Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches

Shruti Agarwal and Hany Farid
University of California, Berkeley
Berkeley, CA USA
{shruti_agarwal,hfarid}@berkeley.edu

Ohad Fried and Maneesh Agrawala
Stanford University
Stanford, CA USA
{ohad,maneesh}@cs.stanford.edu

Abstract

Recent advances in machine learning and computer graphics have made it easier to convincingly manipulate video and audio. These so-called deep-fake videos range from complete full-face synthesis and replacement (face-swap), to complete mouth and audio synthesis and replacement (lip-sync), and partial word-based audio and mouth synthesis and replacement. Detection of deep fakes with only a small spatial and temporal manipulation is particularly challenging. We describe a technique to detect such manipulated videos by exploiting the fact that the dynamics of the mouth shape – visemes – are occasionally inconsistent with a spoken phoneme. We focus on the visemes associated with words having the sound \( M \) (mama), \( B \) (baba), or \( P \) (papa) in which the mouth must completely close in order to pronounce these phonemes. We observe that this is not the case in many deep-fake videos. Such phoneme-viseme mismatches can, therefore, be used to detect even spatially small and temporally localized manipulations. We demonstrate the efficacy and robustness of this approach to detect different types of deep-fake videos, including in-the-wild deep fakes.

1. Introduction

Rapid advances in computer graphics, computer vision, and machine learning have led to the ability to synthesize highly realistic audio, image, and video in which anybody can be made to say and do just about anything. With enough sample recordings, for example, it is possible to synthesize realistic audio in anyone’s voice [22]. With enough sample images, it is possible to synthesize images of people who don’t exist [15, 16]. And, realistic videos can be created of anybody saying and doing anything that its creator wants [27, 9].

There are, of course, many entertaining and useful applications for such synthesized content – so-called deep fakes. This content, however, can also be weaponized; it can be used to create non-consensual pornography, to instigate small- and large-scale fraud, and to produce dis-information designed to disrupt democratic elections and sow civil unrest.

We describe a forensic technique for detecting a specific class of deep-fake videos. We begin by briefly reviewing previous work on the creation and detection of deep-fake videos before describing our technique in more detail.

Creation: Face-swap deep fakes are the most popular form of deep-fake videos, in which one person’s face in a video is replaced with another person’s face. Most of these face-swap deep fakes are created using a generative adversarial network (GAN), including DeepFake FaceSwap [2], Faceswap-GAN [4], and FS-GAN [21]. Other methods rely on more traditional computer-graphics approaches to create deep fakes. Face2Face [26] and FaceSwap [3], for example, allow for the creation of puppet-master deep fakes in which one person’s facial expressions and head movements are mapped onto another person. Neural Textures [25] is an
image synthesis framework that combines traditional graphics rendering with learnable components to create, among other things, lip-sync deep fakes in which a person’s mouth is modified to be consistent with another person’s speech. This work generalizes earlier work that was designed to create lip-sync deep fakes on a per-individual basis [24].

Unlike each of these previous techniques that require either a visual or auditory imposter, text-based synthesis techniques can modify a video on a per-word basis [13]. This type of deep fake poses even more significant challenges for detection, as only a small change need be made to dramatically alter the meaning of a video.

Detection: There is a significant literature in the general area of digital forensics [12]. Here we focus only on techniques for detecting the types of deep-fake videos, broadly categorized as low-level or high-level approaches.

Low-level forensic techniques detect pixel-level artifacts introduced by the synthesis process. Some of these techniques detect generic artifacts [32, 30, 31, 28], while others detect explicit artifacts that result from, for example, image warping [20], image blending [18] and inconsistencies between the image and metadata [14]. The benefit of low-level approaches is that they can detect artifacts that may not be visibly apparent. The drawback is that they can be sensitive to unintentional laundering (e.g., transcoding or resizing) or intentional adversarial attacks (e.g., [8]). In addition, these approaches are generally more effective in detecting face-swap and puppet-master deep fakes in which the entire face is synthesized or rendered, as opposed to lip-sync deep fakes in which only the mouth region is synthesized.

High-level approaches, in contrast, tend to generalize and be more resilient to laundering and adversarial attacks. These techniques focus on semantically meaningful features including, for example, inconsistencies in eye blinks [19], head-pose [29], physiological signals [11], and distinct mannerisms [6]. As with low-level techniques, these approaches are generally more effective when confronted with face-swap and puppet-master deep fakes in which the entire face is manipulated, but are less effective when confronted with complete mouth and audio synthesis and replacement (lip-sync) [24] and partial word-based audio and mouth synthesis and replacement [13].

Table 1. The number of videos, duration of videos, and total number of visemes MBP for each dataset.

<table>
<thead>
<tr>
<th></th>
<th>video (count)</th>
<th>video (seconds)</th>
<th>MBP (count)</th>
</tr>
</thead>
<tbody>
<tr>
<td>original</td>
<td>79</td>
<td>2,226</td>
<td>1,582</td>
</tr>
<tr>
<td>A2V [24]</td>
<td>111</td>
<td>3,552</td>
<td>2,323</td>
</tr>
<tr>
<td>T2V-L [13]</td>
<td>59</td>
<td>308</td>
<td>166</td>
</tr>
<tr>
<td>T2V-S [13]</td>
<td>24</td>
<td>156</td>
<td>57</td>
</tr>
<tr>
<td>in-the-wild</td>
<td>4</td>
<td>87</td>
<td>66</td>
</tr>
</tbody>
</table>

Overview: We describe a forensic technique for detecting lip-sync deep fakes, focusing on high-level techniques in order to be robust to a range of different synthesis techniques and to be more robust to intentional or unintentional laundering. Our technique exploits the fact that, although lip-sync deep fakes are often highly compelling, the dynamics of the mouth shape – so-called visemes – are occasionally inconsistent with a spoken phoneme. Try, for example, to say a word that begins with M, B, or P – mother, brother, parent – and you will notice that your lips have to completely close. If you are not a ventriloquist, you will have trouble properly enunciating “mother” without closing your lips. We observe that this type of phoneme to viseme mapping is occasionally violated, even if it is not immediately apparent upon casual inspection. We describe how these inconsistencies can be leveraged to detect audio-based and text-based lip-sync deep fakes and evaluate this technique on videos of our creation as well as in-the-wild deep fakes.

2. Methods

Datasets: We analyse lip-sync deep fakes created using three synthesis techniques, Audio-to-Video [24] (A2V) and Text-to-Video [13] in which only short utterances are manipulated (T2V-S), and Text-to-Video in which longer utterances are manipulated (T2V-L). The A2V synthesis technique takes as input a video of a person speaking and a new audio recording, and synthesizes a new video in which the person’s mouth is synchronized with the new audio. The T2V synthesis techniques take as input a video of a person speaking and the desired text to be spoken, and synthesize a new video in which the person’s mouth is synchronized with the new words. The videos in the T2V-S dataset are taken directly from the original publication [13]. The videos in the T2V-L dataset are generated using the implementation of [13] generalized from short to longer utterances. We also apply our analysis to four in-the-wild lip-sync deep fakes downloaded from Instagram and YouTube1.

For each lip-sync video, we also collected, when available, the original video that was used to create the fake. For each video, the face in each frame was localized, aligned, and cropped (to 256 × 256 pixels) using OpenFace [7], and resaved at a frame-rate of 30 fps. Shown in Table 1 are the count and duration (in seconds) of the lip-sync and original videos in our testing dataset.

Phonemes and Visemes: In spoken language, phonemes are perceptually distinct units of sound. A viseme, the visual counterpart of a phoneme, corresponds to the mouth shape needed to enunciate a phoneme. Shown in Figure 1 are a subset of six visemes with their corresponding

1www.instagram.com/bill_posters_uk and youtu.be/VWMDacZ3L4
phonemes (a single viseme may correspond to more than one phoneme) [1].

In order to pronounce chair (CH), jar (JH), or shelf (SH), for example, you need to bring your teeth close together and move your lips forward and round them, causing the teeth to be visible through the open mouth. Whereas, in order to pronounce toy (OY), open (UH), or row (UR), the lips again need to be rounded but the teeth are not brought together and therefore not visible through the open mouth. The phoneme group of M (mother), B (brother), and P (parent), on the other hand, requires the mouth to be completely closed for the pronunciation.

The specific shape of various visemes may depend on other speech characteristics like emphasis or volume. The M, B, P phoneme group (MBP), however, always requires the mouth to be completely closed regardless of other speech characteristics (with the exception of ventriloquists). We focus, therefore, our analysis on this consistent phoneme/viseme mapping.

Extracting Phonemes: In order to analyse a viseme during a spoken MBP phoneme, we first extract the location of all phonemes as follows. Google’s Speech-to-Text API [5] is used to automatically transcribe the audio track associated with a video. The transcription is manually checked to remove any errors and then aligned to the audio using P2FA [23]. This alignment generates a sequence of phonemes along with their start and end time in the input audio/video. Here, only the MBP phonemes will be considered. Shown in the last column of Table 1 are the number of MBP phoneme occurrences extracted for each dataset.

Measuring Visemes (manual): For a given MBP occurrence, the associated viseme is searched in six video frames around the start of the occurrence. We consider multiple frames to adjust for small phoneme to audio alignment errors. Only the frames around the start of the occurrence are analysed because the mouth should be closed before the MBP phoneme sound is made.

Given six frames for an MBP occurrence, we take three approaches to determine if the expected mouth-close viseme is present in any of the frames. The first approach is purely manual where an analyst is presented with six video frames and a reference frame from the same video where the mouth is clearly closed. The analyst is then asked to label each presented sequence as “open” or “closed.” A closed sequence is one in which the mouth is completely closed for at least one video frame. This approach provides the ground-truth for an automatic computational approach to determining if the mouth shape associated with a MBP phoneme is open or closed. This type of manual analysis might also be applicable in one-off, high-stakes analyses.

Measuring Visemes (profile): In the second approach, a mouth-close viseme is automatically detected in any of the six frames centered around an MBP occurrence. For each frame, the lip region is extracted from 68 facial landmarks [17]. The extracted lip region is rescaled to 50 × 50 pixels and converted from RGB to grayscale. A vertical intensity profile is then extracted from the middle of the mouth (Figure 2). We expect this intensity profile to be qualitatively different when the mouth is open or closed. Shown in the top middle panel of Figure 1, for example, is a mouth open in which the vertical intensity profile will change from skin tone to bright (teeth), to dark (the back of the mouth), to bright (teeth), and then back to skin tone. In contrast shown in the top right panel of Figure 1, is a mouth closed in which the vertical intensity will be largely uniform skin tone.

The overall profile shape is quantified by computing the sum of the prominences of the local minima, l, and maxima, h, in the intensity profile (as determined using MATLAB’s findpeaks function, with the default parameters). Figure 2. The measurements l and h capture how much the intensity along the profile decreases (e.g., when the back of the mouth is visible) and increases (e.g., when the teeth are visible). These measurements are made for each of the six frames, l_i and h_i, i ∈ [1, 6], and compared to the reference measurements l_r and h_r, in which the mouth is closed, Figure 3. The measure of similarity to a reference frame in the six-frame sequence is the minimum of |l_i - l_r| + |h_i - h_r|, i ∈ [1, 6].

Measuring Visemes (CNN): In a third approach, we explored if a more modern learning-based approach can outperform the hand-crafted profile feature. Specifically, we trained a convolutional neural network (CNN) to classify if a mouth is open or closed in a single video frame. The input to the network is a color image cropped around the mouth and rescaled to a 128 × 128 pixels (Figure 1). The output, c_i, of the network is real-valued number in [0, 1] corresponding to an “open” (0) or “closed” (1) mouth. The open/closed classification in a six-frame sequence is the maximum of c_i, i ∈ [1, 6].
Figure 3. Six sequential frames extracted from a single MBP occurrence in different deep-fake videos. Shown on the right is a reference frame where the mouth is clearly closed. Shown below each frame is a 1-D intensity profile used to automatically classify the mouth as open or close. The bounding box corresponds to a frame that matched the reference frame shown to the right (only the closed-mouth sequences match).
The network is trained using videos of Barack Obama for whom the lip-sync deep fakes were created in the A2V dataset. This training dataset consists of original videos disjoint from the testing videos reported in Table 1. In total, we manually labelled 15,600 video frames where the mouth is open (8258 instances) or closed (7342 instances). In each frame, OpenFace [7] is used to automatically detect, scale, and rotate (in-plane) the face to a normalized pose and resolution.

The Xception architecture [10] is used to train a classifier using 90%/10% images for training/validation. The network is trained for 50,000 iterations with a mini-batch of size 64. In each mini-batch, equal number of images were randomly sampled from each label. The initial learning rate of 0.01 was reduced twice at iterations 20,000 and 40,000. The weights were optimized using Adam optimizer and a cross-entropy loss function.

Global Audio-to-Video Alignment: We previously used P2FA to ensure that the phonemes were correctly synchronized with the underlying audio. Here we also ensure that the audio is correctly synchronized with the underlying video. This audio-to-video alignment is done through a brute-force search of the global shift in the audio (in the range $[-1, 1]$ seconds, in steps of $1/10$ seconds) that creates the best agreement between all MBP phonemes and the correct mouth-closed viseme. This alignment contends with slight audio to video desynchronization that might occur from transcoding or innocuous video editing.

3. Results

Detecting Deep Fakes (manually): We evaluate the efficacy of detecting deep fakes first by using the manual annotation for determining if the phoneme and viseme pairing is correct. Shown in Figure 4 are the percent of MBP phoneme occurrences where the correct viseme is observed. For each dataset, the percent is reported before (blue) and after (orange) the global audio to video alignment. The problem of misalignment is most salient for in-the-wild videos where before alignment only 45.5% of the visemes were correct, as compared to 90.9% after alignment. For each of the other datasets, misalignment was not an issue.

For the four deep-fake data sets (A2V, T2V-S, T2V-L, in-the-wild), the percentage of correct phoneme to viseme pairing (after alignment) ranges from a high of 90.9% of 66 occurrences (in-the-wild), to 76.8% of 2,323 occurrences (A2V), and 70.2% of 57 occurrences (T2V-S), and 18.7% of 166 occurrences (T2V-L). The phoneme to viseme pairing in original videos is correct for 99.7% of 1,582 occurrences (the small number of errors are due either to manual annotation or transcription error).

Shown in Table 2 is the rate (per second) at which MBP phonemes occur (total column) and the rate at which phoneme-viseme mismatches occur (incorrect column). The rate of spoken MBP phonemes varies from 0.71 (original) to 0.37 (T2V-S), and so it is important to compare to the appropriate base rate when considering overall accuracy.

Even a relatively low number of say 10% incorrect phoneme to viseme pairings can, over time, lead to an effective detection strategy. In particular, shown in the left-most panel of Figure 5 is the percent of videos that are correctly identified as fake as a function of video duration, from 1 to 30 seconds. A video is detected as fake if the number of incorrect phoneme to viseme mismatches exceeds the expected number.

<table>
<thead>
<tr>
<th>dataset</th>
<th>correct</th>
<th>incorrect</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>original</td>
<td>0.709</td>
<td>0.001</td>
<td>0.710</td>
</tr>
<tr>
<td>A2V</td>
<td>0.49</td>
<td>0.15</td>
<td>0.64</td>
</tr>
<tr>
<td>T2V-L</td>
<td>0.09</td>
<td>0.43</td>
<td>0.52</td>
</tr>
<tr>
<td>T2V-S</td>
<td>0.26</td>
<td>0.11</td>
<td>0.37</td>
</tr>
<tr>
<td>in-the-wild</td>
<td>0.64</td>
<td>0.05</td>
<td>0.69</td>
</tr>
</tbody>
</table>

Table 2. The average number of correct, incorrect, and total viseme occurrences/second of video.

<table>
<thead>
<tr>
<th>dataset</th>
<th>profile</th>
<th>CNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>original</td>
<td>99.4%</td>
<td>99.6%</td>
</tr>
<tr>
<td>A2V</td>
<td>96.6%</td>
<td>96.9%</td>
</tr>
<tr>
<td>T2V-L</td>
<td>83.7%</td>
<td>71.1%</td>
</tr>
<tr>
<td>T2V-S</td>
<td>89.5%</td>
<td>80.7%</td>
</tr>
<tr>
<td>in-the-wild</td>
<td>93.9%</td>
<td>97.0%</td>
</tr>
</tbody>
</table>

Table 3. The accuracy of the two automatic techniques (profile and CNN) to detect if a mouth is open or closed. The accuracies are computed at a fixed threshold corresponding to average false alarm rate of 0.5% (i.e., misclassifying a closed mouth as open).
is the accuracy of the two automatic detection techniques against two simple laundering operations, recompression and resizing. Each video was laundered using ffmpeg by: (1) reencoding at a lower quality of \( \text{qp}=40 \) (typical videos are encoded at higher quality of \( \text{qp} \in [10, 20] \)); or (2) resizing to half-resolution and scaling back to the original resolution (effectively, blurring each video frame). The average accuracy of the profile and CNN technique in detecting open or closed mouth after recompression is 90.46% and 88.32%. The average accuracy of the profile and CNN technique after resizing is 83.80% and 89.92%.

Resizing has a significant impact on accuracy for the profile technique. This is because resizing reduces the prominence of the local minima and maxima. As a result, the open mouth are more likely to be mis-classified as closed.

For the CNN technique, the video detection accuracy for the original and A2V datasets remains comparable to the manual and profile annotations: at 30 seconds, the accuracy on the original and A2V datasets is 93.4% and 97.8%. For the T2V-S dataset, however, the accuracy drops from 97.4% to 81.0%. This is because the CNN was trained only on videos of Barack Obama exclusively in the A2V dataset, and thus does not generalize well to different people in the T2V-S dataset. We hypothesize that this accuracy can be improved by training a CNN with different people.

**Failures:** Shown in Figure 7 are two six-frame sequences where the profile technique misclassified a closed mouth as open (top) and an open mouth as closed (bottom). The first failure is because the shape of the lips is different from the reference frame. The second failure is because the mouth is asymmetrically open. While these failure cases are somewhat inevitable when using automatic techniques, they are easily flagged by a manual annotator.

**Robustness:** We next examine the robustness of the two automatic detection techniques against two simple laundering operations, recompression and resizing. Each video was laundered using ffmpeg by: (1) reencoding at a lower quality of \( \text{qp}=40 \) (typical videos are encoded at higher quality of \( \text{qp} \in [10, 20] \)); or (2) resizing to half-resolution and scaling back to the original resolution (effectively, blurring each video frame). The average accuracy of the profile and CNN technique in detecting open or closed mouth after recompression is 90.46% and 88.32%. The average accuracy of the profile and CNN technique after resizing is 83.80% and 89.92%.

Resizing has a significant impact on accuracy for the profile technique. This is because resizing reduces the prominence of the local minima and maxima. As a result, the open mouth are more likely to be mis-classified as closed.

For the CNN technique, the video detection accuracy for the original and A2V datasets remains comparable to the manual and profile annotations: at 30 seconds, the accuracy on the original and A2V datasets is 93.4% and 97.8%. For the T2V-S dataset, however, the accuracy drops from 97.4% to 81.0%. This is because the CNN was trained only on videos of Barack Obama exclusively in the A2V dataset, and thus does not generalize well to different people in the T2V-S dataset. We hypothesize that this accuracy can be improved by training a CNN with different people.
Figure 6. Shown is a closed (top) and open (bottom) mouth before (first column) and after recompression (second column) and after resizing (third column). Although our automatic techniques correctly classified the closed-mouth, they misclassified as closed the recompressed and resized open mouth. A human analyst can, however, still identify the small opening between the lips even after recompression or resizing.

For such low quality videos, therefore, manual annotation can be more robust than the automatic detection (Figure 6).

4. Discussion

We described a forensic technique that uses phoneme-viseme mismatches to detect deep-fake videos. Our main insight is that while many visemes can vary, the sounds associated with the M, B, and P phonemes require complete mouth closure, which is often not synthesized correctly in deep-fake videos. For high-stakes cases, we show that an analyst can manually verify video authenticity. For large-scale applications, we show the efficacy of two automatic approaches: one using hand-crafted features that requires no large training data, and one using a CNN.

While we had good reason to look only at MBP phonemes, we believe that including all visemes in the analysis will improve results even further. This extension, however, is not trivial and will require modeling the possible variance of each viseme and co-articulation. It will, however, allow us to use a larger portion of a video for analysis, ultimately leading to better detection.

Our CNN results, trained only on videos of Barack Obama, are person specific and perform much better on videos of Obama. We expect better results using a network that is trained on a large corpus of people. Obtaining such a large labelled dataset is challenging — especially since we care mostly about the hard cases in which a mouth is almost closed or open, with just a few pixel difference. Such labels currently cannot be accurately extracted from face landmark detectors. Thus, it would be beneficial to develop unsupervised methods to automatically differentiate between complete and almost complete mouth closure.

Even with these limitations, our method can already detect state-of-the-art, lip-sync deep fakes. We expect future synthesis techniques to continue the cat-and-mouse game, taking into more careful account the phoneme to viseme matching. We view deep-fake detection using phoneme-viseme mismatches as one more tool in the forensic expert toolkit, to be developed and used together with other complementary techniques.

Acknowledgement

This work was partially supported by the Brown Institute for Media Innovation (Fried and Agrawala). This research was developed with funding from Facebook, Google, and from the Defense Advanced Research Projects Agency (DARPA FA8750-16-C-0166). The views, opinions, and findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government (Agarwal and Farid).

References

Figure 7. Shown are two six-frame sequences where the automatic profile technique failed to correctly classify the mouth as open or closed. The mouth in the upper sequence was incorrectly classified as open, whereas in the lower sequence, the mouth was incorrectly classified as closed. Shown on the far right is the reference frame and shown below each frame is the intensity profile used for classification.


[29] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *IEEE International Con-
