Lost in Translation: Lip-Sync Deepfake Detection from Audio-Video Mismatch

Matyas Bohacek
Stanford University
Stanford, CA USA
maty@stanford.edu

Hany Farid
University of California, Berkeley
Berkeley, CA USA
hfarid@berkeley.edu

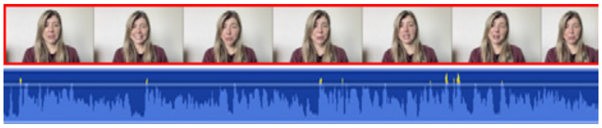
Abstract

Highly realistic voice cloning combined with AI-powered video manipulation allows for the creation of compelling lip-sync deepfakes where anyone can be made to say things they never did. The resulting fakes are being used to entertain, but also for everything from election related disinformation to small- and large-scale fraud. Lip-sync deepfakes can be particularly difficult to detect because only the mouth and jaw of the person talking is modified. We describe a robust and general-purpose technique to detect these fakes. This technique begins by independently translating the audio (using audio-to-text transcription) and video (using automated lip-reading). We then show that the resulting transcriptions are significantly mismatched for lip-sync deepfakes as compared to authentic videos. The robustness of this technique is evaluated against a controlled dataset of our creation and in-the-wild fakes, all of varying length and resolution.

1. Introduction

First it was Instagram video ads of Tom Hanks promoting dental plans. Then it was TV personality Gayle King hawking a sketchy weight-loss plan. Next, Elon Musk was shilling for the latest crypto scam, and, more recently, Taylor Swift was announcing a giveaway of Le Creuset cookware. All, of course, were fake.

Each of these lip-sync deepfake scams are powered by two separate technologies. First, a celebrity's voice is cloned from authentic recordings. Where it used to take hours of audio to convincingly clone a person's voice, today it takes a few minutes of authentic recording [13, 21]. Once the voice is cloned, an audio file is generated from a simple text prompt in a process called text-to-speech. And, there are several inexpensive commercial offerings that make cloning a voice easy with minimal cost (e.g., https://elevenlabs.io).



video transcription: I just had its bread roll it's your presence about the media in a way

audio transcription: I just think it's really feel good and excellent piece of cinema

manual transcription: I just think it's really feel-good and an excellent piece of cinema

Figure 1. An audio/video clip from a lip-sync deepfake in which the participant responds to the question "what is your favorite movie and why?" The mismatch between the video (lip reading) and audio transcriptions reveals evidence of a lip-sync deepfake.

Once a voice has been created, an original video is modified to make the celebrity's mouth region move consistently with the new audio [6, 17, 22]. Tools for this video generation are now readily available online for free or for a nominal cost. These lip-sync deepfakes add to a growing list of sophisticated techniques for creating fake videos [10].

Although the resulting fakes are not yet perfect, they are reasonably convincing, particularly when being viewed on a small mobile screen. The power of these types of video-powered scams is that they can fail 99% of the time and still be highly lucrative for scam artists. More than any other nefarious use of lip-sync deepfakes, it is these types of frauds and scams that seem to have recently gained the most traction.

While a few years ago, lip-sync deepfakes were less concerning, recent advances in voice cloning have catapulted the threat of lip-sync deepfakes. A particularly challenging aspect of these fakes is that unlike other deepfakes, only the mouth region in the video needs to be altered, making detection more difficult.

This paper was written entirely by the authors without the use of ChatGPT. However, based on the abstract, the first three words of the title were suggested by chatGPT4.

While many techniques have attempted to detect lip-sync deepfakes by learning specific synthesis artifacts [11, 25, 27], these learning-based approaches suffer from the usual problems of requiring large amounts of data, struggling with out-of-domain or laundered videos, and a lack of explainablity [12]. In comparison, we describe a robust and explainable technique that leverages semantic-level differences in the underlying audio and video feeds. In particular, we show that a comparison of audio-to-text and video-to-text (lip reading) transcription results in significant differences for lip-sync deepfakes as compared to authentic videos. This technique leverages the fact that lip-sync deepfakes create mouth shapes that are at times inconsistent with the underlying audio. At the same time, our visual system can sometimes be impervious to these inconsistencies.

After framing our work in context to related efforts that analyze lip movements, we describe the collection of two datasets, one in a controlled setting and one of CNN's Anderson Cooper. We evaluate our technique against both of these datasets as well as several in-the-wild deepfakes. We also evaluate the robustness of our technique to video length and resolution.

1.1. Related Work

One of the earliest examples of exploiting mismatches between the audio signal and the mouth shape compared the shape of the mouth (viseme) with a specific phoneme [2]. When uttering a `M, B, P` phoneme, for example, the mouth has to close. In this earlier work, the authors isolate all instances of the `M, B, P` phoneme in the audio signal and then isolate the six corresponding video frames for analysis of the expected closed mouth shape. The strength of this approach is that it focuses on an explainable semantic-level feature. The drawback is that it relies on relatively short video snippets from which it is difficult to accurately measure the shape of the mouth.

Building on this earlier work the authors in [3] exploit a person-specific association between specific words and a speaker's head movements and facial gestures. When, for example, at the onset of President Obama saying "Hi everybody," he tends to tilt his head upwards. The strength of this approach is that it can learn specific characteristics that can be difficult to mimic in a deepfake. The drawback is this approach requires a large amount of data and is only applicable to individuals with a large digital footprint.

These two earlier approaches exploit the correlation between the shape of the mouth, expressions, and the spoken word. Although a little less obvious, correlations have also been shown between movement of the mouth and ear [1]. The strength of exploiting this relationship is that lip-sync (and face-swap) deepfakes do not (yet) synthesize the ears so this part of the head provides a useful feature. The drawback is that this approach requires the head to be rotated somewhat so that one of the ears is visible and the dynamics of the ear are fairly subtle, making measurement difficult in low-resolution video.

Moving beyond just the shape of the mouth, the authors in [5] estimate from an audio the anatomical arrangement of the human vocal tract during speech. The authors find that AI-generated audio yield impossible or highly-unlikely anatomical arrangements. The strength of this approach is that it exploits a detailed model of the vocal tract that today's generative AI cannot easily mimic. The drawback is that this approach does not consider the underlying video which can provide important information as to authenticity.

In contrast to – but also building on – these earlier approaches we propose a technique that takes full advantage of the entire audio and video channel, is applicable regardless of the identity in the video, does not require any training data, and is robust to video length and resolution, making it applicable to a broad range of in-the-wild lip-sync deepfakes.

2. Creation

2.1. Datasets

We created and collected three datasets for the purposes of evaluation: (1) a controlled dataset of 10 individuals responding to questions and reading a script; each video was of variable length and trimmed to 5 minutes in length; (2) a total of 36 minutes of footage from CNN's Anderson Cooper consisting of 64 videos each of length 30 seconds; and (3) 14 in-the-wild lip-sync deepfakes ranging in length from 5 to 122 seconds, for a total of 440 seconds.

For the controlled dataset, 10 volunteers were asked to record themselves from their webcam responding to 18 free-form questions of the form "What did you have for breakfast today?", "Describe something that is in your line of sight?", "If you could time travel, where in the past or future would you like to go?", and "If you could have any superpower, what would it be?" Between each block of three questions, the volunteer was asked to look into the camera and silently count to 10. Each volunteer was then asked to read a series of phonetically rich sentences of the form "The birch canoe slid on the smooth planks," "glue the sheet to the dark blue background," and "It's easy to tell the depth of a well."

Using InsightFace [9] to localize facial landmarks, each video was resized so that the horizontal interpupillary distance (IPD) was, on average, 128 pixels. Because a person tends to move throughout the video, we could not fix the resolution throughout the entire video while preserving natural head movement. Instead, we adjusted the video resolution from the IPD extracted from the first 10 video frames.

For the Anderson Cooper dataset, we downloaded 57 videos from CNN's YouTube channel. The videos were filtered using Deepface [20] to extract 64 clips of length 30 seconds each in which Cooper is the only person in the frame and the only one talking.

For the in-the-wild deepfakes, we collected 14 verified deepfakes from a variety of online sites. These include videos of Joe Biden, Alexandria Ocasio-Cortez (2×), Ron DeSantis, Drake, Kamala Harris, Kim Kardashian, Kari Lake, Barack Obama, Queen Elizabeth, Taylor Swift, Donald Trump, Elizabeth Warren, and Ye.

2.2. Deepfakes

We created lip-sync deepfakes for the controlled and Anderson Cooper datasets using VideoRetalking [7] and Wav2Lip [17], two widely used deepfake engines. With a video and an audio file as input, these synthesis engines render a new mouth area in the video to match the provided audio without any person-specific fine-tuning or need for additional reference data.

For the controlled dataset, we created a total of 90 fakes, combining each of the 10 real videos with the remaining 9 original audios in the dataset. For the Anderson Cooper dataset, we created a total of 640 fakes, combining each of the 64 real videos with 10 different audios of Cooper.

We also created face-swap deepfakes using Facefusion[1]. Unlike lip-sync deepfakes which only modify the mouth region, face-swap deepfakes replace an entire face from eyebrows to chin and cheek to cheek with a new identity.

3. Detection

Audio transcription for the footage in question is extracted using Whisper's English-only pipeline [18]. This pipeline constructs a log-mel spectrogram of the audio channel, which is then inserted into a sequence-to-sequence Transformer, yielding a time-stamped transcript. Whisper can also identify and isolate multiple speakers and background noise. Although in our case this feature was not needed because each video contains only a single speaker, these additional features could be used to automatically analyze videos with multiple speakers.

Video transcription is extracted using Auto-AVSR's lip-reading pipeline [14]. First, the pipeline finds the largest mouth region in the video, defines a bounding box around the mouth, and isolates the mouth in each video frame. These frames (converted to grayscale) are then analyzed by an encoder-decoder model trained on thousands of hours of English-speaker footage. The model predicts timestamps of discrete words in the video. Finally, spoken words are predicted based on the mouth shape and a context window of adjacent words. Auto-AVSR limits the length of the video input so we built a wrapper that partitions a video into shorter clips and stitches the output of Auto-AVSR on each clip into a final transcribed video.

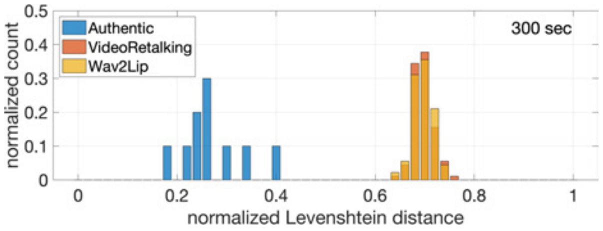


Figure 2. The distance between audio and video transcriptions for full length and resolution authentic and VideoRetalking and Wav2Lip lip-sync videos. A larger value corresponds to larger mismatches. See also Figures 3 and 4.

The difference between the audio and video transcription is quantified using the normalized Levenshtein distance. The standard Levenshtein distance [16] between two strings is computed as the minimum number of single-character insertions, deletions or substitutions required to change one string into the other. The normalized Levenshtein distance [26] converts the standard Levenshtein distance into a proper distance metric bounded into the range $[0, 1]$, with a smaller value corresponding to a better match between the video and audio transcriptions.

As described next, we distinguish authentic from lip-sync deepfake videos using a simple threshold on this normalized Levenshtein distance. For simplicity, we choose a single threshold regardless of video length, resolution, or quality. In practice, however, a more dynamic threshold might be warranted.

4. Results

For the 10 authentic videos in our controlled dataset, (Section 2.1), the median normalized Levenshtein distance is 0.26 with a minimum distance of 0.19 and a maximum distance of 0.40, Figure 2 (as described in the previous section, all distances are normalized into a range of $[0, 1]$)

For the 90 VideoRetalking lip-sync videos, the median normalized Levenshtein distance is 0.69 with a minimum distance of 0.65 and a maximum distance of 0.76, Figure 2. For the full-length videos at the original resolution, there is a significant and perfect separation between the authentic and lip-sync videos with the smallest lip-sync distance 1.6 times larger than the largest authentic distance.

The results for Wav2Lip are similar. For the 90 Wav2Lip videos, the median normalized Levenshtein distance is 0.69 with a minimum distance of 0.65 and a maximum distance of 0.74, Figure 2. For the full-length videos at the original resolution, there is a significant and perfect separation between the authentic and lip-sync videos with the smallest lip-sync distance 1.6 times larger than the largest authentic

[1] https://github.com/facefusion/facefusion

distance.

These clear differences are because the video transcriptions of the lip-sync videos can deviate significantly from the underlying audio, even if this is not always visually apparent. This is particularly true over the relatively long 5-minute videos.

Below are a few representative examples of audio and video transcriptions and mismatches:

manual: I make my breakfast and then I take my tea and breakfast to my favorite place to sit and read the news while I drink tea and eat breakfast

audio: I make my breakfast then I take my tea and breakfast to my favorite place to sit and read the news while I drink tea and eat breakfast

authentic video: I make up at breakfast and then I take my tea at breakfast and my favorite place to sit and with the news while I can try tea at breakfast

lip-sync video: things I used to do as a basket I used to do as a basket I used to do as a basket I used to do as a basket

manual: doing it on Alice would be really good she just is such an interesting person and she has a lot to talk about

audio: doing it on Alice would be really good she just is such an interesting person and she has a to talk about

authentic video: would it be Alex woman relate she's just such an interesting person she has a lot to talk about

lip-sync video: I'm not going to lie I'm not going to lie I'm not going to lie I'm not going to lie

manual: If you could time travel where in the past or the future would you like to go?

audio: If you could time travel where in the past or the future would you like to go?

authentic video: If you could time travel where the best or the future would you like to go?

lip-sync video: I hope you guys enjoy this video if you like this video please subscribe to my channel thank you for watching

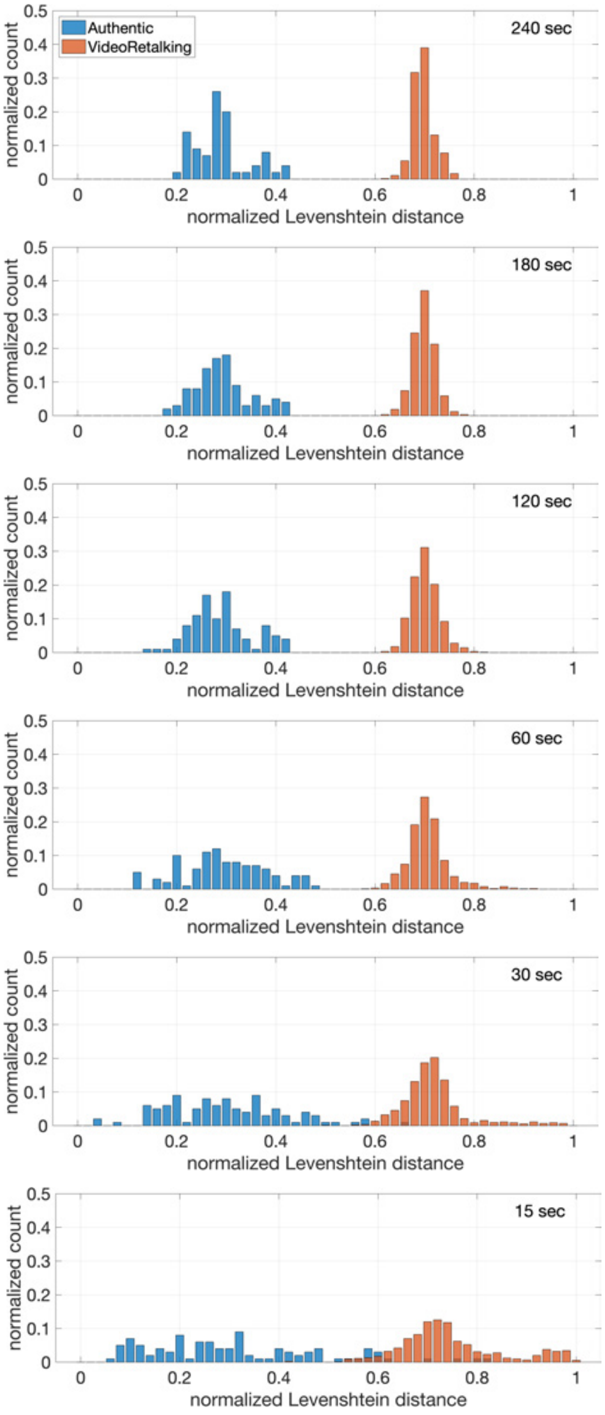


Figure 3. The distance between audio and video transcriptions for authentic and VideoRetalking lip-sync videos of varying length. A larger distance corresponds to a larger mismatch. For videos 60 seconds or longer a distance threshold of 0.5 perfectly separates the authentic from the fake. See also Figure 2.

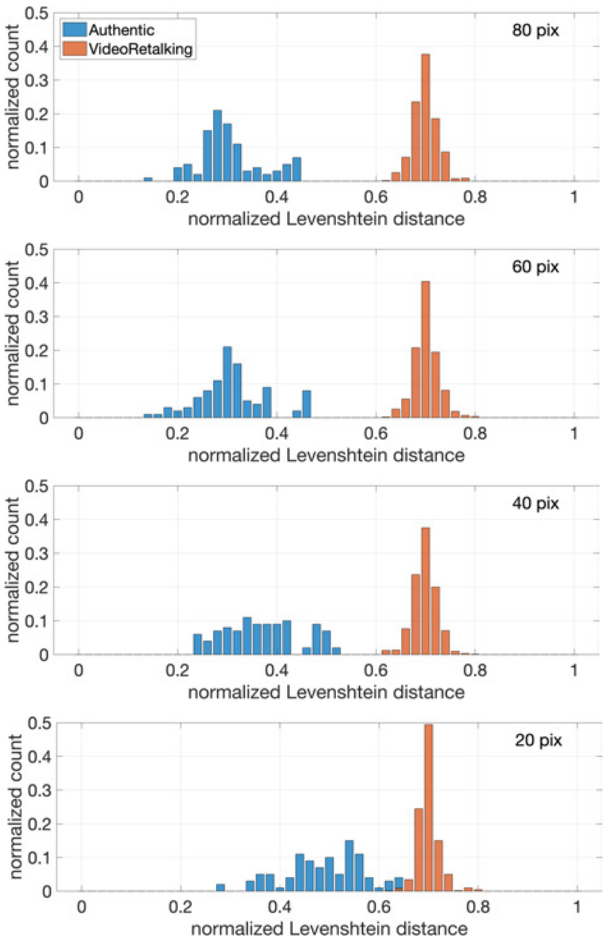


Figure 4. The distance between audio and video transcriptions for authentic and VideoRetalking lip-sync videos of length 60 seconds and of varying resolution (as measured by the speaker's IPD). A larger distance corresponds to a larger mismatch. For videos with an IPD of 60 pixels or larger, a distance threshold of 0.5 perfectly separates the authentic from the fake. See also Figure 2.

4.1. Robustness

Like all forensic techniques, we must consider robustness to a number of real-world factors. Here we consider robustness to the video length and resolution.

The length of each video in our controlled dataset is 5 minutes (300 seconds). For each authentic video and each lip-sync deepfake video in this dataset, we extracted 10 random contiguous clips each of length 240, 180, 120, 60, 30, or 15 seconds, yielding a total of 100 authentic and 900 lip-sync video clips. Shown in Figure 3 is the distribution of normalized Levenshtein distances for the authentic and lip-sync videos. Even at 60 seconds in length – with a distance threshold of 0.5 – the separation between authentic and lip-sync is perfect.

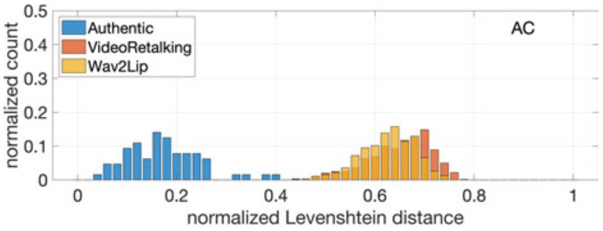


Figure 5. The distance between audio and video transcriptions for authentic and VideoRetalking and Wav2Lip lip-sync videos of Anderson Cooper. A larger value corresponds to larger mismatches. All of the authentic videos fall below a threshold of 0.5, with 97.5% of the VideoRetalking and 98.3% Wav2Lip videos above this threshold.

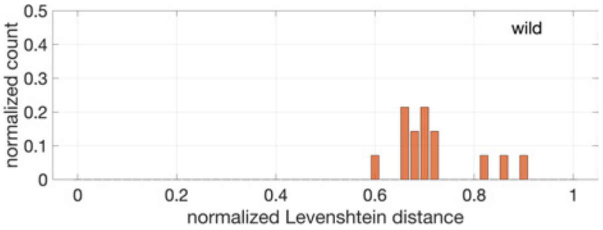


Figure 6. The distance between audio and video transcriptions for 14 in-the-wild lip-sync deepfakes (see Figure 7). A larger value corresponds to larger mismatches. See also Figure 5.

For the video clips of length 30 seconds, 95.0% of authentic videos fall below the 0.5 distance threshold with 99.9% of the lip-sync above this threshold. For videos length 15 seconds, 85% of authentic videos fall below the 0.5 distance threshold with 99.8% of the lip-sync above this threshold. These results suggest that a one-minute video is sufficient to achieve accurate results. We posit that the video transcription worsens with shorter videos due to a reduction in the context of adjacent words (see Section 3).

Next, each video clip of length 60 seconds was resized so that the speaker's IPD is 80, 60, 40, or 20 pixels (the IPD in the original videos is 128 pixels). Shown in Figure 4 is the distribution of normalized Levenshtein distances for the authentic and lip-sync videos. At the original resolution of 128 pixels and a resolution of 80 and 60 pixels – with a distance threshold of 0.5 – the separation between authentic and lip-sync is perfect.

At a resolution of 40 pixels, 94% of authentic videos fall below the 0.5 distance threshold with all of the lip-sync above this threshold. At a resolution of 20 pixels things break down with only 49.0% of authentic videos falling below the 0.5 threshold and all lip-sync above this threshold. These results suggest that a minimum speaker interpupillary distance (IPD) of 60 pixels is sufficient to achieve accurate results.

For both video length and resolution, the pattern is similar and as expected: as the videos degrade, the authentic

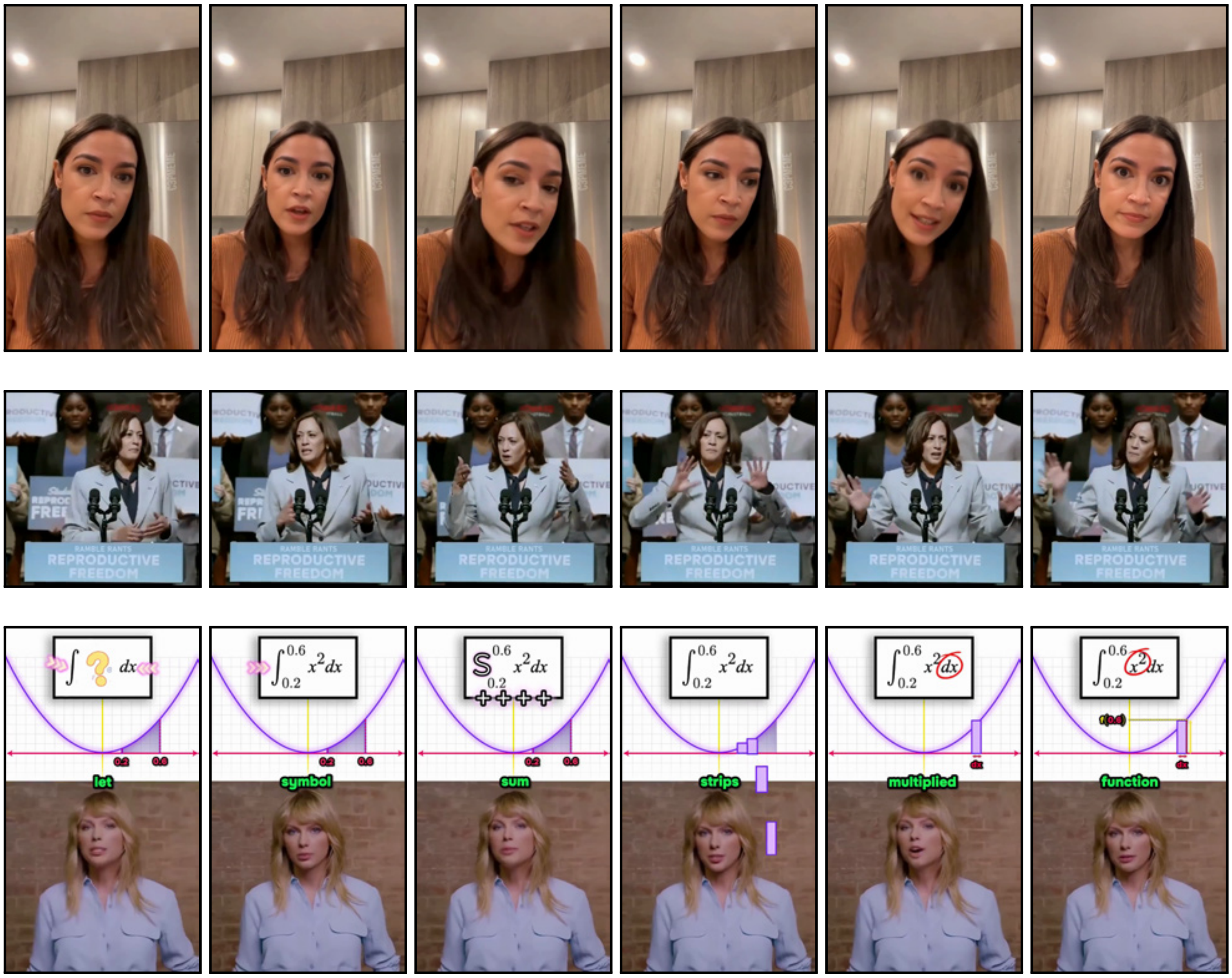


Figure 7. Equally-spaced frames from three of 14 in-the-wild lip-sync deepfakes. The Ocasio-Cortez (top) and Harris (middle) fakes were from apparently politically-motivated deepfakes meant to criticize or ridicule these politicians. The Taylor Swift video (third row) is from a brilliant series of videos in which celebrities' voices and likeness are used to teach math (https://www.tiktok.com/@onlocklearning).

videos become increasingly more difficult to extract reliable video transcriptions from, while the lip-sync videos consistently yield inconsistent audio and video transcriptions.

4.2. Anderson Cooper

Starting with the 64 videos of Anderson Cooper described in Section 2.1, we created 640 VideoRealking and 640 Wav2Lip lip-sync deepfakes where a random audio from one video was swapped into the other.

Shown in Figure 5 is the distribution of normalized Levenshtein distances for the authentic and VideoRetalking and Wav2Lip lip-sync videos. The median distance for the authentic videos is 0.16 with a minimum of 0.05 and a maximum of 0.40. The median distances for the VideoRetalking and Wav2Lip lip-syncs are 0.66 and 0.63 with a minimum distance of 0.43 and 0.44. All of the authentic videos fall below a threshold of 0.5, with 97.5% of the VideoRetalking and 98.3% Wav2Lip videos above this threshold.

4.3. In the Wild

For the 14 in-the-wild deepfakes described in Section 2.1 (see Figure 7), the median normalized Levenshtein distance is 0.70 with a minimum distance of 0.61 and a maximum distance of 0.91, Figure 6. Using the same distance threshold of 0.5 as above, all of the lip-sync videos are confidently classified as fake.

4.3.1 Comparison

By comparison to these in-the-wild accuracies, the viseme-phoneme technique [2] detects between 94% and 97% of

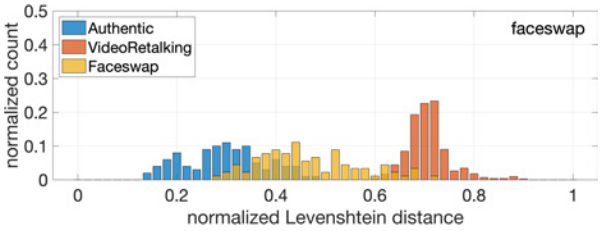


Figure 8. The distance between audio and video transcriptions for 90 face-swap deepfakes. A larger value corresponds to larger mismatches.

in-the-wild lip-sync deepfakes while mis-classifying 0.5% of real videos as fake. The person-specific extension of the viseme-phoneme [3] performs slightly better detecting 98% of lip-sync while mis-classifying 0.5% of real videos as fake. The lip-dynamic learning-based approach [25] reports an overall accuracy for real and fake on the order of 97%.

Across all three of these earlier approaches, we report higher overall detection accuracies with fewer false alarms. In addition, our approach does not require the large amounts of training data required by the second and third approach.

4.4. Face Swap

Our technique is specifically designed to detect lip-sync deepfakes in which the mouth region is synthesized to correspond to a new audio track. We wondered if this same technique would be applicable to face-swap deepfakes in which the speaker's face – eyebrows to chin and cheek to cheek – is replaced with another identity. This type of deepfake is likely to be more challenging for us because the shape of the mouth in the original video is already consistent with the audio.

Using the same original videos clipped to the first 60 seconds (see Section 2.1), we created 90 face-swap deepfakes in which the face in each of the original 10 videos is replaced with each of 9 other faces.

The median normalized Levenshtein distance for these deepfakes is 0.44 with a minimum distance of 0.29 and a maximum distance of 0.72. Using the same distance threshold of 0.5, only 35.6% of the deepfakes are correctly classified. As shown in Figure 8, the distribution of distances for these face-swap deepfakes lies midway between the authentic and lip-sync deepfakes. As expected, these fakes are more difficult to detect with this technique, but some of the lower-quality fakes are detectable.

4.5. Audio-to-Video

A new breed of audio-to-video deepfakes has emerged in which a single image of a person is synthesized so that the mouth, face, and gestures are animated to be consistent with an audio track [8, 23].

We subjected five videos created by VLOGGER[2] using the same analysis as described above (the code for this technique is not yet publicly available so we could not create a larger set of deepfakes). Even though these videos are only 10 seconds in length, the normalized Levenshtein distance between the audio and video transcriptions were 0.63, 0.71, 0.73, 0.77, and 0.87 – well above our threshold of 0.5.

On the other hand, the five English-language videos [3] created by EMO [23] – ranging in length from 25 to 58 seconds – yield distances of 0.15, 0.26, 0.40, 0.48, and 0.51, only one of which is (barely) above our 0.5 threshold.

Because these datasets are so limited in size, it remains to be seen how effective our technique will be in detecting this new type of deepfake.

5. Discussion

When watching a person talk, we take in auditory and visual information. A large body of scientific literature has shown that our perception of speech is fundamentally multisensory [19] meaning that the auditory and visual information are combined to determine our final perception.

This type of multi-sensory integration yields a single coherent percept when the signals are consistent. When there is a mismatch, however, between the auditory and visual signals, it is not necessarily the case that we will notice that something is amiss.

Classic multi-sensory illusions provide insight into how we perform auditory-visual integration. The McGurk effect [15], in which a spoken syllable is perceived to be different depending on the visual shape of the mouth, and the ventriloquist effect [4], in which the source of spoken words is misattributed, each illustrate how our perceptual system can create a coherent percept in the face of inconsistent or ambiguous auditory and visual signals. As such, forensic techniques that independently analyze auditory and visual data streams – unlike our brains – may, therefore, be more sensitive than our perceptual system.

By separately transcribing the auditory and visual signals, we find glaring mismatches that are not always perceptually obvious but that provide a powerful forensic clue. This approach is robust to the length and resolution of the video, is applicable to a variety of different lip-sync synthesis engines, and is not data-intensive in terms of requiring large amounts of training data (and therefore does not suffer from the typical out-of-domain issues).

What is a bit surprising about the efficacy of this approach is that the creation of deepfakes can incorporate explicit modeling of mouth shape to properly match the shape of the mouth to the audio [17, 24]. This matching, however,

[2] https://enriccorona.github.io/vlogger

[3] https://humanaigc.github.io/emote-portrait-alive

is not (yet) perfect. We expect that our approach, alongside other complementary approaches will make detection of lip-sync deepfakes more effective. It remains to be seen if our approach will also be applicable to puppet-master deepfakes in which the movements of one person's entire face and head are driven by another.

A limitation of our approach is that we are not able to video transcribe non-English speech due to a lack of accurate lip-reading for non-English language. We trust, however, that this type of transcription will eventually be available.

Another limitation of our approach is that a sufficiently sophisticated adversary can implement this analysis to determine if their video will be detected as fake. Therefore, as with all forensic techniques, it is important to clarify that a lack of audio-video inconsistency is not proof of an authentic video.

The primary advantage of our technique is that it is simple, explainable, generalizable and does not require intensive data collection. For the latter, unlike the typical learning-based approaches, our approach appears to easily generalize to different synthesis engines. Similar to all techniques, however, this forensic technique will almost certainly have a limited shelf life, and will have to be retired when lip-sync deepfakes perfect the spatial and temporal consistency of the mouth shape and dynamics.

Acknowledgments

This work was supported in part by funding from Adobe, YouTube, and the UC Noyce Initiative. We thank our volunteer participants for their video recordings, and Emily Cooper for many productive and insightful conversations.

References

[1] Shruti Agarwal and Hany Farid. Detecting deep-fake videos from aural and oral dynamics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, pages 981–989, 2021. 2

[2] Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshop*, pages 660–661, 2020. 2, 6

[3] Shruti Agarwal, Liwen Hu, Evonne Ng, Trevor Darrell, Hao Li, and Anna Rohrbach. Watch those words: Video falsification detection using word-conditioned facial motion. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4710–4719, 2023. 2, 7

[4] David Alais and David Burr. The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 14(3):257–262, 2004. 7

[5] Logan Blue, Kevin Warren, Hadi Abdullah, Cassidy Gibson, Luis Vargas, Jessica O'Dell, Kevin Butler, and Patrick Traynor. Who are you (I really wanna know)? detecting audio deepfakes through vocal tract reconstruction. In *31st USENIX Security Symposium*, pages 2691–2708, 2022. 2

[6] Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei Yin, Mingrui Zhu, Xuan Wang, Jue Wang, and Nannan Wang. Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In *SIGGRAPH Asia*, pages 1–9, 2022. 1

[7] Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei Yin, Mingrui Zhu, Xuan Wang, Jue Wang, and Nannan Wang. VideoRetalking: Audio-based lip synchronization for talking head video editing in the wild. In *SIGGRAPH Asia*, pages 1–9, 2022. 3

[8] Enric Corona, Andrei Zanfir, Eduard Gabriel Bazavan, Nikos Kolotouros, Thiemo Alldieck, and Cristian Sminchisescu. VLOGGER: Multimodal diffusion for embodied avatar synthesis. arXiv:2403.08764, 2024. 7

[9] Jiankang Deng, Jia Guo, Zhou Yuxiang, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-stage dense face localisation in the wild. arXiv, 2019. 2

[10] Hany Farid. Creating, using, misusing, and detecting deep fakes. *Journal of Online Trust and Safety*, 1(4), 2022. 1

[11] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5039–5049, 2021. 2

[12] Sarthak Kamat, Shruti Agarwal, Trevor Darrell, and Anna Rohrbach. Revisiting generalizability in deepfake detection: Improving metrics and stabilizing transfer. In *IEEE/CVF International Conference on Computer Vision*, pages 426–435, 2023. 2

[13] Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[14] Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. Auto-AVSR: Audio-visual speech recognition with automatic labels. In *ICASSP International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE, 2023. 3

[15] Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976. 7

[16] Gonzalo Navarro. A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1):31–88, 2001. 3

[17] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV 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*, pages 484–492, 2020. 1, 3, 7

[18] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023. 3

[19] Lawrence D Rosenblum. Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science*, 17(6):405–409, 2008. 7

[20] Sefik Ilkin Serengil and Alper Ozpinar. LightFace: A hybrid deep face recognition framework. In *Innovations in Intelligent Systems and Applications Conference*, pages 23–27. IEEE, 2020. 3

[21] Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. arXiv:2304.09116, 2023. 1

[22] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing Obama: Learning lip sync from audio. *ACM Transactions on Graphics*, 36(4):1–13, 2017. 1

[23] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. EMO: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. arXiv:2402.17485, 2024. 7

[24] Jiadong Wang, Xinyuan Qian, Malu Zhang, Robby T Tan, and Haizhou Li. Seeing what you said: Talking face generation guided by a lip reading expert. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14653–14662, 2023. 7

[25] Chen-Zhao Yang, Jun Ma, Shilin Wang, and Alan Wee-Chung Liew. Preventing deepfake attacks on speaker authentication by dynamic lip movement analysis. *IEEE Transactions on Information Forensics and Security*, 16:1841–1854, 2020. 2, 7

[26] Li Yujian and Liu Bo. A normalized Levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1091–1095, 2007. 3

[27] Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *IEEE/CVF International Conference on Computer Vision*, pages 14800–14809, 2021. 2