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.

Figure 1. Six example visemes and their corresponding phonemes. The phonemes in the top-right (M, B, P), for example, correspond to the sound you make when you say "mother", "brother", or "parent". To make this sound, you must tightly press your lips together, leading to the shown viseme.

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

	video (count)	video (seconds)	MBP (count)
original	79	2; 226	1; 582
A2V [24]	111	3; 552	2; 323
T2V-L [13]	59	308	166
T2V-S [13]	24	156	57
in-the-wild	4	87	66

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

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 most 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].

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 YouTube[1].

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

[1] www.instagram.com/bill_posters_uk and youtu.be/VWMEDacz3L4

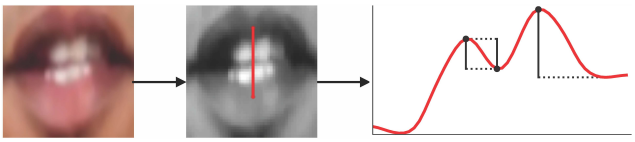

Figure 2. Overview of the profile feature extraction used to measure the mouth-closed viseme. The input image is first converted to grayscale and a vertical intensity profile is extracted from the center of the mouth. Shown on the right is the intensity profile with the location of local minima and maxima (black dots) and their corresponding prominences measured as the height, denoted by the dashed horizontal lines, relative to a neighboring minima/maxima.

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 (UW), 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 measuremtns are made for each of the six frames, l_i and h_i; i 2 [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 $(jl_i \quad l_r j + jh_i \quad h_r j)$; i 2 [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, 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 2 [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).

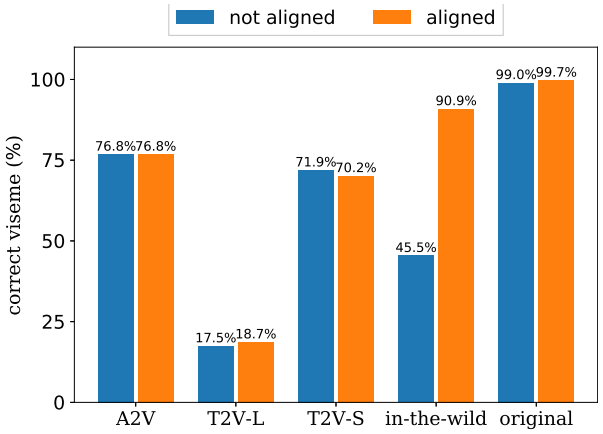

Figure 4. The number of correct MBP phoneme to viseme pairings before (blue) and after (orange) audio to video alignment. The T2V-L lip-sync deep fakes are the least well matched, while the (aligned) in-the-wild deep fakes are correctly matched more than 90% of the time.

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.

dataset	correct	incorrect	total
original	0:709	0:001	0:710
A2V	0:49	0:15	0:64
T2V-L	0:09	0:43	0:52
T2V-S	0:26	0:11	0:37
in-the-wild	0:64	0:05	0:69

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

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 ex-

dataset	profile	CNN
original	99:4%	99:6%
A2V	96:6%	96:9%
T2V-L	83:7%	71:1%
T2V-S	89:5%	80:7%
in-the-wild	93:9%	97:0%

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).

Figure 5. Shown in each panel is the accuracy with which lip-sync deep fakes are detected using mismatched phoneme to viseme pairings. Each solid curve (orange, green, red, and purple) corresponds to a different deep-fake dataset and the dashed curve (blue) corresponds to the original dataset. Each panel corresponds to a different technique for determining if a mouth is open or closed. Detection accuracy improves steadily as the length of the video increases from 1 to 30 seconds.

pected mismatch of 6.3% found in original video (Figure 4. As expected, the detection accuracy increases as the video length increases. At a length of 30 seconds, for example, nearly all of the A2V, T2V-L, and T2V-S videos are classified correctly, while only 4% of original videos are misclassified.

Detecting Deep Fakes (automatically): We next evaluate the accuracy of automatically determining if a mouth is open or closed and how these automatic classifications impact the accuracy of detecting a video as real or fake. Throughout, the manual annotation described above are used as ground truth.

Shown in Table 3 is the accuracy of the two automatic techniques (pro le and CNN) to detect if a mouth is open or closed. Each classifier was configured to have an average false alarm rate of 0.5% (i.e., misclassifying a closed mouth as open). The performance of both the pro le and CNN techniques are high on the A2V dataset with an average accuracy above 96%. On the T2V-L and T2V-S datasets, however, the pro le technique performs better than the CNN which was only trained on videos of Barack Obama (somewhat surprisingly, however, the CNN generalizes to the in-the-wild videos).

Shown in the central and right-most panel of Figure 5 is the video detection accuracy when the manual annotation of mouth open or closed is replaced with the automatic detection based on intensity pro les (center) and CNN classification (right). Using the pro le technique, the video detection accuracy is only slightly degraded as compared to the manual annotation (left-most panel): at 30 seconds, for example, the manual annotation has an accuracy on the original, A2V, and T2V-S datasets of 96.0%, 97.8%, and 97.4%, as compared to the automatic pro le technique with an accuracy of 93.4%, 97.0%, and 92.8%.

For the CNN technique, the video detection accuracy for the original and A2V datasets remains comparable to the manual and pro le annotations: at 30 seconds, the accuracy on the original and A2V datasets is 98.4% 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 pro le 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 qp=40 (typical videos are encoded at higher quality of qp 2 [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 pro le and CNN technique in detecting open or closed mouth after recompression is 90.46% and 88.32%. The average accuracy of the pro le and CNN technique after resizing is 88.80% and 89.92%.

Resizing has a significant impact on accuracy for the pro le 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.

Figure 6. Shown is a closed (top) and open (bottom) mouth before (rst column) and after recompression (second column) and after resizing (third column). Although our automatic techniques correctly classi ed the closed-mouth, they misclassi ed 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 ef cacy 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 M, B, P 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 speci c 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 bene cial 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 ndings expressed are those of the authors and should not be interpreted as representing the of cial views or policies of the Department of Defense or the U.S. Government (Agarwal and Farid).

References

[1] Annosoft lipsync tool. http://www.annosoft.com/docs/Visemes17.html . 3

[2] Deepfakes faceswap. https://github.com/deepfakes/faceswap . 1

[3] Faceswap. https://github.com/MarekKowalski/FaceSwap/ . 1

[4] Faceswap-GAN. https://github.com/shaoanlu/faceswap-GAN . 1

[5] Google speech-to-text. https://cloud.google.com/speech-to-text/docs . 3

[6] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In IEEE Conference on Computer Vision and Pattern Recognition, Workshop on Media Forensics, pages 38–45, 2019. 2

[7] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In IEEE Winter Conference on Applications of Computer Vision, pages 1–10, 2016. 2, 5

[8] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. arXiv: 1608.04644, 2016. 2

[9] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In IEEE International Conference on Computer Vision, 2019. 1

[10] François Chollet. Xception: Deep learning with depthwise separable convolutions. In IEEE Conference on Computer Vision and Pattern Recognition, 2017. 5

[11] Umur Aybars Ciftci and Ilke Demir. Fakecatcher: Detection of synthetic portrait videos using biological signals. arXiv: 1901.02212, 2019. 2

[12] Hany Farid. Photo Forensics. MIT Press, 2016. 2

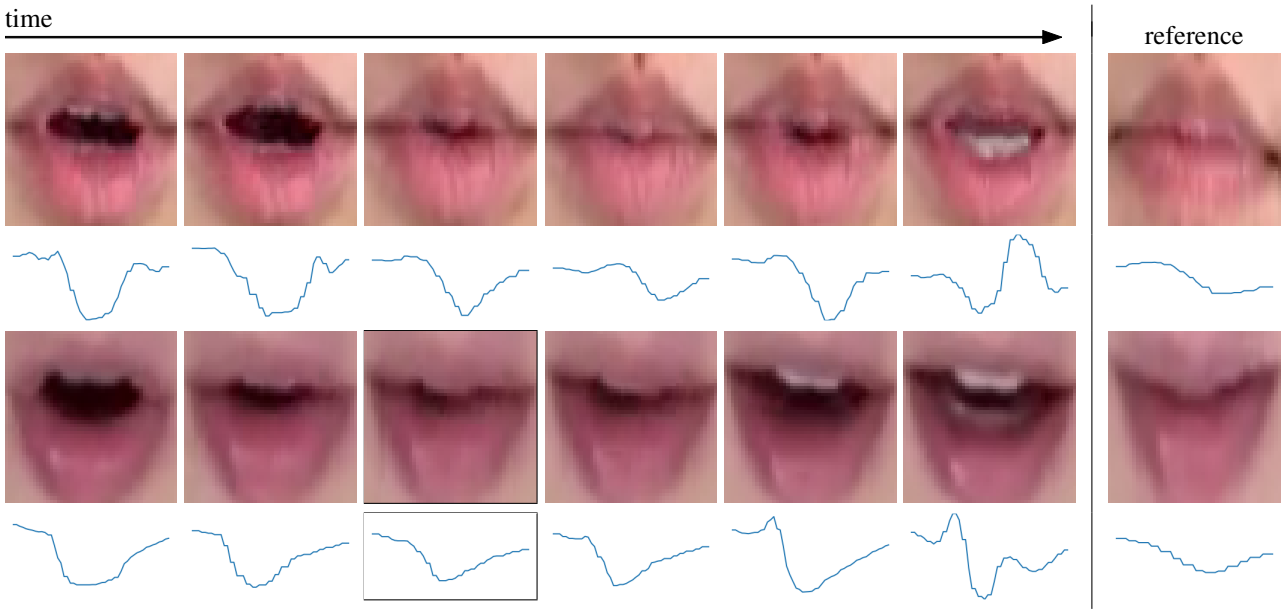

Figure 7. Shown are two six-frame sequences where the automatic profile technique failed to correctly classify the mouth as open or closed. The mouth in the upper sequence was incorrectly classified as open, whereas in the lower sequence, the mouth was incorrectly classified as closed. Shown on the far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

[13] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. Text-based editing of talking-head video. *ACM Transactions on Graphics*, 2019. 2

[14] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A. Efros. Fighting fake news: Image splice detection via learned self-consistency. In *European Conference on Computer Vision*, 2018. 2

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1

[16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. 2019. 1

[17] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 2009. 3

[18] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face X-ray for more general face forgery detection. *arXiv preprint arXiv:1912.13458*, 2019. 2

[19] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing AI created fake videos by detecting eye blinking. In *IEEE International Workshop on Information Forensics and Security*, pages 1–7, 2018. 2

[20] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. arXiv: 1811.00656, 2018. 2

[21] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: Subject agnostic face swapping and reenactment. arXiv: 1908.05932, 2019. 1

[22] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. 2016. 1

[23] Steve Rubin, Floraine Berthouzoz, Gautham J Mysore, Wilmot Li, and Maneesh Agrawala. Content-based tools for editing audio stories. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, 2013. 3

[24] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing Obama: Learning lip sync from audio. *ACM Transactions on Graphics*, 2017. 2

[25] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred Neural Rendering: Image Synthesis using Neural Textures. arXiv: 1904.12356v1, 2019. 1

[26] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1

[27] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. 2020. 1

[28] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. CNN-generated images are surprisingly easy to spot...for now. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

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

ference on Acoustics, Speech and Signal Processing, pages 8261–8265, 2019. 2

[30] Ning Yu, Larry Davis, and Mario Fritz. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints. In *IEEE International Conference on Computer Vision*, 2018. 2

[31] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and Simulating Artifacts in GAN Fake Images. arxiv: 1907.06515, 2019. 2

[32] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Two-stream neural networks for tampered face detection. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017. 2