2. Similarly, λ_2 and λ_3 controls the neighborhood constraint and values are set to 0.1 or 0.2 respectively [48]. The distance d is fixed to be the Euclidean distance. In addition, triplets are selected within the mini-batch only.

4.1. Experimental Protocol

We propose an evaluation protocol for a cross-modal verification method in order to answer Q1. The aim of cross-modal verification task is to verify if an audio sample and a face image belong to the same identity or not based on a threshold value. We report performance on a standard verification metric i.e. Equal Error Rate (EER).

The *MAV-Celeb* dataset is divided into train and test splits consisting of disjoint identities from the same language typically known as *unseen-unheard* configuration [31, 32]. Fig. 4 shows evaluation protocol during training and testing stages. At inference time, the network is evaluated on a *heard* and completely *unheard* language. The protocol is more challenging than previously known *unseen-unheard* configuration due to the presence of an *unheard* language in addition to disjoint identities. The dataset splits EU, EH contains 64–6, 78–6 identities for train and test respectively.

4.2. Experiments and Results

In first set of experiments, we compare the performance of the proposed Two-Branch network on a cross-modal verification application with previous state-of-the-art between faces and voices [31, 35, 50]. We extracted face and voice embedding from pretrained VGG-Face CNN descriptor and VoxCeleb respectively. Finally, we train the Two-Branch network for a cross-modal verification application on top of face and voice embedding. Table 3 shows the result along with previous state-of-the-art methods. It is clear that the performance of our method is comparative with state-of-the-art methods, therefore we configure the approach to evaluate cross-modal verification method across multiple languages on *MAV-Celeb* dataset to establish baseline results.

In the second set of experiments, we evaluate Two-Branch network on cross-modal verification method between faces and voices to measure performance on *heard* and *unheard* configurations of *MAV-Celeb* dataset. Table 4

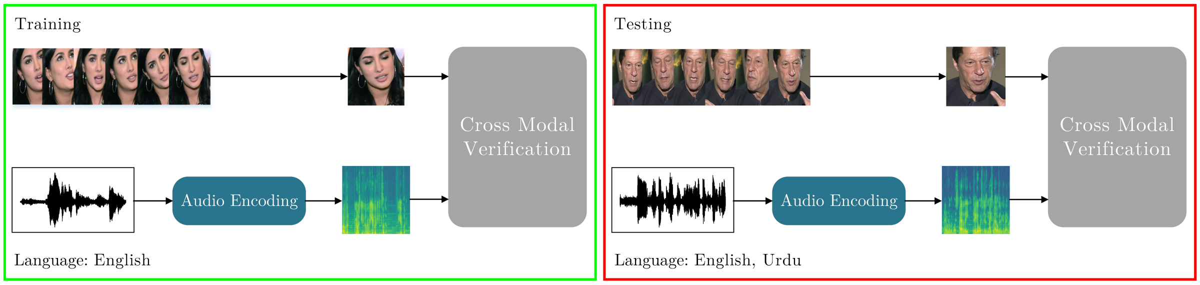

Figure 4. Evaluation protocol to analyze the impact of multiple languages on association between faces and voices. **Green** and the **red** blocks represent training and testing strategies. At test time, the network is evaluated on *unseen-unheard* configuration from the same language (English) *heard* during training along with a completely *unheard* language (Urdu).

Method	EER
Two-Branch (Proposed)	29.0
Learnable Pins-Scratch [31]	39.2
Learnable Pins-Pretrain [31]	29.6
Single Stream Network [35]	29.5
DIMNet [50]	**24.5**

Table 3. Cross-modal verification results with *unseen-unheard* configuration on VoxCeleb dataset. **(EER: lower is better)**

		EU	
Method	Configuration	Eng. test (EER)	Urdu test (EER)
Two-Branch (Proposed)	Eng. train	45.1	48.3↓7.1
	Urdu train	47.0↓6.3	44.3
		EH	
		Eng. test (EER)	Hindi test (EER)
Two-Branch (Proposed)	Eng. train	35.7	36.7↓2.8
	Hindi train	38.9↓4.2	37.3

Table 4. Cross-modal verification between face and voice across multiple language on various test configurations of *MAV-Celeb* dataset. The down arrow(↓) represents percentage decrease in performance. **(EER: lower is better)**

shows the result of face-voice association across multiple languages using the proposed evaluation protocol. We observed performance drop across 2 splits, which clearly demonstrate that the association between faces and voices is not language independent. We observed that the performance degradation is due to different data distributions of the two languages, typically known as domain shift [44]. Moreover, the model does not generalize well to other *unheard* language. However, the performance is still better than random verification, which is not trivial considering the challenging nature and configuration of the proposed evaluation protocol.

5. Speaker Recognition

This section investigates the performance of speaker recognition across multiple languages to answer the following question.

Q2. Can a speaker be recognised irrespective of the spoken language?

For example, consider a model trained with voice samples of one language. At inference time, the model is evaluated with audio samples of the same language and a completely *unheard* language of the same speaker. Therefore, the experimental setup provides a foundation for speaker recognition across multiple languages to answer Q2.

5.1. Baselines

We employed following 3 methods to establish baseline results for speaker recognition across multiple languages using *MAV-Celeb* dataset to answer Q2.

VGG-Vox – Nagrani et al. [34] introduced VGG-Vox network by modifying VGG-M [10] model to adapt to the spectrogram input. Specifically, the fully connected *fc6* layer of VGG-M is replaced by two layers – a fully connected layer and an average pool layer.

Utterance Level – Xie et al. [53] presented a deep neural network based on NetVLAD or GhostVLAD layer that is used to aggregate thin-ResNet architecture frame features.

SincNet – Ravanelli et al. [41] presented a deep neural model to process raw audio samples and learn features. The approach is based on parameterized sinc function for band-pass filtering that is used to convolve the wavefor to extract basic low-level features to be later processed by the deeper layers of the network.

5.2. Experimental Protocol

We proposed an evaluation protocol in order to analyze the impact of multiple languages on speaker recognition to answer Q2. The *MAV-Celeb* dataset is divided into typical

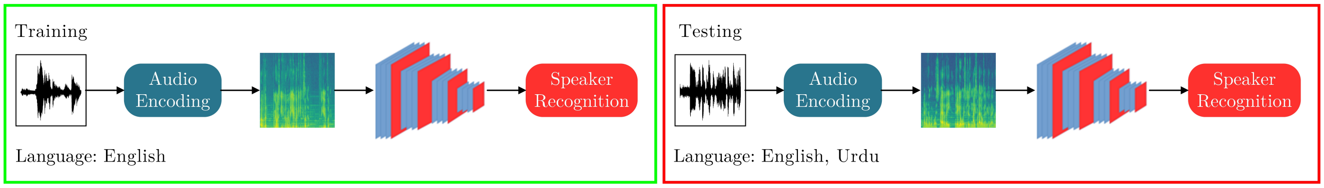

Figure 5. Evaluation protocol to analyze the impact of multiple languages on speaker recognition. **Green** and the **red** blocks represent training and testing strategies respectively. At test time, the network is evaluated on the same language *heard* during training along with completely *unheard* language of the same identities.

		EU	
Method	Configuration	Eng. test (Top-1)	Urdu test (Top-1)
VGGVox-Scratch	Eng. train	56.1	$37.7\downarrow_{32.8}$
	Urdu train	$45.4\downarrow_{19.9}$	56.7
VGGVox-Pretrain(VoxCeleb1)	Eng. train	41.0	$38.0\downarrow_{7.0}$
	Urdu train	$46.0\uparrow_{6.8}$	43.0
SincNet	Eng. train	32.5	$19.3\downarrow_{40.6}$
	Urdu train	$21.3\downarrow_{47.0}$	40.2
		EH	
		Eng. test (Top-1%)	Hindi test (Top-1%)
VGGVox-Scratch	Eng. train	60.2	$55.0\downarrow_{8.6}$
	Hindi train	$47.5\downarrow_{13.6}$	54.7
VGGVox-Pretrain(VoxCeleb1)	Eng. train	43.0	$32.0\downarrow_{25.5}$
	Hindi train	$43.0\downarrow_{10.4}$	48.0
SincNet	Eng. train	23.9	$8.7\downarrow_{63.6}$
	Hindi train	$14.4\downarrow_{43.0}$	25.3

Table 5. Speaker identification results across multiple languages on test configurations of *MAV-Celeb* dataset. The down arrow(↓) and up arrow(↑) represents percentage decrease and increase in performance. (**Top-1: higher is better**)

		EU	
Method	Configuration	Eng. test (EER)	Urdu test (EER)
VVGVox-Scratch	Eng. train	36.5	$41.5\downarrow_{13.7}$
	Urdu train	$40.3\downarrow_{3.0}$	39.1
Utterance Level-Scratch	Eng. train	39.9	$45.5\downarrow_{14.0}$
	Urdu train	$42.5\downarrow_{10.3}$	38.5
		EH	
		Eng. test (EER)	Hindi test (EER)
VVGVox-Scratch	Eng. train	29.6	$37.8\downarrow_{27.7}$
	Hindi train	$32.7\downarrow_{15.9}$	28.2
Utterance Level-Scratch	Eng. train	34.9	$40.6\downarrow_{16.3}$
	Hindi train	$42.7\downarrow_{18.9}$	35.9

Table 6. Speaker verification results across multiple languages on various test configurations of *MAV-Celeb* dataset. The down arrow(↓) represents percentage decrease in performance. (**EER: lower is better**)

classification scenario for speaker identification. However, different voice tracks of the same person are used for train, validation and test [34]. The network is trained with one language and tested with the same language and a completely *unheard* language of same identities. Moreover, the dataset is split into disjoint identities for speaker verification [34]. Fig. 5 shows evaluation protocol for speaker recognition across multiple languages. The protocol is consistent with previous studies on human subjects for speaker identification [39]. For identification and verification, we employed Top-1 accuracy and EER metrics to report performance.

5.3. Experiments and Results

We evaluate the performance of speaker recognition across multiple languages on 3 baseline methods. Table 5 shows speaker identification performance on 2 splits (EU, EH) of *MAV-Celeb* dataset. We note that the performance drop occurred on a completely *unheard* language across all baseline methods for both splits. The speaker identification models (VGG-Vox, SincNet) do not generalize well on *unheard* language and is overfitted on a particular language. However, its performance is quantitatively better than random classification on *unheard* language. Based on these results, we conclude that speaker identification is a language dependent task. Furthermore, these results are inline with the previous studies which show that human's speaker identification performance is higher on people speaking familiar language than people speaking *unknown* language [39].

Similarly, Table 6 shows speaker verification performance on 2 splits (EU, EH) of *MAV-Celeb* dataset. We note that performance drop occurred on a completely *unheard* language for EU and EH across three baseline methods. Therefore, speaker verification is also not language independent.

6. Conclusion

In this work, effect of language is explored on cross-modal verification between faces and voices along with speaker recognition tasks. A new audio-visual dataset consisting of 154 celebrities is presented with language level annotation. The dataset contains 2 splits having same set of identities speaking English/Urdu and English/Hindi. In the cross-modal verification experiment by changing train-

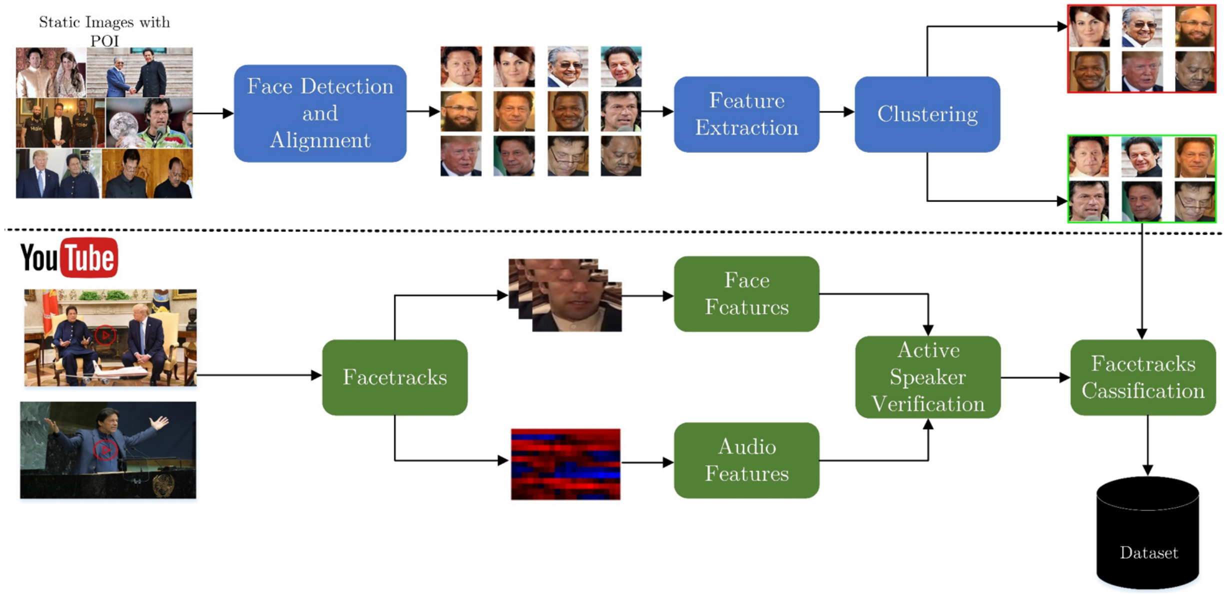

Figure 6. Data collection pipeline. It consists of two blocks with upper block download static images while the bottom block download and process videos from YouTube.

ing and test language, performance drop is observed indicating that face-association is not language independent. In case of speaker recognition, similar drop in performance is observed, thus concludes that speaker recognition is also language dependent task. The reason in performance is due to the domain shift caused by two different languages.

A. Dataset Collection Pipeline

In this section we present a semi-automated pipeline inspired by Nagrani et al. [34] for collecting the proposed dataset. The pipeline is shown in Fig. 6 and various stages are discussed below.

Stage 1 – List of Persons of Interest: In this stage, candidate list of Persons of Interest (POIs) is generated by scraping Wikipedia. The POIs cover over a wide range of identities including sports persons, actors, actresses, politicians, entrepreneurs and singers.

Stage 2 – Collecting list of YouTube links: In this step we used crowd-sourcing to collect lists of YouTube videos. Keywords like “Urdu interview”, “English Interview”,“public speech English”, “public speech Urdu” are appended to increase the likelihood that search results contain an instance of POI speaking. The links of search results are stored in text files. Videos are then automatically downloaded using the links from these text files.

Stage 3 – Face tracks: In this stage, we employed joint face detection and alignment using Multi-task Cascaded Convolutional Networks (MTCNN) for face detection and alignment [54]. MTCNN can detect faces in extreme conditions, and different poses. After face detection and alignment, shot boundaries are detected by comparing color histograms across consecutive frames. Based on key frames from shot boundaries and detected faces, face tracks are generated.

Stage 4 – Active speaker verification: The goal of this stage is to determine the visible speaking faces. We carried out this stage by using ‘SyncNet’ which estimates the correlation between mouth motion and audio tracks [12]. Based on scores from this model, face tracks with no visible speaking faces, voice-over and background speech are rejected.

Stage 5 – Static Images: In this stage, static images are automatically downloaded using Google Custom Search API based on list of POIs obtained from stage 1. MTCNN is employed to detect and align static face images. A clustering mechanism based on a popular density-based clustering algorithm DBSCAN [16] is used to remove false positives from the detected and aligned faces. Interestingly, DBSCAN does not require a priori specification of the number of clusters in the data. Intuitively, the clustering algorithm groups faces of an identity that are closely packed together.

Stage 6 – Face tracks classification: In this stage, active speaker face tracks are classified if they belong to POI or not. We trained an Inception ResNet V1 network [46] on VGGFace2 dataset [9] with center loss [51] to extract discriminative embedding from face tracks and static images. A classifier is trained based on Support Vector Machine with static face embedding. Finally, classification is performed using a score with a threshold obtained from each face track.

References

[1] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Emotion recognition in speech using cross-

modal transfer in the wild. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 292–301, 2018. 2

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 2

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2

[4] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019. 3

[5] Omer Arshad, Ignazio Gallo, Shah Nawaz, and Alessandro Calefati. Aiding intra-text representations with visual context for multimodal named entity recognition. In *2019 15th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019. 2

[6] Roland Auckenthaler, Michael J Carey, and John SD Mason. Language dependency in text-independent speaker verification. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, volume 1, pages 441–444. IEEE, 2001. 2

[7] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*, pages 892–900, 2016. 2

[8] Susanne Burger, Karl Weilhammer, Florian Schiel, and Hans G Tillmann. Verbmobil data collection and annotation. In *Verbmobil: Foundations of speech-to-speech translation*, pages 537–549. Springer, 2000. 3

[9] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018. 8

[10] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014. 6

[11] David Chen, Sam Tsai, Vijay Chandrasekhar, Gabriel Takacs, Huizhong Chen, Ramakrishna Vedantham, Radek Grzeszczuk, and Bernd Girod. Residual enhanced visual vectors for on-device image matching. In *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pages 850–854. IEEE, 2011. 3

[12] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016. 8

[13] Christopher Cieri, Joseph P Campbell, Hirotaka Nakasone, Kevin Walker, and David Miller. The mixer corpus of multilingual, multichannel speaker recognition data. Technical report, PENNSYLVANIA UNIV PHILADELPHIA, 2004. 3

[14] Najim Dehak, Patrick Kenny, Reda Dehak, Ondrej Glembek, Pierre Dumouchel, Lukas Burget, Valiantsina Hubeika, and Fabio Castaldo. Support vector machines and joint factor analysis for speaker verification. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4237–4240. IEEE, 2009. 2

[15] Daniel PW Ellis, Rita Singh, and Sunil Sivadas. Tandem acoustic modeling in large-vocabulary recognition. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, volume 1, pages 517–520. IEEE, 2001. 2

[16] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996. 8

[17] 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)*, volume 5, pages 36–41. IEEE, 2017. 2

[18] Shota Horiguchi, Naoyuki Kanda, and Kenji Nagamatsu. Face-voice matching using cross-modal embeddings. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1011–1019, 2018. 1

[19] Jing Huang and Brian Kingsbury. Audio-visual deep learning for noise robust speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7596–7599. IEEE, 2013. 2

[20] A. Zisserman J. S. Chung, A. Nagrani. Voxceleb2: Deep speaker recognition. In *INTERSPEECH, 2018*, 2018. 1, 3

[21] Karen Jones, Stephanie M Strassel, Kevin Walker, David Graff, and Jonathan Wright. Call my net corpus: A multilingual corpus for evaluation of speaker recognition technology. In *INTERSPEECH*, pages 2621–2624, 2017. 3

[22] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 2

[23] Patrick Kenny. Joint factor analysis of speaker and session variability: Theory and algorithms. *CRIM, Montreal,(Report) CRIM-06/08-13*, 14:28–29, 2005. 2

[24] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. Efficient large-scale multi-modal classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

[25] Changil Kim, Hijung Valentina Shin, Tae-Hyun Oh, Alexandre Kaspar, Mohamed Elgharib, and Wojciech Matusik. On learning associations of faces and voices. In *Asian Conference on Computer Vision*, pages 276–292. Springer, 2018. 1, 2, 4

[26] Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren. A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1695–1699. IEEE, 2014. 3

[27] Liang Lu, Yuan Dong, Xianyu Zhao, Jiqing Liu, and Haila Wang. The effect of language factors for robust speaker recognition. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4217–4220. IEEE, 2009. 2

[28] Jay Mathews. 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, 2019. [Online; accessed 16-April-2021]. 1

[29] Mitchell McLaren, Luciana Ferrer, Diego Castan, and Aaron Lawson. The speakers in the wild (sitw) speaker recognition database. In *Interspeech*, pages 818–822, 2016. 3

[30] Abhinav Misra and John HL Hansen. Spoken language mismatch in speaker verification: An investigation with nist-sre and crss bi-ling corpora. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 372–377. IEEE, 2014. 2

[31] 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)*, pages 71–88, 2018. 1, 2, 4, 5, 6

[32] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8427–8436, 2018. 1, 2, 5

[33] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60:101027, 2020. 1, 3

[34] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017. 1, 3, 4, 6, 7, 8

[35] 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)*, pages 1–7. IEEE, 2019. 1, 2, 5, 6

[36] Shah Nawaz, Muhammad Kamran Janjua, Ignazio Gallo, Arif Mahmood, Alessandro Calefati, and Faisal Shafait. Do cross modal systems leverage semantic relationships? In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[37] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. 2011. 2

[38] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015. 4

[39] Tyler K Perrachione, Stephanie N Del Tufo, and John DE Gabrieli. Human voice recognition depends on language ability. *Science*, 333(6042):595–595, 2011. 7

[40] Sandra Pruzansky. Pattern-matching procedure for automatic talker recognition. *The Journal of the Acoustical Society of America*, 35(3):354–358, 1963. 2

[41] Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with sincnet. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 1021–1028. IEEE, 2018. 6

[42] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1-3):19–41, 2000. 2

[43] Ahmad Salman and Ke Chen. Exploring speaker-specific characteristics with deep learning. In *The 2011 International Joint Conference on Neural Networks*, pages 103–110. IEEE, 2011. 3

[44] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000. 6

[45] Nitish Srivastava and Russ R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012. 2

[46] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017. 8

[47] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 2

[48] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016. 2, 4, 5

[49] Peisong Wen, Qianqian Xu, Yangbangyan Jiang, Zhiyong Yang, Yuan He, and Qingming Huang. Seeking the shape of sound: An adaptive framework for learning voice-face association. *arXiv preprint arXiv:2103.07293*, 2021. 1

[50] Yandong Wen, Mahmoud Al Ismail, Weiyang Liu, Bhiksha Raj, and Rita Singh. Disjoint mapping network for cross-modal matching of voices and faces. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. 1, 2, 5, 6

[51] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016. 8

[52] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6292–6300, 2019. 2

[53] 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)*, pages 5791–5795. IEEE, 2019. 3, 6

[54] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 8

[55] Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. Adaptive co-attention network for named entity recognition in tweets. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2