Detecting Deep-Fake Videos from Aural and Oral Dynamics

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

Abstract

A face-swap deep fake replaces a person's face – from eyebrows to chin – with another face. A lip-sync deep fake replaces a person's mouth region to be consistent with an impersonated or synthesized audio track. An overlooked aspect in the creation of these deep-fake videos is the human ear. Statically, the shape of the human ear has been shown to provide a biometric signal. Dynamically, movement of the mandible (lower jaw) causes changes in the shape of the ear and ear canal. While the facial identity in a face-swap deep fake may accurately depict the co-opted identity, the ears belong to the original identity. While the mouth in a lip-sync deep fake may be well synchronized with the audio, the dynamics of the ear motion will be de-coupled from the mouth and jaw motion. We describe a forensic technique that exploits these static and dynamic aural properties.

1. Introduction

One of the earliest examples of what we now generically call deep-fake videos dates back to 1997 [8]. In this seminal work, video of a person mouthing words she did not speak are synthesized by reordering the mouth region in training video to match specific phonemes in a new audio track (today, we would call this a lip-sync deep fake). The intervening two decades has seen tremendous advances in computer-graphics and -vision based rendering, synthesis, and understanding.

The term "deep fake" emerged in 2017 when a Reddit user named "deepfakes", along with other Reddit users, began using advances in machine learning to digitally insert celebrity faces into sexually explicit material. Over the intervening few years, the sophistication, quality, and ease of generating synthetic audio and video has accelerated leading today to commercially available apps that can be used to more quickly and easily create compelling manipulated video and audio. The general consensus today is, while synthetic video can be entertaining, it can also easily be weaponized in the form of non-consensual pornography, fraud, misinformation, and may lead to a general lack of trust in what we see and hear online.

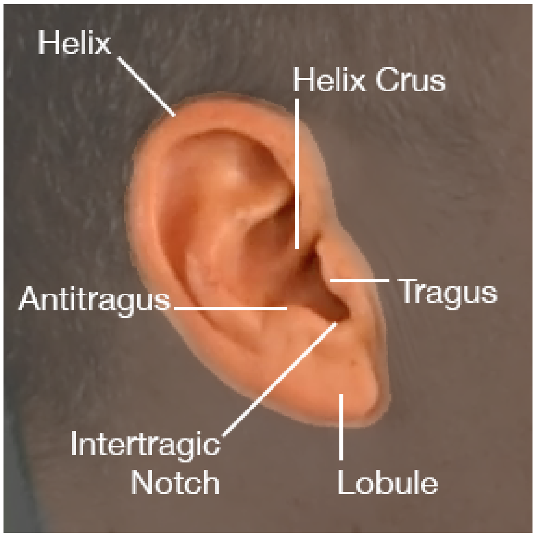

Figure 1. The human ear and a few of its parts.

In response to deep fakes and, more generally, manipulated content, several authentication techniques have emerged that can be roughly categorized into one of three categories:

1. **Forensic Analysis**: based on the assumption that manipulation or synthesis will leave behind some statistical, geometric, or physical artifact, this class of approaches analyzes content for explicit traces of manipulation or synthesis [18]. The benefit of this approach is it can be applied to a broad class of content and requires little to no prior assumptions. The drawback is, to date, most forensic techniques cannot operate at a speed or accuracy for deployment at an internet scale of billions of daily uploads.

2. **Digital Signatures**: this class of approaches tackles the authentication from a different direction, focusing on authenticating content at the point of recording [1].

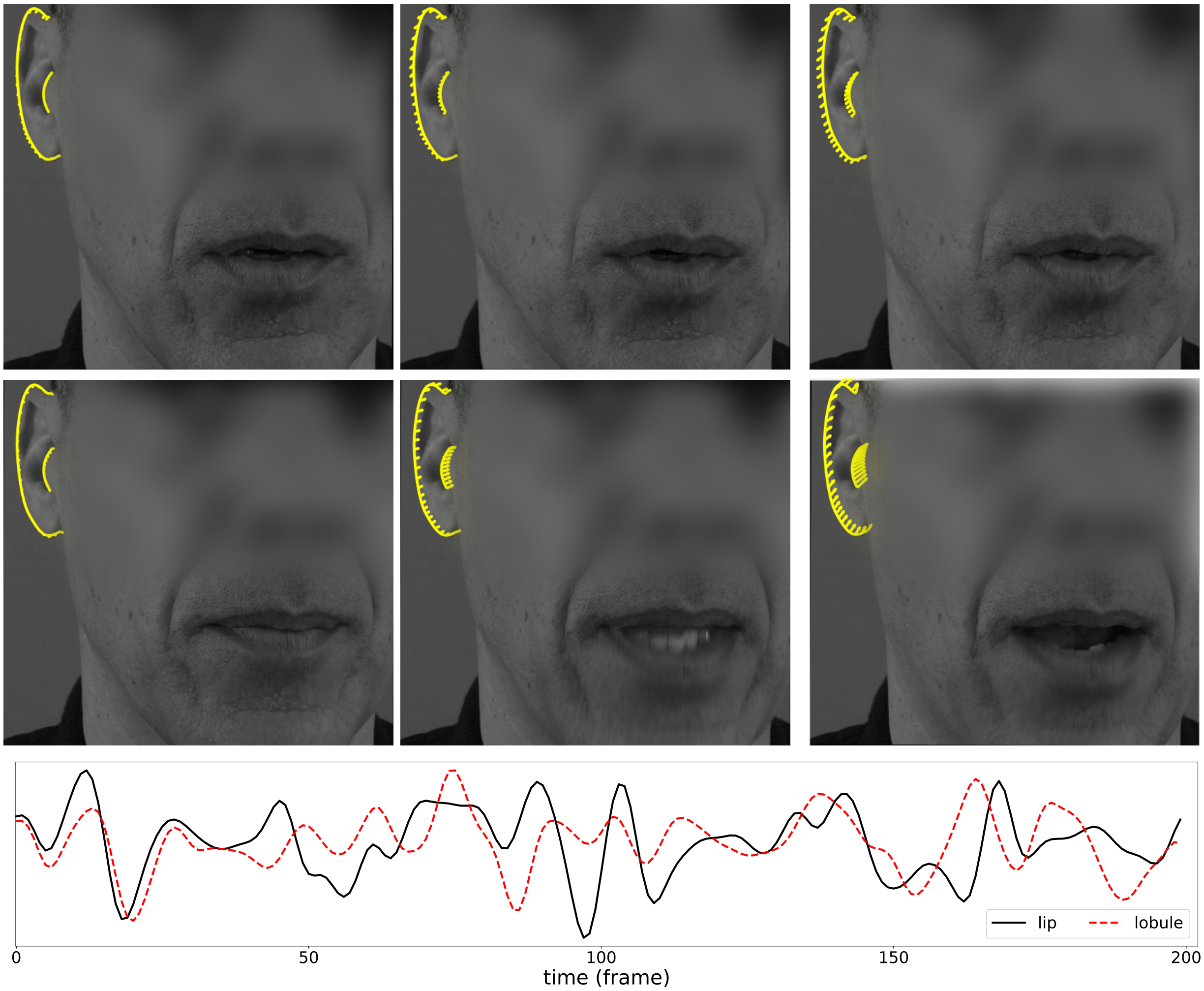

Figure 2. Shown in the first two rows are three equally-spaced frames in which the subject is speaking. Shown in each panel is a tracked Bezier curve corresponding to the ear's helix and lobule (larger outer curve) and tragus (smaller inner curve). The small vectors along each curve correspond to the estimated local motion (scaled by 5x), revealing how the ear moves during speech. Facial expressions such as raised eyebrows, smiling, and surprise induce similar aural motion. Shown in the lower panel is the measured horizontal lobule motion (red, dashed) and vertical lip distance (black, solid), revealing a correlation (r = 0.34) between these two signals.

In either software or hardware, a specialized camera app extracts date/time, geo-location, and pixel data at the point of recording, and hashes and cryptographically signs this data. The resulting digital signature can be used downstream to verify that the content has not been altered from the time of recording, and localize where and when the content was recorded. The benefit of this approach is it can, at internet-scale, verify recorded content quickly and accurately. The drawback of this approach is it requires a specialized camera app and is unable to verify content not recorded through such an app.

3. **Digital Watermarks**: this class of approaches incorporates directly into a synthesis pipeline a digital watermark that can be used downstream to identify deep-fake content [39, 38]. The benefit of this approach is it can quickly and accurately identify deep-fake content. The drawback is it requires a specific infrastructure incorporated into all synthesis pipelines and is vulnerable to attacks designed to remove (or add) watermarks [33].

Here we describe a forensic technique falling into category 1. Most of the focus on creating deep-fake videos has been on facial expressions, the mouth, and audio-video synchronization. The creation of a lip-sync deep fake, for example, requires a detailed synthesis of the mouth region, teeth, and tongue, all the while making sure the mouth is properly synthesized with the audio and spoken

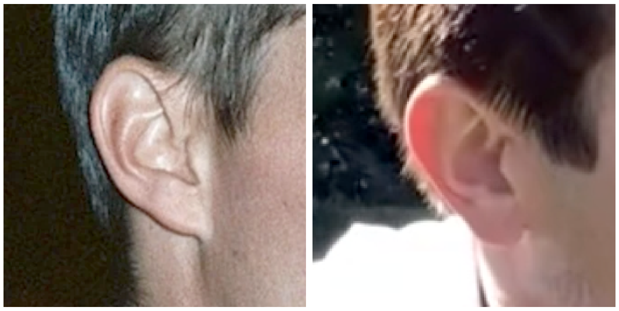

Figure 3. The right ear of the real Tom Cruise (left) and @deeptomcruise of TikTok fame [16] (right), from which we see significant differences to the overall shape and earlobe connectivity to the upper jaw.

phonemes [5]. An overlooked aspect in the creation of these deep-fake videos is the human ear.

The reason for this is probably two-fold. The structure and movement of the human ear is complex, and it is likely our attention is not drawn to a person's ear when they are talking. Eye tracking studies on face perception have consistently revealed a Y-shaped pattern of fixations over the eye, nose and mouth regions [34, 21]. Janik et al. [22] found subjects spend 40% of the time looking at the eyes while free viewing facial photographs.

Both statically – the shape of the human ear provides a biometric signal [9, 10, 32, 3, 17] – and dynamically – movement of the mandible (lower jaw) causes changes in the shape of the ear and ear canal [28, 19, 15] – the human ear provides a rich source of forensic information. Specifically, while the facial identity in a face-swap deep fake may accurately depict the co-opted identity, the ears belong to the original identity. And, while the mouth in a lip-sync deep fake may be well synchronized with the audio, the dynamics of the ear motion will be de-coupled from the mouth and jaw motion. We describe a forensic technique that exploits these static and dynamic aural properties.

In the next section, we place our work in context relative to previous forensic techniques. We then describe our underlying methodology and show the efficacy of our approach across simulated lip-sync deep fakes, production-quality deep fakes, and in-the-wild deep fakes.

2. Related Work

Forensic techniques for detecting deep fakes can be broadly categorized into low- and high-level approaches. Low-level techniques detect pixel-level, synthesis artifacts, including generic artifacts [41, 37, 40, 35], warping artifacts [26], and blending artifacts [24]. These low-level techniques are attractive because they can detect a variety of fakes with relatively high accuracy. The drawback, however, is they can be sensitive to unintentional laundering (e.g., transcoding or resizing) or intentional adversarial attacks (e.g., [12, 11]).

High-level approaches, in contrast, tend to be more resilient to laundering and adversarial attacks. These techniques focus on semantically meaningful features including inconsistencies in eye blinks [25], head-pose [36], physiological signals [13], mouth shape and movement [5], and distinct mannerisms [6, 4]. Because current synthesis techniques are frame-based, incorporating these types of semantic and temporal dynamics is essential to staying ahead of the synthesis-and-detection, cat-and-mouse game.

Biometric identification based on aural features has a well-established literature dating back as far as the 1890s [9, 10, 32, 3, 17]. The general, albeit not unanimous, conclusions today is certain aural features are distinct and stable over a person's lifetime. It remains unclear, however, if these aural features are distinct enough to work on a large scale, and if extraction of these features is sufficiently robust to work in the wild. For our purposes of deep-fake detection, however, the demands of distinctiveness are significantly less than in a biometric setting, and with our focus on video, feature extraction should be more robust than from only a single image.

3. Methods

We describe the aural dynamics methodology and data set, followed by the aural biometric methodology.

3.1. Aural Dynamics

The human ear has three primary sections: the inner- and middle-ear, and the outer-ear consisting of visible features like the lobule, tragus, and helix, Figure 1. Movement in the ear canal – connecting the outer-ear and middle-ear – has been studied in relationship with the movement of the mandible (lower jaw) [28, 19, 15]. Additional studies reveal the middle ear muscles to be responsive to face and head movements, onset of vocalization, yawning, swallowing, coughing, and laughing [30].

We observe such physiological movements in the middle ear can also be observed in movements of the outer-ear's lobule, tragus, and helix, Figure 2. We hypothesize that because deep fakes focus on the synthesis of the face, these aural movements will be absent or disrupted in deep fakes. We next describe techniques for measuring aural motion and correlating this motion to oral signals consisting of facial movements and auditory signals.

We describe the estimation of aural motion in a video in which it is assumed a single person is talking with their left or right ear visible throughout the video segment. This estimation is composed of four parts, as enumerated below.

Face Alignment: For each video frame, 68, 2D facial landmarks are extracted using Dlib [23]. Using these landmarks, the face in each frame is aligned such that the endpoints of

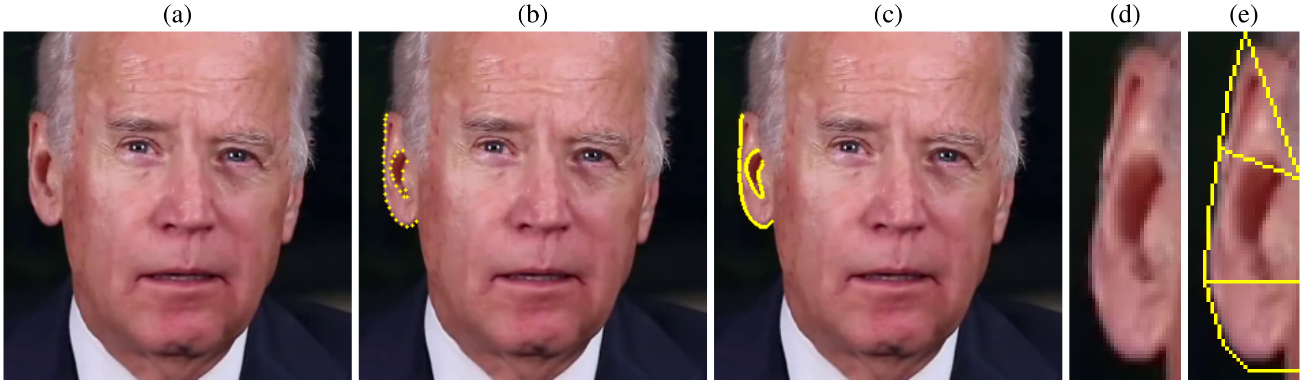

Figure 4. Shown are (a) a single video frame where the face has been tracked, aligned, and cropped; (b) 35 manually annotated aural landmarks; (c) 100 points on each of two Bezier fitted curves; (d) rotated and cropped ear; and (e) three regions from which local aural motion are averaged.

the jaw (landmarks 0 and 16) lie on a horizontal line, are scaled to have a fixed distance of 164 pixels, and translated to a fixed location (pixel locations $(46, 90)$ and $(210, 90)$). After this alignment, a 256×256 pixel region is cropped around the face and ears, Figure 4(a).

Feature Tracking: In order to localize the ears in each frame, we begin by manually annotating 35 aural landmarks on the first aligned video-frame. The first set of 20 landmarks are on the outer portion of the ear, from the helix to the lobule, and the remaining set of 15 landmarks are around the tragus, Figure 4(b). We fit to each of these sets of landmarks, a Bezier curve of order 8 and 10, respectively. A total of 100 points are uniformly sampled from each of these curves, Figure 4(c). Lastly, the 200 Bezier points are tracked across all frames using the Kanade-Lucas-Tomasi (KLT) tracker [31].

Aural Alignment: In order to measure the local aural motion due to facial expressions and speech, we first eliminate the global motion due to head movements. In each frame, the tracked aural landmarks are affine-aligned to the landmarks in the previous frame. Each frame is then rotated such that the axes of the bounding box containing all of the aural landmarks are parallel to the image axes, Figure 4(d).

Motion Estimation: The local aural motion due to facial expressions and speech is estimated using dense optical flow between each consecutive aligned frames. The average 2D motion in the horizontal and vertical directions is computed in three aural regions around the helix, tragus, and lobule, Figure 4(e), yielding a total of six estimated aural motions. We next describe how these motions are correlated to oral signals consisting of facial expressions and speech.

Aural/Oral Correlations: We observe the per-frame aural movements are correlated to the per-frame vertical distance between the lips (i.e., the openness of the mouth) and audio root mean square energy (RMSE) (i.e., the loudness of the speech). While there are other facial and auditory correlations, we focus here on just these two. For each video frame, 68, 3D facial landmarks are estimated using the OpenFace toolkit [7]. The landmarks corresponding to the center of the top and bottom lip (landmarks 51 and 57) are used to compute vertical distance between the lips.

The audio RMSE is measured using the open-source python package LibROSA [27] over a sliding 0.032-second window and a hop length of 0.033 seconds. Given an audio with 16kHz sampling rate and a corresponding video of 30 fps, this yields a single audio RMSE value for each video frame.

The Pearson correlation between the horizontal and vertical aural motion in each of three ear regions and the above two oral signals is computed over a sliding 10-second segment with a 0.033-second shift. This yields a total of 12 correlations per each video segment. Shown in Figure 2 (bottom panel) is a representative example of the measured tragus horizontal movement (red) and lip vertical distance (black), from which the correlation is computed.

3.2. Data Set

A total of 64 videos were downloaded from YouTube of Joe Biden, Angela Merkel, Donald Trump, and Mark Zuckerberg. These videos spanned in length between 12 and 59 seconds, Table 1. We ensured the left or right ear was visible throughout each video. For each frame of each video, we measured the aural motion, the vertical distance between the lips, and the audio RMSE. Due to large head movements, the feature tracking occasionally failed (4-5 times per video) and was corrected by manually re-annotating the necessary features.

In a lip-sync deep fake, the mouth movements of an existing video are modified to match a new audio. We used the following three strategies to generate such lip-sync deep

	video (count)	total (seconds)	minimum (seconds)	maximum (seconds)
Joe Biden	21	769	12	59
Angel Merkel	10	228	12	42
Donald Trump	16	547	16	54
Mark Zuckerberg	17	484	20	29

Table 1. The number of videos in our data set, along with the total, minimum, and maximum video duration for each of four individuals.

fakes: (1) a lip-sync deep fake is simulated by simply correlating the aural movements from one video segment to the oral signal from a randomly selected segment of the same length; (2) visually compelling lip-sync deep fakes were generated for Biden, Merkel, Trump, and Zuckerberg, in which the mouth region is GAN-synthesized to be consistent with a new audio and optimized for visual quality and temporal coherence (courtesy of Kristof Szabo, Zoltan Kovacs, and Dominik Mate Kovacs). A total of six fakes were created for each of the four identities by swapping the original audio with a randomly selected audio from the same individual; and (3) three in-the-wild lip-sync deep fakes were downloaded from YouTube and Instagram, two for Donald Trump [1] and one for Mark Zuckerberg [2].

3.3. Aural Biometrics

The aural dynamics described above are designed to detect lip-sync deep fakes in which the aural and oral signals are desynchronized. In a face-swap deep fake, however, these signals are likely to be consistent with the original speaker. But, in a face-swap deep fake the ears in the video belong to the original identity and not to the person it purports to depict. As a result, we can leverage aural biometrics to verify the true identity in the video. There is a significant literature on aural biometrics including 2D, image-based features [42], 3D model-based features [14], and learned features [29].

Here we adopt a simple approach based on 2D, image-based features which capture the general shape of the ear. Any of a number of other techniques would be equally viable. In our approach, we first manually annotate 20 landmarks equally spaced from the helix to the lobule, and another 15 landmarks equally spaced around the tragus, Figure 7(c). The overall shape of the helix and tragus are characterized using two Bezier curves of order 8 and 10.

The shape of an ear is compared for similarity to a reference ear by first aligning the 35 aural landmarks, as described in the previous section. Because the ear may be imaged from any camera angle, the resulting perspective projection can significantly alter its appearance in the image. We assume, therefore, a reference ear in which the ear is parallel to the imaging plane, thus minimizing any perspective distortion. The comparison ear is then aligned to this reference ear using a planar homography [20] applied to the 35 aural landmarks. Although the ear is not perfectly planar, this homography is reasonable given the relatively small depth change along the ear as compared to a typical distance to the camera.

Once aligned, two ears are compared for similarity by measuring the average Euclidean distance between 100 equally sampled points on each of two Bezier curves and their closest point in the reference ear. This average distance is used as our measure of biometric similarity between two ears.

	Biden	Merkel	Trump	Zuckerberg
training (all)	0.97	0.90	0.93	0.87
simulated	0.97	0.82	0.93	0.87
GAN-generated	0.78	0.90	0.98	0.78
in-the-wild	–	–	0.70	0.71
training (ind)	0.99	0.96	0.99	0.97
simulated	0.96	0.80	0.98	0.86
GAN-generated	0.97	0.85	0.97	0.82
in-the-wild	–	–	0.76	0.77

Table 2. The performance (reported as area under the curve, AUC) for a single model trained on all four individuals (top), and separate models trained on each individual (bottom). All video segments are 10 seconds in length. Results are reported for the training dataset, and three different types of fakes: simulated, GAN-generated, and in-the-wild.

4. Results

For all Biden, Merkel, Trump, and Zuckerberg videos, the distribution of audio, facial, and aural correlations are shown in Figures 5 and 6. These correlations are computed between the horizontal (Figure 5) and vertical (Figure 6) motion in the helix, tragus, or lobule with the facial (lip vertical) or audio (RMSE) signal. Shown in the last row of these figures are the correlations for simulated fakes in which the aural movements from one video segment are paired to the oral signal from a randomly selected video segment.

The nature of the correlations are somewhat person-specific. For horizontal aural motion, for example, the tragus motion is strongly positively correlated with audio for Trump, but weakly negatively correlated for Biden, Merkel, and Zuckerberg. Similarly, the horizontal lobule motion is strongly negatively correlated for Trump, but not for the others. Additionally, the horizontal tragus motion is positively correlated to the lip vertical distance for Biden, Merkel, and Zuckerberg, but not Trump. For vertical aural motion, this basic pattern continues. The tragus motion is

[1] www.instagram.com/p/ByPhCKuF22h/, youtu.be/VWMEDacz3L4

[2] youtu.be/cnUd0TpuoXI

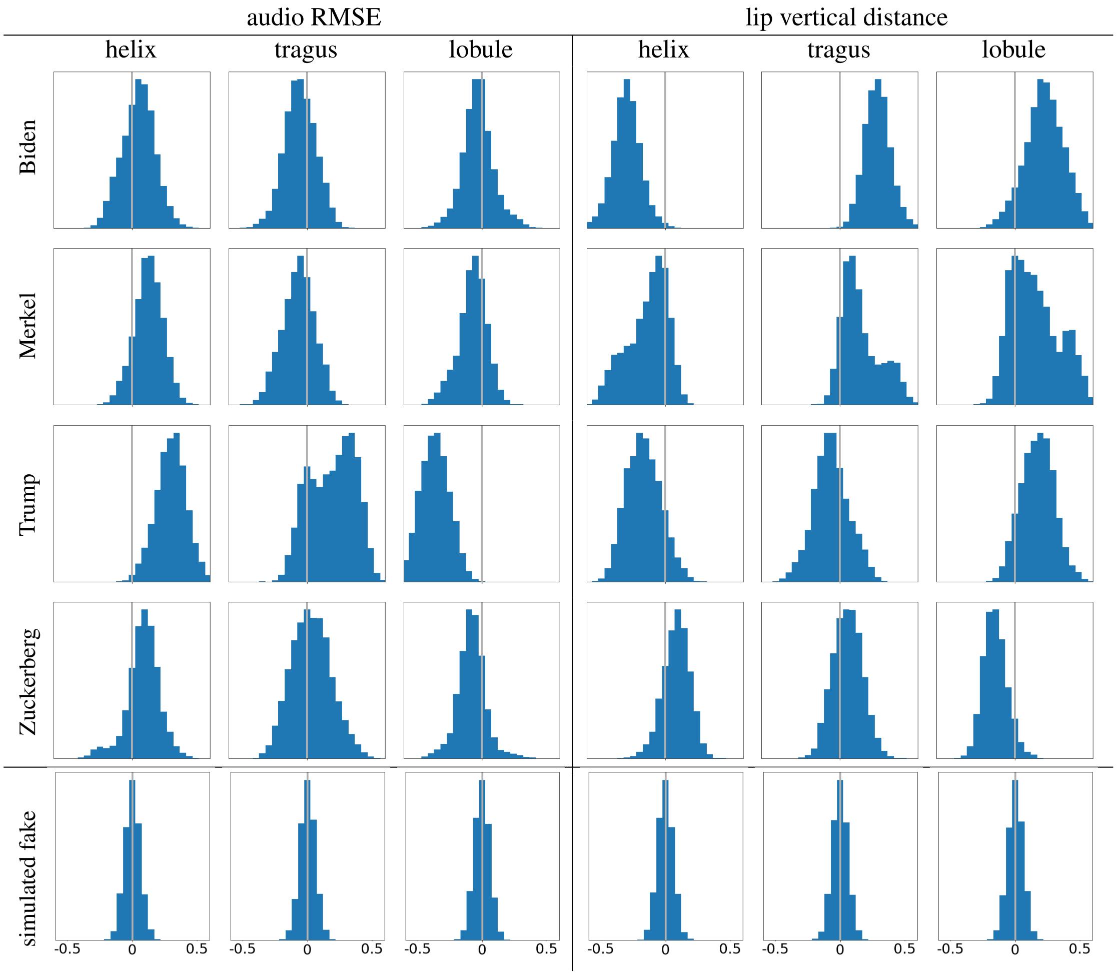

Figure 5. Shown are the distribution of correlations between audio (left) and lip vertical distance (right) and the *horizontal* motion of three aural areas. From top to bottom are the results for four individuals, and simulated fakes. While the fakes have no correlation, we see strong, but not necessarily consistent, correlations across individuals.

strongly negatively correlated with audio for Trump, but not for the others.

By comparison, in all cases, the simulated fakes (last row of Figures 5 and 6), we see a complete lack of correlation between these aural and oral signals.

In order to evaluate the efficacy of these dynamic aural features to detect lip-sync deep fakes, a linear classifier is trained as follows. For each individual, the available videos are split into non-overlapping, $80\%/20\%$ training and testing sets. A logistic regression model is trained on the 12 aural/oral correlations for the original videos and simulated-fake videos. This model was then evaluated on the testing original videos, and all three types of fake videos: simulated, GAN-generated, and in-the-wild.

Shown in Table 2 (top), is the average accuracy reported as the area under the curve (AUC) for 20 random training/testing splits. The average training AUC is 0.91, and the average testing AUC is 0.84, ranging from a low of 0.70 for the Trump in-the-wild fakes, to a high of 0.98 for the GAN-generated Trump fakes. This predictor was trained on all four identities. As we saw in Figures 5 and 6, however, the nature of the correlations is somewhat person-specific.

Shown in Table 2 (bottom) are the results of training four separate logistic regression models, trained on original and fake videos from one individual with, again, 20 random training/testing splits. With this person-specific training, the average training AUC increases from 0.91 to 0.98, and the average testing accuracy increases from 0.84 to 0.87.

Despite only analyzing short 10-second segments, overall accuracy is fairly high. This accuracy can be improved by integrating over an entire video with a simple majority rule.

4.1. Aural Biometrics

We demonstrate the use of our aural shape features on the TikTok viral, deep fake videos of Tom Cruise created

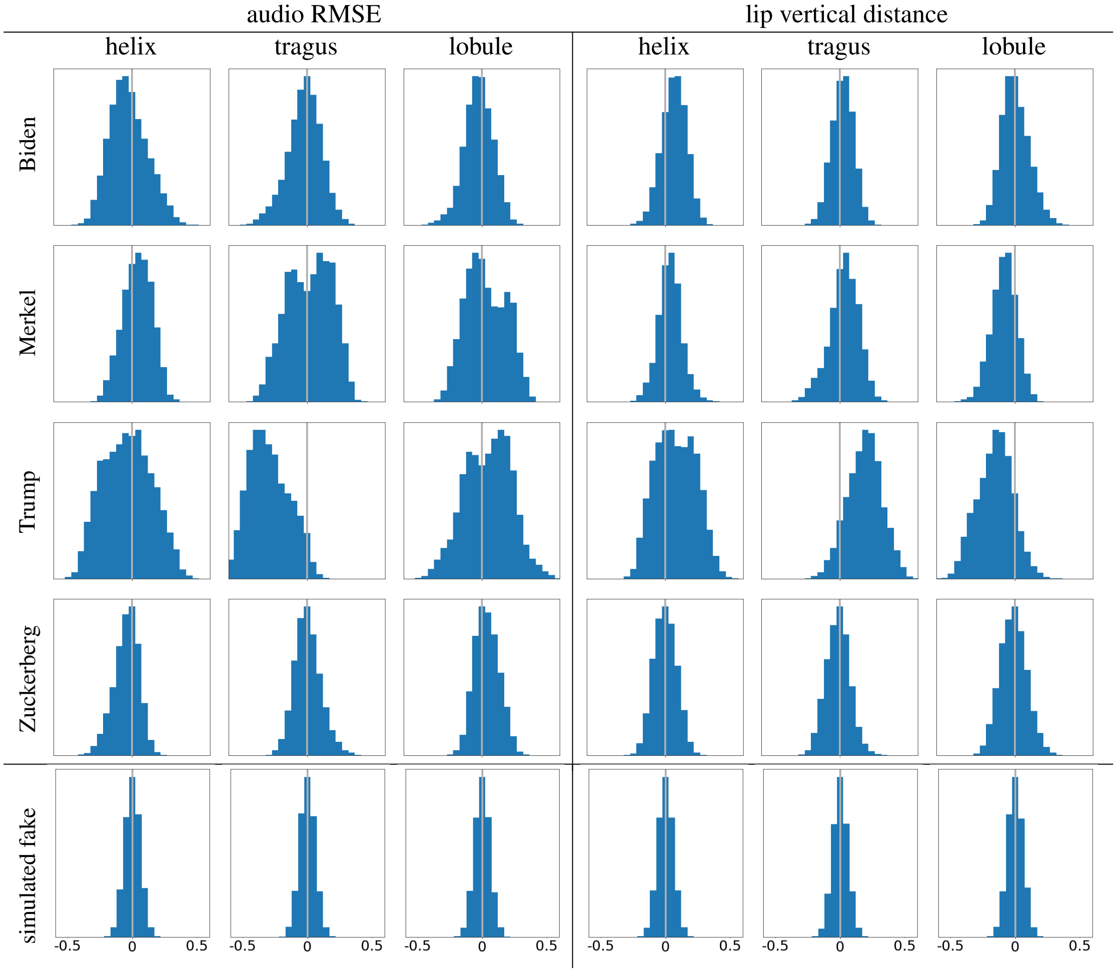

Figure 6. Shown are the distribution of correlations between audio (left) and lip vertical distance (right) and the *vertical* motion of three aural areas. From top to bottom are the results for four individuals, and simulated fakes. While the fakes have no correlation, we see strong, but not necessarily consistent, correlations across individuals.

by @deeptomcruise [16]. We collected 13 images of Cruise from various internet sources where the left or right ear was visible. Although it has been shown that the left and right ears exhibit some symmetry, there also exist some asymmetries [2]. Despite these asymmetries, we compare all ears to a single right-ear reference, Figure 7(c).

Shown in Figure 7 (top), is the comparison of all 12 ears to the reference ear. The average difference, as described in Section 3.3, across all ears is 0.28 with a minimum and maximum difference of 0.19 and 0.37 (these differences are unitless because the aural landmarks are normalized into a range of $[-1, 1]$).

By comparison, shown in Figure 7(a-b) is a comparison between the @deeptomcruise ears and the reference ear. Here, the shape difference is 0.51 and 0.58, more than 35% larger than the largest difference across authentic ears.

5. Discussion

TikTok's @deeptomcruise recently produced what is arguably some of the most compelling and sophisticated deep-fake videos to date [16]. Beyond the excellent face-swap synthesis, these videos also benefit from a talented performer who resembles the real Tom Cruise and is capable of imitating his mannerisms and voice. The impersonator, however, left behind biometric clues to his identity: the shape and structure of the ears and – not discussed here, but worthy of further investigation – distinct characteristics of the hands. Because deep-fake synthesis has understandably focused on the face, these additional biometric signals should prove a useful addition to the forensic analyst's toolkit.

Our methodology of exploiting aural biometrics and aural and oral correlations are, however, not without limitations. Long hair, for example, will impede any measurements of the shape or dynamics of the ear; large head movements make tracking and aural motion estimation challenging; large head movements may bring the ear into and out of view; the static biometric analysis requires a reference ear of the individual in question; and the dynamic aural motion analysis is most effective with video of the individual in question. Lastly, accurate tracking of the ears has proven to be challenging, requiring some human assistance to correct for tracking slippage. Our approach would benefit from more robust tracking.

A benefit of our dynamic aural and oral analysis is the measured signal unfolds over hundreds of frames, whereas current synthesis techniques typically operate on one or only a few video frames. In addition to the two oral correlations explored here (mouth movement and audio), other facial and audio signals can be exploited including raised eyebrows, smiling, frowning, and audio pitch.

More generally, focusing on high-level, soft- and hard-biometric signals such as the ear, hand, mannerisms, and iris provide a rich forensic signal, striking at the heart of all forms of deep fakes that simply don't depict the person they purport to be.

References

[1] Content Authenticity Initiative. https://contentauthenticity.org/. 1

[2] Ayman Abaza and Arun Ross. Towards understanding the symmetry of human ears: A biometric perspective. In *Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 1–7, 2010. 7

[3] Ayman Abaza, Arun Ross, Christina Hebert, Mary Ann F Harrison, and Mark S Nixon. A survey on ear biometrics. *ACM Computing Surveys*, 45(2):1–35, 2013. 3

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

[5] Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 660–661, 2020. 3

[6] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 38–45, 2019. 3

[7] Tadas Baltrušaitis, 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. 4

[8] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: Driving visual speech with audio. In *24th Annual Conference on Computer Graphics and Interactive Techniques*, pages 353–360, 1997. 1

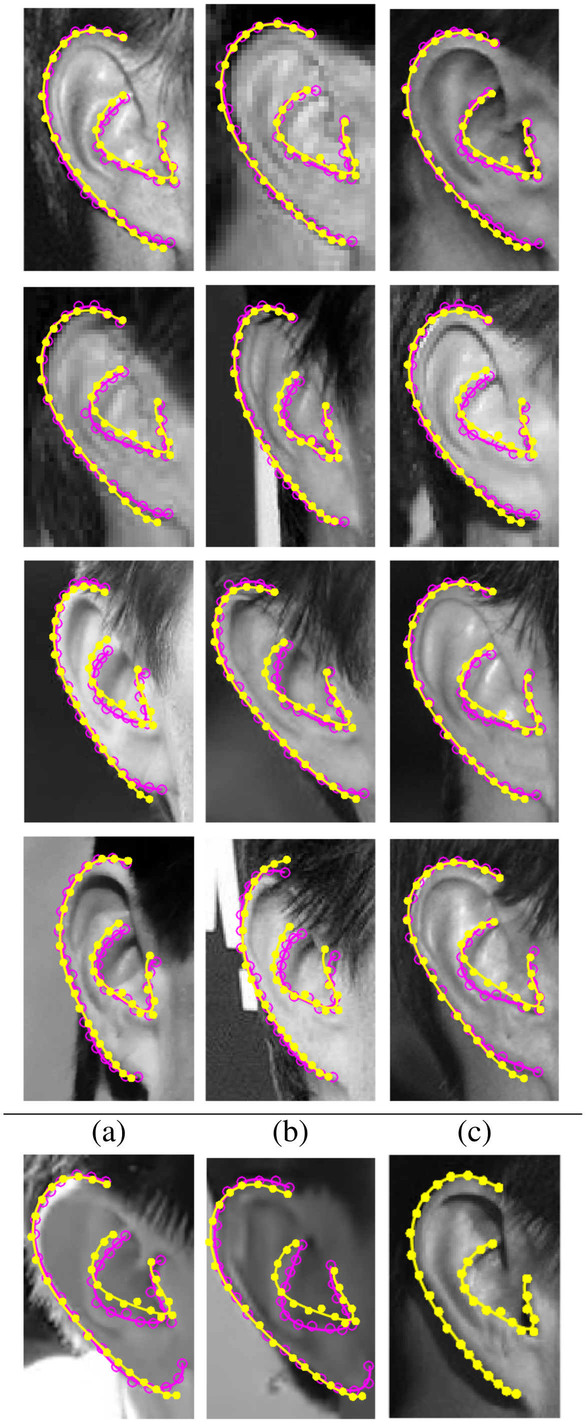

Figure 7. Twelve images of Tom Cruise's ear (top), two images of @deeptomcruise's ear (a-b), and a reference ear for the real Cruise (c). The yellow annotation (filled circle) corresponds to the shape of the reference ear, and the magenta (open circle) corresponds to the comparison ear's shape. A total of 20 landmarks are annotated from the helix to lobule, and 15 along the tragus, to which two Bezier curves are fit.

[9] Mark Burge and Wilhelm Burger. Ear biometrics. In *Biometrics*, pages 273–285. Springer, 1996. 3

[10] Mark Burge and Wilhelm Burger. Ear biometrics in computer vision. In *International Conference on Pattern Recognition*, volume 2, pages 822–826, 2000. 3

[11] Nicholas Carlini and Hany Farid. Evading deepfake-image detectors with white-and black-box attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 658–659, 2020. 3

[12] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. arXiv: 1608.04644, 2016. 3

[13] Umur Aybars Ciftci and Ilke Demir. FakeCatcher: Detection of synthetic portrait videos using biological signals. arXiv: 1901.02212, 2019. 3

[14] Hang Dai, Nick Pears, and William Smith. A data-augmented 3D morphable model of the ear. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 404–408, 2018. 5

[15] Sune Darkner, Rasmus Larsen, and Rasmus R Paulsen. Analysis of deformation of the human ear and canal caused by mandibular movement. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 801–808. Springer, 2007. 3

[16] deeptomcruise. https://www.tiktok.com/@deeptomcruise. 3, 7

[17] Žiga Emeršič, Vitomir Štruc, and Peter Peer. Ear recognition: More than a survey. *Neurocomputing*, 255:26–39, 2017. 3

[18] Hany Farid. *Photo Forensics*. MIT press, 2016. 1

[19] Malcolm J Grenness, Jon Osborn, and W Lee Weller. Mapping ear canal movement using area-based surface matching. *The Journal of the Acoustical Society of America*, 111(2):960–971, 2002. 3

[20] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003. 5

[21] John M Henderson, Carrick C Williams, and Richard J Falk. Eye movements are functional during face learning. *Memory & Cognition*, 33(1):98–106, 2005. 3

[22] Stephen W Janik, A Rodney Wellens, Myron L Goldberg, and Louis F Dell'Osso. Eyes as the center of focus in the visual examination of human faces. *Perceptual and Motor Skills*, 47(3):857–858, 1978. 3

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

[24] 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: 1912.13458, 2019. 3

[25] 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. 3

[26] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. arXiv: 1811.00656, 2018. 3

[27] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *14th Python in Science Conference*, volume 8, pages 18–25, 2015. 4

[28] Robert J Oliveira, Bruce Hammer, Arthur Stillman, John Holm, Catherine Jons, and Robert H Margolis. A look at ear canal changes with jaw motion. *Ear and Hearing*, 13(6):464–466, 1992. 3

[29] Ramar Ahila Priyadharshini, Selvaraj Arivazhagan, and Madakannu Arun. A deep learning approach for person identification using ear biometrics. *Applied Intelligence*, pages 1–12, 2020. 5

[30] Gerhard Salomon and Arnold Starr. Electromyography of middle ear muscles in man during motor activities. *Acta Neurologica Scandinavica*, 39(2):161–168, 1963. 3

[31] Carlo Tomasi and Takeo Kanade. Detection and tracking of point. Technical report, CMU-CS-91-132, Carnegie, Mellon University, 1991. 4

[32] Barnabas Victor, Kevin Bowyer, and Sudeep Sarkar. An evaluation of face and ear biometrics. In *Object Recognition Supported by User Interaction for Service Robots*, volume 1, pages 429–432, 2002. 3

[33] Sviatolsav Voloshynovskiy, Shelby Pereira, Thierry Pun, Joachim J Eggers, and Jonathan K Su. Attacks on digital watermarks: classification, estimation based attacks, and benchmarks. *IEEE Communications Magazine*, 39(8):118–126, 2001. 2

[34] Gail J Walker-Smith, Alastair G Gale, and John M Findlay. Eye movement strategies involved in face perception. *Perception*, 6(3):313–326, 1977. 3

[35] 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. 3

[36] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8261–8265, 2019. 3

[37] 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. 3

[38] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial GAN fingerprints: Rooting deepfake attribution in training data. arXiv: 2007.08457, 2020. 2

[39] Ning Yu, Vladislav Skripniuk, Dingfan Chen, Larry Davis, and Mario Fritz. Responsible disclosure of generative models using scalable fingerprinting. arXiv: 2012.08726, 2021. 2

[40] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in GAN fake images. arXiv: 1907.06515, 2019. 3

[41] 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. 3

[42] Yuxiang Zhou and Stefanos Zaferiou. Deformable models of ears in-the-wild for alignment and recognition. In *12th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 626–633. IEEE, 2017. 5