Watch Those Words: Video Falsification Detection Using Word-Conditioned Facial Motion

Shruti Agarwal*[1], Liwen Hu[2], Evonne Ng[1], Trevor Darrell[1], Hao Li[2], and Anna Rohrbach[1]

[1]University of California, Berkeley
[2]Pinscreen, Inc.

Abstract

In today's era of digital misinformation, we are increasingly faced with new threats posed by video falsification techniques. Such falsifications range from cheapfakes (e.g., lookalikes or audio dubbing) to deepfakes (e.g., sophisticated AI media synthesis methods), which are becoming perceptually indistinguishable from real videos. To tackle this challenge, we propose a multi-modal semantic forensic approach to discover clues that go beyond detecting discrepancies in visual quality, thereby handling both simpler cheapfakes and visually persuasive deepfakes. In this work, our goal is to verify that the purported person seen in the video is indeed themselves by detecting anomalous facial movements corresponding to the spoken words. We leverage the idea of attribution to learn person-specific biometric patterns that distinguish a given speaker from others. We use interpretable Action Units (AUs) to capture a person's face and head movement as opposed to deep CNN features, and we are the first to use word-conditioned facial motion analysis. We further demonstrate our method's effectiveness on a range of fakes not seen in training including those without video manipulation, that were not addressed in prior work.

1. Introduction

Humans tend to trust what they see, especially when it comes to video. Historically, video has been the best proof that an event has indeed occurred. However, in the rapidly evolving misinformation landscape of the present digital era, this may not be true for long. Video manipulation techniques are more accessible than ever, while the reach of internet and social media enables rapid spread of falsified content. Recent headlines, such as *"XR Belgium posts deepfake of Belgian premier linking Covid-19 with climate crisis"* [43], *"Dutch MPs in video conference with deepfake imitation of Navalny's Chief of Staff"* [44][1], *"When virtual turns fake: Danish politicians 'meet' Belarusian opposition figure"* [39] are examples of how both deepfakes and cheapfakes (e.g. lookalikes) pose real threats and have serious consequences, especially if targeted at people in power.

To protect the public against potential disinformation

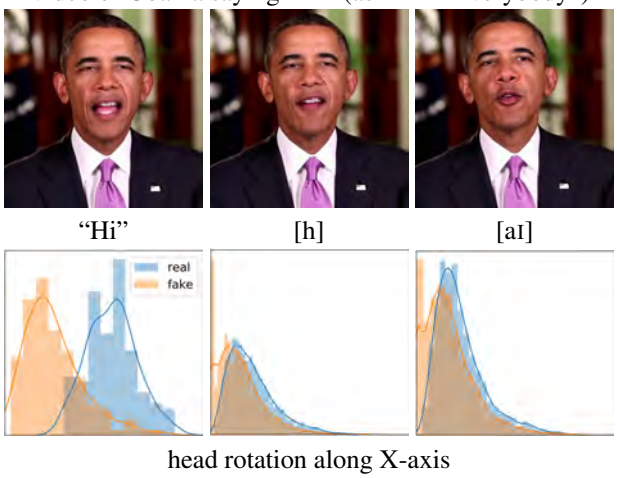

Figure 1. We are looking for inconsistencies between *what they say* and *how they move*. One way to address this is via audio, or sounds (phonemes), however this may miss some important semantic cues. Consider an example of Obama saying "Hi" (as in "Hi Everybody"). We noticed that each time he says that, he rotates his head along X-axis. If we plot the amount of "head rotation along X-axis" for each occurrence of the word "Hi", we see a clear separation between the real and fake Obama videos, allowing us to successfully detect falsification. At the same time, the two phonemes, [h] and [aɪ], that constitute the word "Hi", do not have any correlation with head rotation. This is intuitive since these phonemes also occur in many other words, where Obama does not move his head. This motivates us to focus on spoken *words* to discover biometric word-specific patterns.

*shrutia@mit.edu

[1]A later report explained that in fact it was an impersonator [45].

campaigns, new deepfake detection methods are being introduced to combat new and more advanced deepfake techniques [41, 32, 25, 34, 19, 35, 36, 42]. Not only is detection increasingly challenging, most methods are ineffective against cheapfakes falsified through conventional techniques (e.g., speeding up or slowing down a video), or with no video manipulation at all (a lookalike, audio dubbing).

In this work, we aim to detect video falsifications related to a person's identity. Specifically, we aim to detect if the purported person "seen" in a video is indeed themselves. This is distinct from the deepfake detection problem, where the goal is to distinguish between pristine (non-manipulated) and generated/altered videos. For such methods, a video of an impersonator would be wrongly identified as "real". Similarly, a non-manipulated video with dubbed or edited speech would also be considered "real" by many deepfake detectors, since often they do not take speech into account. In contrast, our problem statement is more general, as it includes both deepfakes and falsified pristine videos. Furthermore, as the deepfakes' quality improves, detecting visual flaws will also become increasingly difficult. Our key insight is to use *semantic, person-specific cues as an alternative, and generalizable solution to detecting video falsifications.*

Biometric-based techniques [5] have been recently introduced to identify falsified videos, that are either not manipulated or extremely realistic. [5] analyze the authenticity of a person by correlating the head and facial movements in the existing footage of the person. Even though several hours of training video is required, they are well-suited for public figures such as celebrities and world leaders who are often a target. However, these person-specific methods are ineffective against cutting-edge audio-to-lip synthesis techniques such as [41, 36] or commercial video dialog replacement solutions developed by Synthesia[2] or Canny AI[3], which only manipulate the mouth.

To this end, we propose a semantic, multimodal detection approach that integrates speech transcripts into person-specific gesture analysis. We leverage interpretable Action Units (AUs) [6] to model a person's face and head movement. Our approach is to analyze the *word-conditioned facial movement* captured with AUs to learn per-word models for classifying real and fake videos. Our intuition is that each individual may have identifying, unique patterns in how their speech, facial expressions, and gestures co-occur. This is distinct from using raw audio or sounds, as illustrated in Figure 1. The individual phonemes lack the semantics and thus may not capture high-level idiosyncratic regularities of facial gestures associated with specific words. Our word-level models distinguish real and falsified videos using the facial patterns associated with specific words. At test time, we compute classification scores for each word in a video clip, and aggregate them into a final score.

We experiment using real/fake videos of world leaders and TV talk-show hosts, where we consider a full spectrum of cutting edge video manipulation techniques [36, 35, 41], as well as fakes found in the wild. We compare our approach to several prominent prior works, and we show that we achieve the best performance across the entire range of fakes. No other method that we consider demonstrates such general capability, as they tend to suffer on audio dubbing or in-the-wild lip-sync fakes.We additionally compare using phonemes instead of words within our model design, and see that while it performs well on cases with audio-visual inconsistencies, it struggles to recognize the fakes that require biometric features. We perform ablation studies to confirm that the key advantage of our method is indeed the word-conditioned analysis. Lastly, an added benefit of our approach is interpretability: we are able to capture human understandable, person-specific word-movement patterns predictive of a video being real or falsified (e.g., common in real videos but absent in the fakes ones).

Our contributions are as follows. (a) We present a new, general problem statement: given a video, predict whether a person is authentic, regardless if falsification is a deep- or cheapfake. (b) We propose the first semantic person-specific approach to address this problem that leverages word-conditioned facial movements. (c) We perform a comparative study across multiple fake types, ranging from deepfakes to impersonators and audio dubbing. Unlike prior work, our approach shows strong generalization across all types of fakes. Namely, our method exhibits two key capabilities: recognizing speech-video inconsistency, while also capturing biometric features. (d) Our approach also offers interpretability, allowing us to expose the person-specific predictive word-gesture patterns.

2. Related Work

We identify two types of detection techniques: (1) Person-generic methods analyze whether manipulation occurred regardless of the person's identity; (2) Person-specific methods verify whether the characteristics of the seen individual match the real person. Person-generic approaches are often trained on large datasets with real and fake videos, and rely on either **low-level features** or **high-level semantics**. Person-specific methods, on the other hand, typically require additional **biometric-based** data for identification.

Low-Level Feature-Based Forensics. These methods (often CNN classifiers) are typically person-generic and focus on visual artifacts or statistical anomalies learned implicitly from images or videos [1, 21, 33, 51, 30, 40, 52, 53, 33, 38, 46]. While many techniques struggle to generalize to new video manipulation techniques or unseen deep-

[2]https://www.synthesia.io/

[3]https://www.cannyai.com/

fake videos [16], some focus on artifacts that appear also for unseen fakes. [29] detect the warping artifacts, [27] identify blending traces during synthesis, and [47] leverage inconsistencies between images and meta-data. While promising detection capabilities have been shown, these methods are often susceptible to deteriorations like compression, resolution reduction, or adversarial perturbations and attacks [8, 24].

High-Level Semantic-Based Forensics. Person-generic and high-level semantic-based techniques focus on explicit anomalies of person's characteristic, such as the absence of eye blinking [28], the inconsistencies in head pose [50], human physiological signals like heart beat [18, 37], ear movements [2], and other biological signals [10]. These approaches often generalize better to unseen deepfakes and are more resilient to laundering. However, a reliable extraction of high-level features is often difficult to achieve in unconstrained settings and short video clips. Several recent methods focus on temporal inconsistencies in facial performances [22, 9, 31, 4] but rely on robust 3D face tracking.

Similar to our proposed work, multi-modal techniques [9, 31, 54] exploit both the audio and visual signal to detect deepfakes. While audio signals can provide cues like emotions and *how* a person is talking, our work focuses on spoken words, which provide a more direct information about *what* is being said. For example, different head nods associated with words convey agreement, disagreement or greetings in many cultures [13]. We believe that our approach is complementary to audio-based methods, since it captures word-specific patterns, inaccessible to raw audio, as illustrated in Figure 1. In [4], the authors exploit the shape of the lips when phonemes 'P', 'B', or 'M' are being pronounced. Whereas in [22], the authors use only visual signal to detect if the lip movements are 'readable" in a video. Even though these techniques can detect deepfakes where the lips are modified, since they are not person-specific they will struggle identifying video falsifications using impersonators.

Biometric-Based Forensics. Biometric-based detection methods [11, 48, 5, 12, 3, 49, 26] are person-specific as they try to verify the authenticity of a person using known identity priors. These works are the most relevant to our technique and many of them exploit person-specific facial movement over time to detect deepfakes. In [26, 49], the authors use visual appearance and movement of lips to perform speaker verification and detect person-specific deepfakes. In these previous works, the authors analysed only a small set of words which will restrict their approach to fakes where those are being said. In contrast, we include the facial movements from the entire face and use a much larger vocabulary size, enabling our approach to handle in-the-wild deepfakes. The method of [5] introduced a biometric approach for public figures, where person-specific facial movements in a video are compared with those of pristine videos. Despite the requirement for hours of training data of a known person, this approach is resistant to realistic deepfakes, or even to lookalikes, when no video manipulation is used. More advanced techniques that incorporate CNN-based behavior classification using optical flow [3] or 3DMM-based facial tracking [12] have shown improved performance for deepfake detection. Nevertheless, recent advancements in speech-to-lip synthesis [41, 36] show that it is possible to produce highly convincing speech manipulations without altering global facial characteristics. In this work, we introduce a multi-modal semantic approach that exploits the fact that spoken words may be associated with distinct person-specific facial movements. In particular, these movements involve the entire face/head and not only the lip region of a person, and are difficult to disguise even for skilled impersonators.

3. Word-Conditioned Facial Analysis

Given an input video of an individual, our goal is to classify it as real or fake. We leverage the key insight that individuals often use identifying gestures associated with specific interactions like greeting, disagreement, etc. In our approach, we represent these conversational units in terms of words and analyze the facial gestures associated with them. Considering conversational units at the granularity of words gives us a good trade-off between the number of occurrences of each unique unit and speech semantics. Using N-grams or unique sentences would result in fewer occurrences while phonemes would result in less meaningful speech semantics. We include an empirical comparison of our approach to its phoneme-conditioned counterpart in Section 5.2.

As shown in Figure 2, we first transcribe the audio and then extract the corresponding per-frame facial movements of the speaker represented by AUs [17]. We encode the speaker motion as the amount of change in the AUs that happens within the window of the word's occurrence. Finally, word-level classifiers are trained to detect whether the visual movement match with the spoken words.

Word-Aligned Facial Feature Extraction. We denote $\mathbf{F}_{1:T} = \{f\}_{i=1}^{T}$ as a set of frames f from a video of length T. Given a video, we transcribe the audio of the video to get the phonation time of each word w, expressed as the start f_s and end f_n frames, where $d = n - s$ is the duration of the phonation. To associate an individual's facial expression and head motion with the corresponding word, we extract AUs for the window $\mathbf{F}_{s:n}$. In contrast to 3-D or 2-D facial landmarks, these AUs represent semantically meaningful micro-expressions such as the strength of a cheek or chin movement (e.g. "chin raiser"). For a given word spoken within the frame range $\mathbf{F}_{s:n}$, we extract a 25-D facial feature $\vec{g}_i$ at each timestep to obtain $\mathbb{G}_{s-t:n+t} = \{\vec{g}\}_{i=s-t}^{n+t}$.

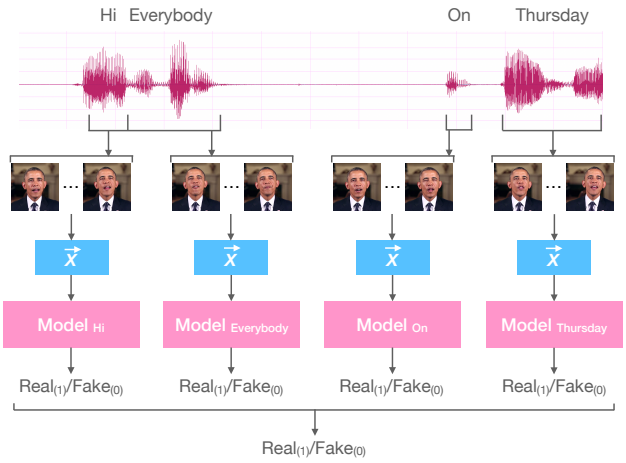

Figure 2. Given an input video, we first transcribe the audio and obtain per-frame alignments of word. For each word, an AU-based feature vector $\vec{x}$ is extracted from the corresponding frames $\mathbb{F}$ and then evaluated by a word-specific classifier $\vec{\theta}$. A final score for the possibility of "Real" is computed for the video using the geometric mean of all the scores.

A padding of t=3 frames is added to account for small misalignment between the words and the frames. However, we assume that there is no large misalignment between the video and audio signal. Each 25-D facial feature is composed of 4 components: (1) the intensities of the 17 AUs, (2) the 3-D head rotations and 3-D head translations along X, Y, Z axis, (3) the 3-D horizontal distance between mouth corners (lip-hor), and (4) the 3-D vertical distance between the lower and upper lip (lip-ver). Instead of using a variable-length feature $\mathbb{G}_{s-t:n+t} \in \mathbb{R}^{d\times 25}$, we use the deltas between the maximum and the minimum values extracted during the word phonation window. The facial feature of each word occurrence is then expressed as:

$$\vec{x}_w = \max_{\vec{g}_i \in \mathbb{G}_{s-t:n+t}} (\vec{g}_i) - \min_{\vec{g}_i \in \mathbb{G}_{s-t:n+t}} (\vec{g}_i), \tag{1}$$

where $\vec{x}_w \in \mathbb{R}^{25\times 1}$ is used for building a word-specific model. Intuitively, these features capture the maximum range of movement happening when a word is spoken (how much the head moves up when saying "Hi") regardless of the temporal misalignment of features. E.g., in the real videos of Obama the word "Hi" spans a minimum of 9 frames and a maximum of 27 frames. By using the range of motion as the feature, we are thus avoiding temporal variability across different utterances of the same word.

Word-Specific Classifiers. We train linear per-word classifiers to tell if the given gesture features belong to the given words. Instead of using more complex learning-based approaches with high-dimensional features, we use linear classifiers to highlight the efficacy of our interpretable features in a simple model. Given real videos where the words-specific facial movements are correct, we create simulated fakes where words are deliberately matched with random facial movements (speech transcript matched to a wrong video). In addition to that, we create synthesized fakes using the recent lip-sync generation method Wav2Lip [36]. As before, words are deliberately matched with random facial movements but now the lips are synthesized to say the words. By using these synthesized fakes, we ensure that our classifier does not rely only on lip reading errors in order to detect fakes. Using this real and fake data, we train person-specific word-specific logistic regression classifiers.

Let $\vec{x}_w \in \mathbb{R}^{25\times 1}$ be the facial feature corresponding to word w. Let $y_w \in [0, 1]$ be the ground truth label of $\vec{x}_w$ where $y_w = 1$ if $\vec{x}_w$ is from a real video sequence. We learn the model parameters $\boldsymbol{\theta}_w \in \mathbb{R}^{25\times 1}$ for a linear classifier that maximizes the following objective function $L_{\boldsymbol{\theta}_w}$:

$$L_{\boldsymbol{\theta}_w} = \prod_{i=1}^{M} P(y_i|\vec{x}_i), \tag{2}$$

where $P(y_i|\vec{x}_i)$ is the probability of y_i given $\vec{x}_i$ and M is the number of total occurrences of w in the training data.

$$P(y_i|\vec{x}_i) = [\sigma(\boldsymbol{\theta}_w^\top \cdot \vec{x}_i)]^{y_i}.[1 - \sigma(\boldsymbol{\theta}_w^\top \cdot \vec{x}_i)]^{1-y_i} \tag{3}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function.

Testing. During evaluation, given a test video of a purported individual, we extract the features as described above. The transcribed words not seen in the training set are discarded. For each remaining word w, the corresponding facial features $\vec{x}_w$ are examined using the target word classifier $\boldsymbol{\theta}_w$ for the given individual. A score s_w in the range of 0 (fake) and 1 (real) for $\vec{x}_w$ is computed as:

$$s_w = \sigma(\boldsymbol{\theta}_w^\top \cdot \vec{x}_w). \tag{4}$$

A final score is computed for a given video using the geometric mean of scores across all trained words in the video.

4. Dataset

To validate our proposed approach on the general problem statement that includes both deepfakes and non-manipulated fakes, we compile the following dataset. We consider four US politicians (Barack Obama, Donald Trump, Joe Biden, Kamala Harris) and two TV talk-show hosts (John Oliver, Conan O' Brien). Further we provide the details for the types of data that we use.

Real: The real videos of the politicians were taken from the World Leaders Dataset (WLDR) [3] and the videos of the talk-show hosts were taken from [20]. The total hours and example frames are shown in Table 1 (Column 1) and Figure 3 (Column 1).

Dubbing: Using the real videos for each individual, we simulate the dubbing scenario by mismatching the video and the audio. For every real video, a new dubbed video

Figure 3. Examples of the data used in our work, spanning different types of falsified video, from deepfakes to fakes with non-manipulated video.

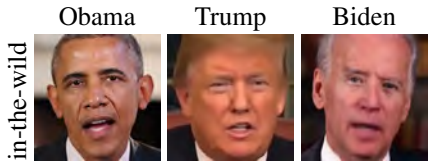

Figure 4. Examples of the data used in our work, specifically, in-the-wild lip-sync examples for three individuals.

is created by matching it with a random audio of the same length. We produce the same number of hours for the dubbed videos as we have of real videos, Table 1.

Wav2Lip: Using the real videos, we create lip-sync deepfakes where the lip region in the video is modified to match a random audio. We use the off-the-shelf implementation of Wav2Lip [36] to create these fakes. The example frames are shown in Figure 3 (Column 2).

Impersonator: The person-specific impersonator videos are obtained from Saturday Night Live on YouTube. The impersonator videos for Obama, Biden, and Trump are from WLDR, and for Harris, Oliver and O'Brien from YouTube. The total hours and example frames are shown in Table 1 (Column 4) and Figure 3 (Column 3).

FaceSwap: The FaceSwap deepfakes are created using the impersonator videos by replacing impersonator's face with the target person's face. The videos for Obama, Biden, and Trump are from WLDR and for Harris, Oliver and O'Brien are created using the FaceSwap library [35]. The total hours and example frames are shown in Table 1 (Column 5) and Figure 3 (Column 4).

	Real	Dubbing	Wav2Lip	Impersonator	FaceSwap	itw
Obama	12.5	12.5	12.5	0.16	0.11	0.99
Trump	6.1	6.1	6.1	0.19	0.19	0.08
Biden	5.1	5.1	5.1	0.04	0.14	0.12
Harris	2.5	2.5	2.5	0.05	0.05	-
O' Brien	14.5	14.5	14.5	0.08	0.08	-
Oliver	17.8	17.8	17.8	0.04	0.04	-

Table 1. Number of hours of video for each of the six individuals for different types of video falsification scenarios.

	Training		Number of unique words during tesing			
	Number of Words	Number of models	Real/ Dubbing/ Wav2Lip	Impersonator	FaceSwap	itw
Obama	4,925	918	812	248	211	543
Trump	3,664	817	543	296	282	81
Biden	3,985	816	523	133	145	121
Harris	2,270	844	346	125	124	-
O' Brien	6,306	657	548	196	187	-
Oliver	10,330	739	670	118	98	-

Table 2. Columns 2-3: Training data statistics in terms of the total number of unique words and the number of word-specific models that we train. Columns 4-7: number of unique word models tested in each sub-task.

In-the-wild (itw): The in-the-wild lip-sync videos for Obama, Trump, and Biden are collected from [2, 5, 41]. The total hours and example frames are shown in Table 1 (Column 6) and Figure 4.

5. Experiments

We evaluate our approach on five falsification scenarios and compare it with the state-of-the-art deepfake detection methods and a phoneme-based baseline. We also provide several ablations and analysis studies. We end with showcasing our method's interpretability.

5.1. Implementation Details

Data Preprocessing: Each video is first preprocessed so that only the person of interest is retained. Given one frame of an input video, we first use a single-stage face detector [14] to localize all the faces. Then a face recognition network ArcFace [15] is used to check whether each face is the target person, and the outliers are masked out. (For impersonator videos, the face of impersonator is used instead of the target person.) For transcription, we used an open-source implementation of DeepSpeech [23]. For AU extraction, we use the facial behavior analysis toolkit OpenFace2 [6, 7].

Training Details: In our experiments, we use logistic regression to solve the binary classification problem of real/fake video. To train our person-specific word classi-

	Dubbing	Wav2Lip	Impersonator	FaceSwap	itw
Obama	1.00	1.00	0.95	0.90	0.98
Trump	0.95	0.99	0.89	0.92	0.98
Biden	0.84	0.93	0.98	0.73	0.95
Harris	0.90	0.89	0.82	0.93	-
O' Brien	0.91	0.88	0.90	0.84	-
Oliver	0.94	0.93	0.86	0.87	-
Avg	0.92	0.94	0.90	0.87	0.97

Table 3. Accuracy in terms of AUC on 10-second video clips for the six individuals and five video falsification scenarios. The average AUC across individuals is given in the last row.

fiers, we use 90% of the real videos for the "Real" class, and 90% of the Dubbing and Wav2Lip lip-sync videos for the "Fake" class. The number of unique words present in the speech for each individual is given in Table 2 (column 2). The word-specific models are trained for the words with average frequency of once per hour in training dataset. For example, the overall duration of videos for Harris/Oliver is 7.5/53.4 hours. Therefore, in the case of Harris/Oliver, a word classifier is trained if the word frequency is greater than equal to 7/53, respectively. Shown in Table 2 (column 3) is the total number of word models trained for each individual. On average, 799 word models are trained, with the smallest/largest number of models trained for O'Brien/Obama.

Testing Details: We test our approach on remaining 10% of real, audio dubbing and Wav2Lip lip-sync videos. Additionally, we test on all the videos with Impersonators, FaceSwap, and in-the-wild lip-sync deepfakes which were *not seen during training* (as introduced in Section 4). Each test video is divided into overlapping 10-second video clips (30 fps) with a shift window of two seconds. Shown in the Table 2 (columns 4-7), is the total number of unique words that were evaluated in each of the test datasets, based on the occurred words within the trained words-set in testing time.

Evaluation Metric: We report the Area Under the Curve (AUC) score for the 10-second test videos. For the previous methods that perform analysis on a temporal window less than 10 seconds, we average predictions over 10 seconds.

Methods: As our approach does not analyze the audio signal for our detection model and uses only person-specific visual features conditioned on words, we compare our approach with other visual feature-based forensic techniques. Moreover, while there is prior work on audio-visual deepfake detection, we were unable to find any publicly available code bases. Thus, we select the following prior methods with available code bases: the low-level feature-based method in XceptionNet [40]; the high-level semantic-based approach in LipForensics [22]; the biometric-based techniques in Protecting World Leaders (PWL) [5] and ID-Reveal [12]. At the same time, we are interested in empirically assessing whether using words provides some additional benefits over using sounds in the audio. To address that, we construct a version of our method that uses *phonemes*[4] instead of words for determining visual windows and training classifiers. Since phonemes correspond to sounds made during speech, this serves as a proxy to audio-visual methods.

5.2. Results

Shown in Table 3, is the performance of our method in terms of AUC for each individual test case. The average AUC across all the individuals is shown in the bottom row. Our approach works the best for Obama with the average AUC of 0.97 across all types of falsification scenarios and the worst for O'Brien with an average AUC of 0.88. This is expected as the Obama videos have higher quality and better consistency in facial movement during the formal weekly addresses. The videos of O'Brien are of lower visual quality and have a wider range of facial movements during the informal interviews, monologues, and audience interactions during the talk-show. This makes it more difficult for our word-conditioned model to learn consistent facial movement patterns from O'Brien videos.

Comparison with State-Of-The-Art: Shown in Table 4 are the average AUCs across all the individuals for each method and video falsification scenarios. Our approach performs the best across all the video falsification scenarios except in case of Wav2Lip where LipForensics obtains the best performance of 0.98. All the previous methods fail to detect the dubbing video falsification scenario as there is no video manipulation performed in this case. The non-biometric techniques fail to detect impersonators' video. Even though the related biometrics-based methods are able to detect FaceSwaps and impersonators, they perform poorly on lip-sync videos. This is because these techniques only use the visual cues of a person identity, most of which are preserved in the lip-sync videos. This shows the advantage of our approach, i.e. using words in combination with the visual cues. When comparing word- to phoneme-conditioning, we see that phonemes have strong ability to detect audio-visual inconsistency, but fail in capturing person-specific features needed for Impersonator and FaceSwap fakes. This is intuitive, since phonemes correspond to sounds made over short spans and shared across many words, thus missing some semantic and idiosyncratic clues that can be leveraged via word-conditioning. To sum up, word-conditioning enables us to capture both the audio-visual inconsistency and the biometric features.

Effect of Using Words: We further analyse the effect of training the word-specific classifiers by training two different versions of our approach. In the first version (Fixed Window), we do not use the word information and compute the 25-D visual gesture features using all the non-

[4]We use the CMU Pronouncing Dictionary (https://github.com/cmusphinx/cmudict) which breaks the words from video transcripts into phonemes, of which there are 70.

	Dubbing	Wav2Lip	Impersonator	FaceSwap	itw
XceptionNet [40]	0.50	0.78	0.57	0.54	0.49
LipForensics [22]	0.50	**0.98**	0.43	0.81	0.95
PWL [5]	0.50	0.63	0.86	0.85	0.60
ID-Reveal [12]	0.50	0.66	0.85	0.78	0.61
Ours w/ Phonemes	**0.95**	0.96	0.61	0.58	**0.98**
Ours w/ Words	0.92	0.94	**0.90**	**0.87**	0.97

Table 4. The performance in terms of AUCs on 10-second video clips. For each method and video falsification scenario, shown above are the average AUCs across all six individuals.

	Dubbing	Wav2Lip	Impersonator	FaceSwap	itw
Fixed Window	0.50	0.91	0.81	0.68	0.87
Word Window	0.79	0.88	0.72	0.68	0.94
Ours w/ Words	0.92	0.94	0.90	0.87	0.97

Table 5. The average AUC performance across all individuals for two ablations of our method, see text for details.

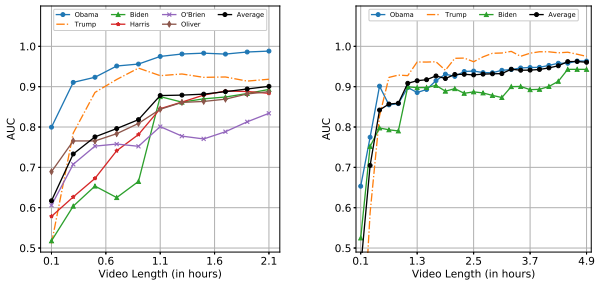

Figure 5. Effect of number of hours of training video. Left plot is evaluated on Wav2Lip fakes, the right one is evaluated on the in-the-wild fakes.

Wav2Lip	Obama	Trump	Biden	Harris	O'Brien	Oliver
Ours w/ Phonemes	0.73	0.66	0.63	0.67	0.69	0.81
Ours w/ Words	0.68	0.65	0.62	0.64	0.56	0.72

Table 6. "Transfer" performance in terms of AUCs for models trained on five people and tested on a held-out sixth person (pristine vs. Wav2Lip), reported per test person, averaged.

overlapping fixed windows of 30 frames. This window size is chosen as 95% of the words have a duration smaller or equal to 30 frames. Using the corresponding gesture features, we train a single linear classifier to predict Real vs. Fake. In the second version (Word Window), the gesture features are extracted using word intervals, as in our approach, but we train a single linear classifier instead of word-specific classifiers. Shown in Table 5, are the average AUCs across all individuals for the two ablations and our approach. While the word intervals already improve over the fixed window case, the word-specific training helps improve the performance on each type of video falsifications. This clearly shows that the key advantage of our approach is indeed in leveraging the word-conditioned facial gesture analysis.

Model Transfer and Person Specific Features: Based on the previous experiments, we can already hypothesize that our word-conditioned method captures some person-specific features, as evident from its high performance on Impersonator and FaceSwap falsified videos. To further analyze the degree of "person-specificity", we conduct a model "transfer" experiment. Namely, we use models trained for five people and test them on a held-out sixth person. Intuitively, we expect these models not to do well in distinguishing the real vs. Wav2Lip fakes of that person. The average AUC scores are reported in Table 6. First, comparing the scores to Table 3, we see that the overall "transfer" is rather poor, as expected (e.g., a model trained for Obama achieves AUC 1.0 when tested on Obama, while the "transferred" models only give AUC 0.68.) Second, the phoneme-conditioning consistently gets higher "transfer" scores, showing that while it is person-specific, it captures more person-agnostic features, e.g., generic sound-to-lips alignment.

Effect of Training Data Size: We analyse the effect of number of hours of real videos used for training person-specific word models. The effect of training size is evaluated on: 1) Wav2Lip lip-sync fakes which on average have 72% vocabulary overlap with the training dataset and 2) in-the-wild lip-sync fakes which on average have only 28% vocabulary overlap with the training dataset.

Shown in Figure 5 are the AUCs for individuals as the function of training size ranging from 0.1 to 2.1/5.0 hours of real training videos. For each real training size, we use the equal number of hours of fake videos from audio dubbing and Wav2Lip training datasets. The evaluation in the left/right plot is performed on Wav2Lip/in-the-wild lip-sync fakes. Shown with the black curve is the average AUC across all individuals as a function of training size. In each of these evaluation scenarios, the performance improves with the number of hours in training. In case of Wav2Lip, the average performance improves from 0.62 to 0.88 (42%) from 0.1 to 1.3 hours and then improves from 0.88 to 0.90 (2.0%) for training size greater than 1.3 hours. Similarly, for the in-the-wild lip-sync fakes, the average performance of 0.91 is achieved with 1.3 hours of training videos with only a slight improvement after that. This shows that while we used several hours of video per individuals, a relatively smaller training dataset (≈1.5 hours) can provide a similar performance.

Qualitative Results and Interpretability: Here we present qualitative results showing the regularity of facial movements associated with words. Shown in Figure 6 are word-based facial movements of two individuals. For each one, we select one word from the top-5 performing words. (The performance of the word-based classifiers is evaluated on our training data in terms of word-level AUC.) For each selected word and individual, two occurrences are shown, from real (top row) and Wav2Lip fake (bottom row) videos. Shown in the last column is the distribution of one gesture feature (AU) in real and fake training data. We see that

Figure 6. Qualitative examples of the facial movement for specific words that were used to predict real vs. fake. For each word and individual, we show two examples of facial movement from real and fake (Wav2Lip) videos. In the last column we show the distribution of a gesture feature in real and fake training dataset of the individual. E.g., for Trump, the lip rounding and chin raise actions during the word “tremendous” are missing in the fakes. This is supported by the distributions of AU17 and lip-hor AU: the average strength of these movements is lower in fake videos of Trump than the real ones.

Trump, while saying the word “tremendous”, rounds his lips and then presses the lips together before finally opening the lips apart. This rounding action of the lips is absent in the fake examples, even though the lips are closed once in the sequence. This difference in real and fake utterances can also be seen in the distributions of change in chin-raise (AU17) and change in “lip-hor” AUs. For Oliver, the word “billion” is associated with the creation of dimples on the cheeks, which is violated in the fake frames shown here. Thus, in addition to showing good generalization across a range of video falsifications, our method provides interpretability, offering insight into what words/gestures may be responsible for classifying a video as a fake. This is an important capability for an analyst using this tool.

6. Discussion and Limitations

We proposed a novel multi-modal, semantic-based approach for detecting falsified videos. We leverage the idea of learning person-specific associations between the speaker’s facial gestures and spoken words to verify the purported person’s identity. Our experiments show that inconsistent head movements and facial expressions can be identified reliably when an impersonator is used for falsification. Moreover, we demonstrate the effectiveness and robustness of our approach on a wide range of deep and cheapfakes, outperforming all other methods in most cases. Since we do not attempt to detect video manipulation artifacts, our method will still work for more advanced future deepfakes. While other multi-modal detection techniques have shown that audio is an important cue for revealing falsifications, our semantic approach of using words can be an important addition, especially for cases when similar sounding words with different meanings are used. Our experiments with word vs. phoneme conditioning support that.

Our current approach relies on the accuracy of 3-D facial tracking via AU extraction. While this is feasible for our dataset, where the speakers are often front-facing, for unconstrained videos, deep learning based features may be more reliable. Although our method seems to behave rather sample efficiently (Figure 5), it is person-specific and thus requires sufficient ($\approx$ 1 hour) training data to be effective. This data requirement can easily be satisfied for celebrities and world-leaders who are the most vulnerable to deep-fake attacks. Furthermore, while the AUs allow us to obtain interpretable results, denser 3-D facial features could allow for detecting more subtle anomalies. Finally, we have only validated our method for English speech. In the future, we would like to explore how well our word-conditioned technique would work with other languages.

Falsified media is a threat to society, so we envision positive impact from our work. At the same time, almost any method for fake detection may be adapted to create more robust fakes. As the visual quality of fakes keeps improving, it will be increasingly important to build such biometric models to mitigate the harm of deepfakes.

Acknowledgements. This work was supported in part by DoD including DARPA’s SemaFor, PTG and/or LwLL programs, as well as BAIR’s industrial alliance programs. We thank Sarthak Kamat for his valuable feedback during the project.

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. MesoNet: A compact facial video forgery detection network. In *IEEE International Workshop on Information Forensics and Security*, 2018.

[2] Shruti Agarwal and Hany Farid. Detecting deep fakes from aural and oral dynamics. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2021.

[3] Shruti Agarwal, Hany Farid, Tarek El-Gaaly, and Ser-Nam Lim. Detecting deep-fake videos from appearance and behavior. In *IEEE International Workshop on Information Forensics and Security*, 2020.

[4] Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[5] 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 Workshops*, 2019.

[6] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2015.

[7] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. OpenFace 2.0: Facial behavior analysis toolkit. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2018.

[8] Nicholas Carlini and Hany Farid. Evading deepfake-image detectors with white-and black-box attacks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[9] Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. Not made for each other-audio-visual dissonance-based deepfake detection and localization. In *ACM International Conference on Multimedia*, 2020.

[10] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. FakeCatcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[11] J.F. Cohn, K. Schmidt, R. Gross, and P. Ekman. Individual differences in facial expression: stability over time, relation to self-reported emotion, and ability to inform person identification. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, pages 491–496, 2002.

[12] Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15108–15117, October 2021.

[13] Charles Darwin. *The expression of the emotions in man and animals*. University of Chicago press, 2015.

[14] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[15] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[16] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (DFDC) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019.

[17] Paul Ekman and Wallace V Friesen. Measuring facial movement. *Environmental Psychology and Nonverbal Behavior*, 1976.

[18] Steven Fernandes, Sunny Raj, Eddy Ortiz, Iustina Vintila, Margaret Salter, Gordana Urosevic, and Sumit Jha. Predicting heart rate variations of deepfake videos using neural ode. In *IEEE International Conference on Computer Vision Workshops*, 2019.

[19] 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.

[20] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[21] David Güera and Edward J. Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2018.

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

[23] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[24] Shehzeen Hussain, Paarth Neekhara, CRISTIAN CANTON FERRER, JULIAN MCAULEY, and FARINAZ KOUSHANFAR. Exposing vulnerabilities of deepfake detection systems with robust attacks. *Digital Threats: Research and Practice*, 2021.

[25] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Nießner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt. Deep video portraits. *ACM Transactions on Graphics*, 2018.

[26] Pavel Korshunov and Sébastien Marcel. Speaker inconsistency detection in tampered video. In *European Signal Processing Conference*, 2018.

[27] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face X-ray for more general face forgery detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[28] 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*, 2018.

[29] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[30] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92, 2019.

[31] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *ACM International Conference on Multimedia*, 2020.

[32] Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. PaGAN: Real-time avatars using dynamic textures. *ACM Transactions on Graphics*, 2018.

[33] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2019.

[34] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: Subject agnostic face swapping and reenactment. In *IEEE International Conference on Computer Vision*, 2019.

[35] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, et al. Deepfacelab: A simple, flexible and extensible face swapping framework. *arXiv preprint arXiv:2005.05535*, 2020.

[36] 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 *ACM International Conference on Multimedia*, 2020.

[37] Hua Qi, Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Wei Feng, Yang Liu, and Jianjun Zhao. DeepRhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In *ACM International Conference on Multimedia*, 2020.

[38] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*, 2020.

[39] Reuters. When virtual turns fake: Danish politicians 'meet' belarusian opposition figure. https://www.reuters.com/article/us-denmark-belarus/when-virtual-turns-fake-danish-politicians-meet-belarusian-opposition-figure-idUSKBN26T2V4, 2020.

[40] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *IEEE International Conference on Computer Vision*, 2019.

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

[42] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *European Conference on Computer Vision*, 2020.

[43] Brussels Times. Xr belgium posts deepfake of belgian premier linking covid-19 with climate crisis. https://www.brusselstimes.com/news/belgium-all-news/politics/106320/xr-belgium-posts-deepfake-of-belgian-premier-linking-covid-19-with-climate-crisis/, 2020.

[44] NL Times. Dutch mps in video conference with deep fake imitation of navalny's chief of staff. https://nltimes.nl/2021/04/24/dutch-mps-video-conference-deep-fake-imitation-navalnys-chief-staff, 2021.

[45] The Verge. 'deepfake' that supposedly fooled european politicians was just a look-alike, say pranksters. https://www.theverge.com/2021/4/30/22407264/deepfake-european-polticians-leonid-volkov-vovan-lexus, 2021.

[46] Junke Wang, Zuxuan Wu, Jingjing Chen, and Yu-Gang Jiang. M2TR: Multi-modal multi-scale transformers for deepfake detection. *arXiv preprint arXiv:2104.09770*, 2021.

[47] 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.

[48] George Williams, Graham Taylor, Kirill Smolskiy, and Chris Bregler. Body motion analysis for multi-modal identity verification. In *2010 20th International Conference on Pattern Recognition*, pages 2198–2201, 2010.

[49] 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*, 2020.

[50] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2019.

[51] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7556–7566, 2019.

[52] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[53] 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.

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