Cross-modal Speaker Verification and Recognition: A Multilingual Perspective

Shah Nawaz\textsuperscript{1}\textsuperscript{*}, Muhammad Saad Saeed\textsuperscript{4}\textsuperscript{*}, Pietro Morerio\textsuperscript{1}, Arif Mahmood\textsuperscript{5}, Ignazio Gallo\textsuperscript{3}, Muhammad Haroon Yousaf\textsuperscript{4}, Alessio Del Bue\textsuperscript{1,2}

\textsuperscript{1}Pattern Analysis & Computer Vision (PAVIS) - Istituto Italiano di Tecnologia (IIT),
\textsuperscript{2}Visual Geometry & Modelling (VGM) - Istituto Italiano di Tecnologia (IIT),
\textsuperscript{3}University of Insubria,
\textsuperscript{4}University of Engineering and Technology Taxila,
\textsuperscript{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.

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?

\textsuperscript{*}Equal contribution

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Condition</th>
<th>Free Language annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Mixer Corpus [13]</td>
<td>Telephone, Microphone</td>
<td>✓</td>
</tr>
<tr>
<td>Vermobil [8]</td>
<td>Telephone, Microphone</td>
<td>✓</td>
</tr>
<tr>
<td>Call My Net Corpus [21]</td>
<td>Telephone</td>
<td>✓</td>
</tr>
<tr>
<td>SITW [29]</td>
<td>Multimedia</td>
<td>✓</td>
</tr>
<tr>
<td>VoxCeleb [33, 20, 34]</td>
<td>Multimedia</td>
<td>✓ X</td>
</tr>
<tr>
<td>MAV-Celeb (Proposed)</td>
<td>Multimedia</td>
<td>✓ ✓</td>
</tr>
</tbody>
</table>

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).

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

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
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 $256 \times 256$. 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 non-linearities 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.
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 $m$:

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

(1)

Eq. (1) is modified for a voice $y_i'$:

$$d(f_i', y_i') + m < d(f_i', y_i') \quad \forall f_i' \in X_i'^+, \forall f_i' \in X_i'^-.$$  

(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:

$$L(X, Y) = \sum_{i,j,k} \max[0, m + d(f_i, y_j) - d(f_i, y_k)] + \lambda_1 \sum_{i',j',k'} \max[0, m + d(f_i', y_{i'}) - d(f_k', y_{i'})] + \lambda_2 \sum_{i,j,k} \max[0, m + d(f_i', x_j) - d(f_i, x_k)] + \lambda_3 \sum_{i',j',k'} \max[0, m + d(y_i', y_{i'}) - d(y_{i'}, y_{k'})].$$  

(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 $\lambda_1$ is fixed to 2. Similarly, $\lambda_2$ and $\lambda_3$ 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
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 bandpass filtering that is used to convolve the waveform 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

---

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

<table>
<thead>
<tr>
<th>Method</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-Branch (Proposed)</td>
<td>29.0</td>
</tr>
<tr>
<td>Learnable Pins-Scratch [31]</td>
<td>39.2</td>
</tr>
<tr>
<td>Learnable Pins-Pretrain [31]</td>
<td>29.6</td>
</tr>
<tr>
<td>Single Stream Network [35]</td>
<td>29.5</td>
</tr>
<tr>
<td>DIMNet [50]</td>
<td>24.5</td>
</tr>
</tbody>
</table>

**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)**

<table>
<thead>
<tr>
<th>Method</th>
<th>Configuration</th>
<th>Eng. test</th>
<th>Urdu test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-Branch (Proposed)</td>
<td>Eng. train</td>
<td>45.1</td>
<td>48.3\downarrow 3</td>
</tr>
<tr>
<td></td>
<td>Urdu train</td>
<td>47.0\downarrow 3</td>
<td>44.3</td>
</tr>
<tr>
<td>EH</td>
<td>Eng. test</td>
<td>35.7</td>
<td>36.7\downarrow 0.0</td>
</tr>
<tr>
<td></td>
<td>Hindi train</td>
<td>38.9\downarrow 2</td>
<td>37.3</td>
</tr>
</tbody>
</table>

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).
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 for 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.

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-


[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


[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


[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


