Lip Sync Matters: A Novel Multimodal Forgery Detector

Sahibzada Adil Shahzad*†, Ammarah Hashmi*‡, Sarwar Khan*†, Yan-Tsung Peng†, Yu Tsao§, Hsin-Min Wang*
* Social Networks and Human Centered Computing Program, Taiwan International Graduate Program,
Institute of Information Science, Academia Sinica
E-mail: {adilshah275, sarwar, whm}@iis.sinica.edu.tw
† Department of Computer Science, National Chengchi University, Taipei, Taiwan
E-mail: ytpeng@cs.nccu.edu.tw
‡ Institute of Information Systems and Applications, National Tsing Hua University, Taiwan
E-mail:hashmiammarah0@gmail.com
§ Research Center for Information Technology Innovation, Academia Sinica, Taiwan
E-mail: yu.tsao@citi.sinica.edu.tw

***Abstract*—Deepfake technology has advanced a lot, but it is a double-sided sword for the community. One can use it for beneficial purposes, such as restoring vintage content in old movies, or for nefarious purposes, such as creating fake footage to manipulate the public and distribute non-consensual pornography. A lot of work has been done to combat its improper use by detecting fake footage with good performance thanks to the availability of numerous public datasets and unimodal deep learning-based models. However, these methods are insufficient to detect multimodal manipulations, such as both visual and acoustic. This work proposes a novel lip-reading-based multimodal Deepfake detection method called "Lip Sync Matters." It targets high-level semantic features to exploit the mismatch between the lip sequence extracted from the video and the synthetic lip sequence generated from the audio by the Wav2lip model to detect forged videos. Experimental results show that the proposed method outperforms several existing unimodal, ensemble, and multimodal methods on the publicly available multimodal FakeAVCeleb dataset.**

I. Introduction

There are several types of fake media, of which Deepfakes are the most common nowadays. Deepfake content is primarily generated by deep learning algorithms, of which Generative Adversarial Networks (GANs) are best known for creating super-realistic forged content that is barely detectable by the naked human eyes. Forged videos can be generated by various techniques, such as Faceswap [1], FSGAN [2], Wav2lip [3], and real-time voice clones (SV2TTS) [4]. These technologies can manipulate and alter facial identities and lip movement, and even clone a target person's voice from a source person. Deepfakes can be used in many forms, among which spreading false political propaganda, generating fake adult videos, synthesizing Deepfake calls, generating fake news, stealing identities for financial gain, and slandering others have recently become common. Due to its inappropriate use, timely detection is important to avoid harm to society and individuals. A robust and efficient method is strongly desired to detect forgeries irrespective of Deepfake generation methods. Most existing

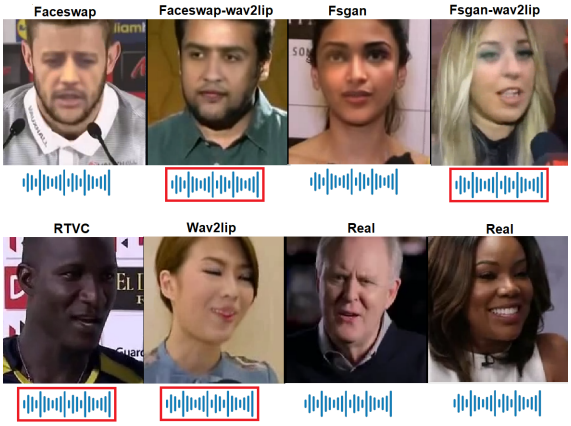

Fig. 1: Real and manipulated samples from the FakeAVCeleb dataset [5]. For Faceswap and Fsgan, only the facial part is manipulated, while the audio modality is real. For Faceswap-wav2lip, Fsgan-wav2lip, and wav2lip, both (audio and video) modalities are manipulated. For RTVC, the visual part is real, while the audio is cloned. For the real category, both modalities are real. The waveforms marked with red boxes represent fake audio with corresponding forged video.

Deepfake detectors are unimodal, and they are suitable for detecting manipulation in one modality. Recently, a new kind of Deepfakes has emerged on social networks and online, in which both audio and video modalities are manipulated, which makes such content more challenging to detect due to their multimodal manipulations. Detecting multimodal manipulations is very challenging for unimodal Deepfake detectors because they are primarily designed to detect one type of manipulation. These unimodal Deepfake detectors process one input at a time, either visual or acoustic, and thus utilize only limited information to distinguish between genuine and forged videos. Additionally, the lack of well organized and labeled multimodal Deepfake datasets is also an issue for designing robust and general multimodal Deepfake detection

978-616-590-477-3 ©2022 APSIPA

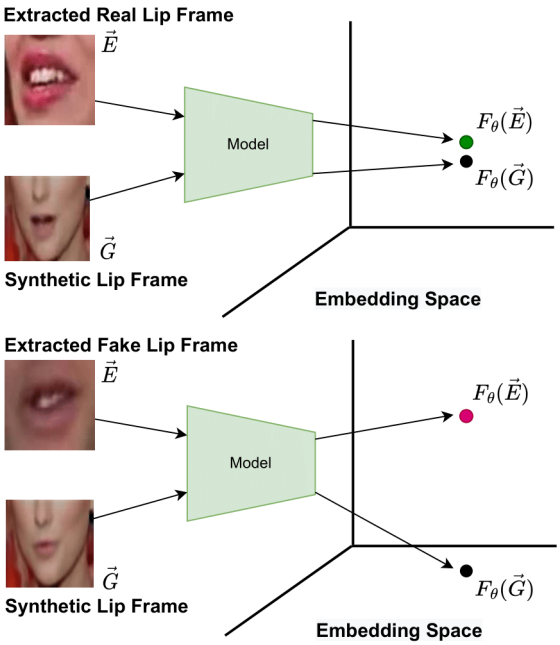

Fig. 2: Embedding space representation of Extracted Real Lips $\vec{E}$ from video with its corresponding Synthetic Lips $\vec{G}$ from audio modality (top). Similarly Extracted Fake Lips $\vec{E}$ from deepfake video with its corresponding Synthetic Lips $\vec{G}$ from audio modality (Bottom). $F_\theta(.)$ represents the forgery detection model that outputs a low dimensional vector $F_\theta(\vec{E})$ for each input Extracted Lips frame $\vec{E}$ and similarly, outputs $F_\theta(\vec{G})$ for each input Synthetic Lips frame $\vec{G}$.

systems. Fortunately, a group of researchers recently released a multimodal dataset called FakeAVCeleb [5]. This dataset is generated from the VoxCeleb2 [6] dataset by selecting videos of 500 celebrities. Each real video is a clean video with only one person's frontal face without occlusion. The FakeAVCeleb dataset is fairly balanced in terms of gender, race, geography, and visual and audio manipulations. Additionally, it covers many Deepfake generation techniques; thus, deep learning models trained with this balanced and diverse dataset can generalize well. Fig. 1 shows several real and manipulated samples from the FakeAVCeleb dataset. We propose a novel multimodal method called "Lips Sync Matters." Compared with state-of-the-art methods, the proposed approach is novel because we utilize visual and acoustic features in a novel way to examine synchronization between modalities. We assume that in the low-dimensional embedding space, for real videos, the embeddings of lip movements are close to those of the corresponding synthetic lip sequence, while for fake videos they are relatively separated, as shown in Fig. 2. Previous multimodal Deepfake detectors used visual information in the form of facial features, lip features, and emotional features and acoustic features, such as spectrogram and mel-frequency cepstral coefficients (MFCCs). Our proposed method mainly exploits two input modalities. First, we extract its lip sequence from the video. Second, we convert the audio modality to the visual modality by synthesizing the corresponding lip sequence from the audio stream using the Wav2lip student model [7]. Current multimodal Deepfake detectors typically train two or more models to handle multiple modalities, namely video and audio models. Although our proposed model also exploits two modalities, we take advantage of one model with a weight sharing strategy like Siamese Network [8] to train the Deepfake detector. Furthermore, a novel spatiotemporal Audio-Visual Lip-Sync Model is fine-tuned with lip sequences extracted from videos and synthesized from audio streams in videos. The proposed Audio-Visual Lip-Sync Model aims to capture high-level semantic features as well as spatial and temporal information from the input video to detect whether the extracted and synthesized lip sequences are synchronized. Audio-Visual consistency is still challenging for current Deepfake generation methods, which provides an opportunity to exploit the inconsistency between the two modalities to discover Deepfakes. The lip movement of forged videos are often out of sync with their audio stream, which we exploit in the form of synthetic lip sequences. The mismatch between the two lip movement is a reliable clue to detect whether the video content is manipulated or not.

The main contributions of our work are as follows:

- An audio-visual lipreading-based model is proposed, which can capture high-level semantic features to distinguish between genuine and forged videos.
- The proposed method is robust to noise as it exploits the visual form of audio features generated from a visual noise filter.
- The proposed method outperforms state-of-the-art methods on the multimodal FakeAVCeleb dataset.

II. Related Work

It is common to use fake images or videos to delude, defame or entertain others. People have been involved in contriving forged images and videos using various editing tools since the invention of photographs and films. With the advent of deep learning, advanced tools have become more available, and generating realistic fake content is much easier and faster than ever. Deepfake [9] caught the attention of the online community in 2017, when non-consensual adult videos with faces maliciously swapped from porn actors to popular actors were quickly shared on a platform called Reddit. Forged videos have been ubiquitous for years. Deepfakes are synthetic media that can be manipulated to generate compelling clips of people who say/do anything they never said/did, anywhere. Current Deepfake algorithms can generate forged images and videos and make it difficult for humans to distinguish between original and synthetic ones. Synthetic media generation tools are powered by Generative Adversarial Networks (GANs) [10], Autoencoders(AEs) [11], and Variational Autoencoders (VAEs) [11].

Deepfakes are usually generated by swapping the source person's face with the target person's face, aiming to make the target person do what the source person does. These artificial intelligence-synthesized media content can be roughly grouped

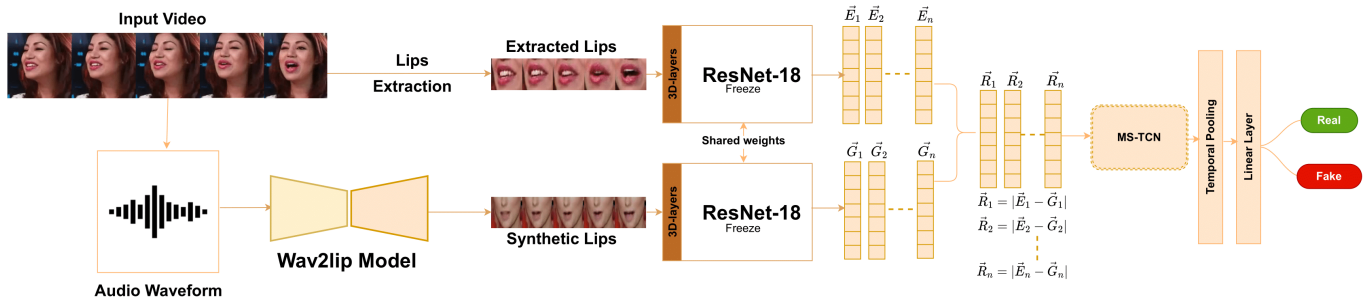

Fig. 3: The proposed architecture for multimodal forgery detection. The Lips sequence is extracted from the input (Real or Fake) video using Dlib pre-trained model. The audio modality from the video is extracted and used as input to Wav2lip Model to generate accurate lip movements. The backbone bone network has initial 3D-Conv layers followed by ResNet-18 pre-trained on a lipreading task. In the extracted lips sequence, a stack of lips frames $\vec{E}$ is fed to the feature extractor, which outputs low dimensional feature 512-D embeddings for each input frame. Similarly, the backbone model's feature extractor is used to feed the synthetic lips sequence (stack of synthetic lips frame $\vec{G}$) to generate 512-D embeddings for each input frame. Both the temporal network and linear classifier are fine-tuned using the absolute difference feature vector from both Extracted and Synthetic embeddings for final predictions.

into three [12] categories, namely Lip-syncing, Head puppetry, and Face swapping. Lip syncing is a video generated in a way that keeps the mouth motion consistent with a specific speech recording, so only the lip region is manipulated. While in the case of Head puppetry or Puppet-master, the target person is the puppet and the person whose action is followed is the master. The puppet-master video animates in a way that the puppet follows the expressions and head and eye motions of the master. Face swapping refers to replacing a source face with a target person's face without manipulating facial expressions. The remarkable progress in Deepfake generation raises serious security concerns and eventually becomes a threat to privacy matters [13]. Its misuse can shake out political propaganda, reduce journalism trust, or simply defame others [13]. Nowadays, not only is it easy to impersonate a person through image manipulation, voice cloning using deep learning is also available. These alerts prompted researchers to actively engage in Deepfake detection research.

A. Deepfake Detection

Fake video detection models have achieved excellent results due to various available datasets, such as DeepfakeTIMIT [14], UADVF [15], FaceForensics++ (FF++) [16], Celeb-DF [17], Google DFD [16], DFDC [18], DeeperForensics [19], KoDF [20], and the recently released multimodal FakeAVCeleb [5] dataset. Additionally, fake video detection is empowered by many unimodal video/image methods, such as Capsule Forensics [21], HeadPose [22], Xception [23], LipForensics [24], Meso-4 and MesoInception4 [25]. Likewise, to address audio spoofing attacks on automatic speaker verification systems, various methods utilizing different acoustic features have also been proposed [26], [27], [28]. Furthermore, some multimodal methods exploit faces as visual features and MFCCs as acoustic features [29], or detect forged videos based on facial and speech emotions [30].

III. Proposed Architecture

Fig. 3 shows our proposed model architecture for multimodal forgery detection. The lip sequence extracted from an input video is transformed into a vector representation sequence, $\{\vec{E}_i\}_{i=1}^{N}$, by a pre-trained ResNet-18 model. A pre-trained Wav2lip model is used to convert the audio track in the input video into a synthetic lip sequence, which is then transformed into a vector representation sequence, $\{\vec{G}_i\}_{i=1}^{N}$, by the same pre-trained ResNet-18 model. By sequentially subtracting each pair of vectors ($\vec{E}_i$ and $\vec{G}_i$) corresponding to the same time and taking the absolute value, a vector representation sequence, $\{\vec{R}_i\}_{i=1}^{N}$, of the input video can be obtained. Finally, a MS-TCN model and a temporal pooling layer are used to extract a single vector representation of the input video, and a linear layer is used for the final prediction. Next, we describe the core modules in detail.

A. Wav2lip Model

The Wav2lip model was adapted from [7], which was originally designed to assist speech enhancement systems when there is no real visual flow or face and lip regions are occluded. The Wav2lip model generates auxiliary information from lip movement from noisy speech. The Lip-sync generator [3] considered as the parent model is highly inaccurate on noisy speech input. The Wav2lip model is trained on a single identity to generate an accurate lip sequence for clean/noisy input speech. The student Wav2lip model is trained by a teacher Lip-sync expert [3], a pre-trained lip generating model that synthesizes lip movement on a static face by feeding clean speech. The student Wav2lip model was trained to mimic the Lip-sync expert model by feeding noisy speech with a static face image as input. The student model is trained on a clean speech from the LRS3 dataset with noise from the VGG-Sound dataset. The trained student model also works like a visual noise filter, so even in the presence of background noise, it generates an accurate lip sequence. Unless stated otherwise, we use the pre-trained model publicly available

here [1]. The proposed system for forgery detection exploits the synchronization between the extracted lip sequence and the generated lip sequence, so the Wav2lip student model is beneficial in our multimodal forgery detection task. The Wav2lip student model generates accurate lip movement in an unconstrained environment.

B. Audio-Visual Lip-Sync Model

We adapted the spatiotemporal CNN model [24] primarily designed for the unimodal forgery detection task. The unimodal Deepfake detector is inspired from a lipreading based model [31] and exploits lip sequences in videos to detect abnormal lip movements for forgery detection. The spatiotemporal model has a spatiotemporal feature extractor (2D ResNet18 with an initial 3D convolutional layer) that outputs a 512-D embedding for each input video frame. The feature extractor is followed by a multi-scale temporal convolutional network (MS-TCN) module to capture long- and short-term temporal information. MS-TCN is followed by a global average pooling layer and a linear classifier to output class probabilities.

For the multimodal forgery detection task, we modified the lipreading model to the Audio-Visual Lip-Sync Model. We froze the feature extractor part (2D ResNet18 with an initial 3D convolutional layer), pre-trained on the lipreading task. The spatiotemporal feature extractor outputs a 512-D embedding for each input video frame (Extracted and Synthetic). We hypothesize that a forged video's lip sequence is inconsistent with its audio counterpart. The proposed model exploits the inconsistency between visual and audio modalities. Furthermore, as shown in Fig. 2, we assume that in the low-dimensional embedding space, for real videos, the embeddings of extracted lip images are close to those of the corresponding synthetic lip images, while for fake videos, the embeddings of extracted lip images and synthetic lip images are relatively separated. Once the frozen feature extractor generates a low-dimensional representation for each frame, the corresponding visual and acoustic feature vectors ($\vec{E}_i$ and $\vec{G}_i$) are differentiated by calculating the absolute difference between the visual and acoustic representations as $\vec{R}_i$, i.e.,

$$\vec{R}_i = |\vec{E}_i - \vec{G}_i|, i = 1, ..., N. \tag{1}$$

Finally, the sequence difference vectors $\{\vec{R}_i\}_{i=1}^{N}$ is fed into MS-TCN, followed by a temporal pooling layer and a classifier to output the class label, either genuine or forged.

We pre-train the lipreading model on the Lipreading in the Wild (LRW) dataset to map lip image sequences to corresponding word sequence and use the feature extractor module to generate the representation sequences ($\vec{E}_i$ and $\vec{G}_i$). Let $D = \{(x_v^j, y_v^j)\}_{j=1}^{L}$ denote the training set, where x_v^j is a real or fake video, $y_v^j \in \{0, 1\}$ denotes whether the video is fake $\left(y_v^j = 0\right)$ or real $\left(y_v^j = 1\right)$, and L is the number of training videos. The prediction of the Audio-Visual Lip-Sync Model $F_{avlip}(.)$ is represented as $\hat{y}_v^j$:

$$\hat{y}_v^j = F_{avlip}(x_e^j, x_g^j), \tag{2}$$

where x_e^j denotes the extracted lip image sequence from video $x_v^j \in D$ and is represented by features $\{\vec{E}_i^j\}_{i=1}^{N}$, and x_g^i denotes the synthetic lip image sequence from the audio modality of the same video and is represented by features $\{\vec{G}_i^j\}_{i=1}^{N}$. The objective function used to optimize the model's trainable parameters is the cross entropy loss:

$$CEL = -\frac{1}{L}\sum_{j=1}^{L} y_v^j \log \hat{y}_v^j + \left(1 - y_v^j\right) \log\left(1 - \hat{y}_v^j\right). \tag{3}$$

IV. Experiments and Results

A. Dataset

We chose the recently released multimodal FakeAVCeleb dataset [5] for the task of multimodal forgery detection. There are several reasons for choosing this dataset for our experiments. The first and most important reason is that it contains multimodal manipulations, i.e., visual and acoustic manipulations. Furthermore, this dataset has manipulated videos from several Deepfake generation methods, including Wav2lip [3], Fsgan [2], Faceswap [1], Real-Time-Voice-Cloning (RTVC), Fsgan-wav2lip, and Facewap-wav2lip. In addition to covering diverse manipulation methods, it is gender-balanced and race-balanced, which eliminates bias in training machine learning models. The FakeAVCeleb dataset contains frontal face videos of celebrities, with one person in each video. The dataset contains 500 real videos and 20000+ fake videos, which is extremely imbalanced for training machine learning models. To deal with the imbalance, we added more real videos from the VoxCeleb1 dataset [32] to the real class and applied an oversampling strategy during training. The training set contains 4000 real videos and 17050 fake videos. For the test split, we designed eight different test sets to evaluate our proposed model. Six test sets were designed based on the manipulation method, namely Faceswap, Faceswap_wav2lip, Fsgan, Fsgan_wav2lip, RTVC, and Wav2lip test sets. The remaining two are generic test sets, called Test-set-1 and Test-set-2. Test-set-1 covers all manipulation methods and contains the same number of samples from Faceswap, Faceswap_wav2lip, Fsgan, Fsgan_wav2lip, RTVC, and Wav2lip in the fake class. Test-set-2 contains the same number of samples from RVFA (Real-Video-Fake-Audio), FVRA (Fake-Video-Real-Audio), and FVFA (Fake-Video-Fake-Audio) in the fake class. All the test sets are balanced in terms of real and fake samples and contain 70 samples per class (real and fake), which brings a total of 140 samples per test set.

B. Preprocessing

The proposed method mainly exploits lip features. Thus, the frontal face is required to extract the lip region using facial landmarks. For lip extraction, face detection is the primary step. The frontal face in the video is detected using a pre-trained CNN-based face detector in the Dlib toolkit [33], and

[1] https://github.com/Sindhu-Hegde/pseudo-visual-speech-denoising

TABLE I
RESULTS OF THE PROPOSED AUDIO-VISUAL LIP-SYNC METHOD ON VARIOUS TEST SETS.

Manipulation Method	Model	Class	Precision	Recall	F1-score	Accuracy
Faceswap	AV-Lip-Sync	Real	0.80	0.96	0.87	0.86
		Fake	0.95	0.76	0.84	
Faceswap_wav2lip	AV-Lip-Sync	Real	1.00	0.96	0.98	0.98
		Fake	0.96	1.00	0.98	
Fsgan	AV-Lip-Sync	Real	0.75	0.96	0.84	0.82
		Fake	0.94	0.69	0.79	
Fsgan_wav2lip	AV-Lip-Sync	Real	0.99	0.96	0.97	0.97
		Fake	0.96	0.99	0.97	
RTVC	AV-Lip-Sync	Real	0.85	0.96	0.90	0.89
		Fake	0.95	0.83	0.89	
Wav2lip	AV-Lip-Sync	Real	0.94	0.90	0.92	0.96
		Fake	0.96	0.96	0.96	
Test-set-1	AV-Lip-Sync	Real	0.92	0.96	0.94	0.94
		Fake	0.96	0.91	0.93	
Test-set-2	AV-Lip-Sync	Real	0.93	0.96	0.94	0.94
		Fake	0.96	0.93	0.94	

then the 96×96 RGB lip region is extracted according to the facial landmarks. The lip image frames are converted to grayscale before feeding to the model. During training, only 25 frames are selected from the entire video. We used the same number of frames as LipForensics [24] to make a fair comparison with the baseline model. The input shape of the extracted lip feature is $1 \times 25 \times 96 \times 96$ $(C \times F \times H \times W)$, where C stands for the number of channels, F for the number of frames, H and W are the height and width of each frame, respectively. The synthetic lip image sequence is synthesized from the audio modality using the Wav2lip model [7]. The input shape of the synthetic lip feature is also $1 \times 25 \times 96 \times 96$. It is worth mentioning that our input data size is twice that of LipForensics [24] and other unimodal [25] methods, but about eight times less than that of the recently proposed multimodal forgery detection method [29], which is $3 \times 25 \times 224 \times 224$.

C. Metrics

We evaluated our multimodal forgery detector in terms of four metrics, namely Accuracy, Precision, Recall, and F1-score. They are computed by

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4}$$

$$Precision = \frac{TP}{TP + FP}, \tag{5}$$

$$Recall = \frac{TP}{TP + FN}, \tag{6}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{7}$$

where TP, TN, FP, and FN stand for True Positive, True Negative, False Positive, and False Negative, respectively. For a fair comparison, we reported clip-level accuracy instead of frame-level accuracy.

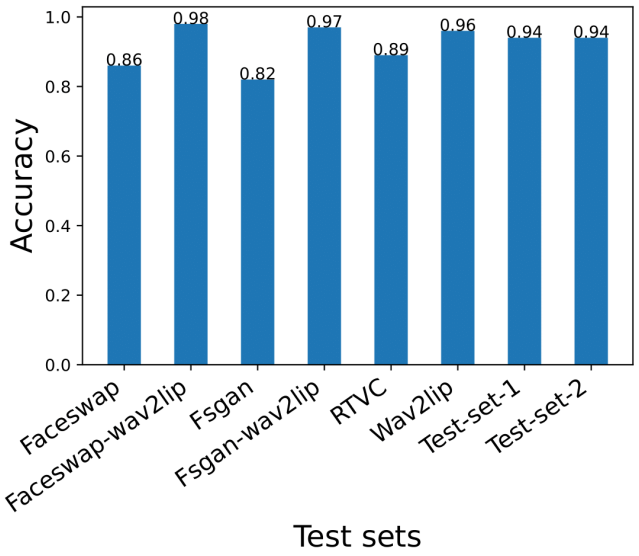

Fig. 4: Bar graph shows the accuracy of proposed Audio-Visual Lip-Sync Model with respect to various test sets reported in Table I. On the x-axis, we have different test sets and y-axis show respective accuracy.

D. Hyperparameters in Training

Our model was trained by the Adam optimizer with a learning rate of 0.0002 and batch size of 32. To deal with the imbalanced data issue, we added real videos from the VoxCeleb1 dataset [32] to the real class and applied oversampling during training.

E. Results

1) Evaluation of Audio-Visual Lip-Sync Model: The results of the proposed method evaluated on the eight test sets in terms of four evaluation metrics, namely precision, recall, F1-score, and accuracy, are shown in Table I. Fig. 4 shows a bar graph comparing accuracy on various test sets. From Table I and Fig. 4, it is clear that our multimodal Audio-Visual Lip-Sync Model performed well in most test sets. For example, in the

TABLE II
RESULTS OF OUR PROPOSED MULTIMODAL FORGERY DETECTION (AUDIO-VISUAL LIP-SYNC MODEL) METHOD COMPARED TO BASELINE UNIMODAL, ENSEMBLE AND MULTIMODAL METHODS. DFD REFERS TO THE DEEPFAKE DETECTION METHOD IN FIRST COLUMN. THE "V", "A" AND "AV" STANDS FOR VISUAL, AUDIO AND AUDIO-VISUAL MODALITY FOR EACH METHOD.

DFD Method	Model	Modality	Class	Precision	Recall	F1-score	Accuracy
Unimodal [34]	VGG16	V	Real	0.6935	0.8966	0.7821	0.8103
			Fake	0.8724	0.7750	0.8208	
Unimodal [34]	Xception	A	Real	0.8750	0.6087	0.7179	0.7626
			Fake	0.7033	0.9143	0.7950	
Ensemble (Soft-Voting) [34]	VGG16	AV	Real	0.6935	0.8966	0.7821	0.7804
			Fake	0.8948	0.6894	0.7788	
Ensemble (Hard-Voting) [34]	VGG16	AV	Real	0.6935	0.8966	0.7821	0.7804
			Fake	0.8948	0.6894	0.7788	
Multimodal-1 [34]	Multimodal-1	AV	Real	0.000	0.000	0.000	0.5000
			Fake	0.496	1.000	0.663	
Multimodal-2 [34]	Multimodal-2	AV	Real	0.710	0.587	0.643	0.674
			Fake	0.648	0.760	0.700	
Multimodal-3 [34]	CDCN	AV	Real	0.500	0.068	0.120	0.515
			Fake	0.500	0.940	0.651	
Multimodal-4 [29]	Not-made-for-each-other	AV	Real	0.62	0.99	0.76	0.69
			Fake	0.94	0.40	0.57	
Unimodal (E-lips) [24]	LipForensics	V	Real	0.70	0.91	0.80	0.76
			Fake	0.88	0.61	0.72	
Unimodal (G-lips) [24]	LipForensics	A	Real	0.71	0.34	0.46	0.60
			Fake	0.57	0.86	0.68	
Multimodal (ours)	**AV-Lip-Sync**	**AV**	**Real**	**0.93**	**0.96**	**0.94**	**0.94**
			Fake	**0.96**	**0.93**	**0.94**	

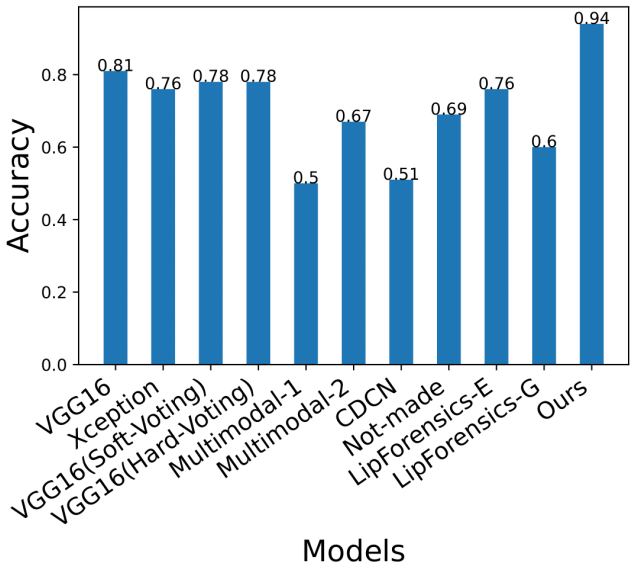

Fig. 5: Bar graph shows the accuracy with respect to various Deepfake detection models reported in Table II. On the x-axis, we have different forgery detection models and the y-axis shows respective accuracy.

case of Faceswap_wav2lip, Fsgan_wav2lip, Wav2lip, Test-set-1, and Test-set-2, it achieved 98%, 97%, 96%, 94%, and 94% accuracy, respectively. However, it performed slightly worse on the Faceswap, Fsgan, and RTVC test sets. There are two reasons for the poor performance on the Faceswap test set. First, the Faceswap manipulation category has fewer training examples than the other categories, with only 632 training examples. Second, Faceswap does not contain audio manipulations, which makes the proposed model less discriminative for this category, as it utilizes both modalities for final prediction. Our proposed model performed the worst on the Fsgan test set with only 82% accuracy. This is because the category contains synthetic faces using GANs without lip and audio manipulations in the forged video, which are hard to detect compared to the test sets using other manipulation methods. Considering that there are only 430 training examples for the RTVC category, where only the audio modality is manipulated, our model has reasonable performance and achieves 89% accuracy.

The results on all test sets indicate that our proposed Audio-Visual Lip-Sync Model successfully exploits high-level semantic features in the form of lip movements. Specifically, by detecting the consistency of speech and lip movements, it can distinguish between abnormal lip movements and real lip movements and whether lip movements match speech. The reasonable performance on the Faceswap and Fsgan test sets also indicates the effectiveness of the proposed model, even with limited lip manipulations and limited training samples. The results on the RTVC test set demonstrate that the proposed model succeeds in exploiting the synchronization issue between audio and video modalities in the case of voice cloning. Note that although our model performed worse on the Faceswap, Fsgan, and RTVC test sets than on the other test sets, it still outperformed unimodal (audio-only or video-only), ensemble, and multimodal models on these three test sets, and we will show the comparison results of different models later.

2) Comparing Audio-Visual Lip-Sync Model with other models: Next, we aimed to compare our model with various baseline unimodal, ensemble, and multimodal methods

following [34], where the authors evaluated different models on the multimodal FakeAVCeleb dataset. All the models were evaluated on the diverse test set, Test-set-2, which contains the same number of forged videos from FVFA (Fake-Video-Fake-Audio), FVRA (Fake-Video-Real-Audio), and RVFA (Real-Video-Fake-Audio) in the fake class. The results are shown in Table II and Fig. 5. The unimodal video models were trained on video features only, and the unimodal audio models were trained on the MFCC features. Ensemble and multimodal models were trained on both visual frames and MFCC features. From Table II, we can see that the unimodal method (VGG16) exploiting visual modality outperformed the ensemble and multimodal methods and achieved the highest accuracy of 81.03%. The results indicate that these ensemble and multimodal method are not effective. In contrast, our proposed multimodal Audio-Visual Lip-Sync Model, which was trained following the same training strategy as in [34], outperformed all the models evaluated in [34]. It achieved 94% accuracy, which is 12% higher than the accuracy of the best unimodal method (VGG).

Additionally, we evaluated the LipForensics model [24], which is a lipreading-based unimodal forgery detection method, on Test-set-2. The extracted (E_lips) and synthetic lips (G_lips) were used to train the model separately. Using the extracted lips (original lips from the video), the LipForensics model achieved 76% accuracy, about the same level of performance as the other models. The reason for not achieving high accuracy is that the LipForensics model only utilizes visual features, which is insufficient to detect multimodal Deepfake videos containing audio-visual manipulations or Deepfake videos containing audio-only manipulations. Using the synthetic lips, the LipForensics model achieved 60% accuracy. Synthetic lips are less discriminative because they are generated from the audio modality from a single identity, so the LipForensics model may only be effective in detecting Deepfake videos with visual-only manipulations. Furthermore, we trained the audio-visual dissonance-based model proposed by Komal Chugh *et al.* [29] on the FakeAVCeleb dataset. The model achieved an accuracy of 69%. The result confirms that simply utilizing spatiotemporal acoustic and visual features is not sufficient for the task of multimodal forgery detection. The high-level semantic features and synchronization between modalities are more effective and provide strong clues for detecting forgeries in multimodal data. Overall, our proposed multimodal Audio-Visual Lip-Sync Model outperformed all state-of-the-art forgery detection models by exploiting spatiotemporal artifacts and high-level semantic features that benefit from lip movements in audio and visual modalities.

3) Discussion: From the above experiments, we can see that the proposed audio-visual Deepfake detector performs favorably under various visual and audio manipulation conditions. Nonetheless, it still has some limitations that need to be addressed in future work. Our system exploits lip motions; therefore, the frontal face is required to extract the lip sequence to check the synchrony with the synthetic lip sequence. Furthermore, in the case of multimodal Deepfake detection, audio modality is required to generate an accurate lip sequence. Therefore, lip occlusion, far frontal face, and adversarial attack may lead to poor performance of the proposed Deepfake detector. In our future work, we will try to fuse our model with other models. It is hoped that these models can complement each other's deficiencies, and thus achieve better overall detection performance.

V. Conclusion

In this paper, we have proposed a novel approach for multimodal forgery detection. We introduced a synthetic lip sequence to benefit the multimodal Deepfake detector. Our proposed method not only exploits two modalities but also captures semantic features to ensure audio-visual consistency between the original lip sequence extracted from the forged video and the synthetic lip sequence generated from the audio modality of the same video. Furthermore, to confirm the robustness of the proposed multimodal Deepfake detector, extensive experiments have been performed to evaluate the model on multiple test sets with various Deepfake generation techniques using the recently released multimodal FakeAVCeleb dataset. Experimental results show that our model outperforms all the state-of-the-art unimodal, ensemble, and multimodal forgery detection methods compared in the paper.

Acknowledgment

This work was supported by the NSTC-Taiwan Grant 111-2221-E-001-002, 111-2221-E-004-010, and 110-2622-E-004-001.

References

[1] I. Korshunova, W. Shi, J. Dambre, L. Theis, Fast face-swap using convolutional neural networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 3677–3685.

[2] Y. Nirkin, Y. Keller, T. Hassner, Fsgan: Subject agnostic face swapping and reenactment, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 7184–7193.

[3] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, C. Jawahar, A lip sync expert is all you need for speech to lip generation in the wild, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 484–492.

[4] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu, et al., Transfer learning from speaker verification to multispeaker text-to-speech synthesis, Advances in neural information processing systems 31 (2018).

[5] H. Khalid, S. Tariq, M. Kim, S. S. Woo, FakeAVCeleb: A novel audio-video multimodal deepfake dataset, in: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.
URL https://openreview.net/forum?id=TAXFsg6ZaOl

[6] J. S. Chung, A. Nagrani, A. Zisserman, VoxCeleb2: Deep Speaker Recognition, in: Proc. Interspeech 2018, 2018, pp. 1086–1090. doi:10.21437/Interspeech.2018-1929.

[7] S. B. Hegde, K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, C. Jawahar, Visual speech enhancement without a real visual stream, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1926–1935.

[8] G. Koch, R. Zemel, R. Salakhutdinov, et al., Siamese neural networks for one-shot image recognition, in: ICML deep learning workshop, Vol. 2, Lille, 2015, p. 0.

[9] M. Westerlund, The emergence of deepfake technology: A review, Technology Innovation Management Review 9 (11) (2019).

[10] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, A. A. Bharath, Generative adversarial networks: An overview, IEEE Signal Processing Magazine 35 (1) (2018) 53–65.

[11] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning internal representations by error propagation, Tech. rep., California Univ San Diego La Jolla Inst for Cognitive Science (1985).

[12] S. Lyu, Deepfake detection: Current challenges and next steps, in: 2020 IEEE international conference on multimedia & expo workshops (ICMEW), IEEE, 2020, pp. 1–6.

[13] B. Chesney, D. Citron, Deep fakes: A looming challenge for privacy, democracy, and national security, Calif. L. Rev. 107 (2019) 1753.

[14] P. Korshunov, S. Marcel, Deepfakes: a new threat to face recognition? assessment and detection, arXiv preprint arXiv:1812.08685 (2018).

[15] X. Yang, Y. Li, S. Lyu, Exposing deep fakes using inconsistent head poses, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 8261–8265.

[16] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics++: Learning to detect manipulated facial images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1–11.

[17] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-df: A large-scale challenging dataset for deepfake forensics, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3207–3216.

[18] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C. C. Ferrer, The deepfake detection challenge (dfdc) dataset, arXiv preprint arXiv:2006.07397 (2020).

[19] L. Jiang, R. Li, W. Wu, C. Qian, C. C. Loy, Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 2889–2898.

[20] P. Kwon, J. You, G. Nam, S. Park, G. Chae, Kodf: A large-scale korean deepfake detection dataset, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10744–10753.

[21] H. H. Nguyen, J. Yamagishi, I. Echizen, Capsule-forensics: Using capsule networks to detect forged images and videos, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 2307–2311.

[22] K. Lutz, R. Bassett, Deepfake detection with inconsistent head poses: Reproducibility and analysis, arXiv preprint arXiv:2108.12715 (2021).

[23] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.

[24] A. Haliassos, K. Vougioukas, S. Petridis, M. Pantic, Lips don't lie: A generalisable and robust approach to face forgery detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5039–5049.

[25] D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, Mesonet: a compact facial video forgery detection network, in: 2018 IEEE international workshop on information forensics and security (WIFS), IEEE, 2018, pp. 1–7.

[26] L. Wang, Y. Yoshida, Y. Kawakami, S. Nakagawa, Relative phase information for detecting human speech and spoofed speech, in: Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[27] M. Todisco, H. Delgado, N. W. Evans, A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients., in: Odyssey, Vol. 2016, 2016, pp. 283–290.

[28] T. B. Patel, H. A. Patil, Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech, in: Sixteenth annual conference of the international speech communication association, 2015.

[29] K. Chugh, P. Gupta, A. Dhall, R. Subramanian, Not made for each other-audio-visual dissonance-based deepfake detection and localization, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 439–447.

[30] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, D. Manocha, Emotions don't lie: An audio-visual deepfake detection method using affective cues, in: Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 2823–2832.

[31] B. Martinez, P. Ma, S. Petridis, M. Pantic, Lipreading using temporal convolutional networks, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 6319–6323.

[32] A. Nagrani, J. S. Chung, A. Zisserman, Voxceleb: a large-scale speaker identification dataset, in: INTERSPEECH, 2017.

[33] D. E. King, Dlib-ml: A machine learning toolkit, The Journal of Machine Learning Research 10 (2009) 1755–1758.

[34] H. Khalid, M. Kim, S. Tariq, S. S. Woo, Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors, in: Proceedings of the 1st workshop on synthetic multimedia-audiovisual deepfake generation and detection, 2021, pp. 7–15.