Detecting Deep Fakes With Mice: Machines vs Biology

Jonathan Saunders (email:saunder@uoregon.edu)
Alex Comerford, (github:@cmfrd)
George Williams (twitter:@cgeorgewilliams)

Introduction

On October 31, 1938, newspapers around the U.S. all reported a “panic” caused by a fake radio broadcast that occurred the day before. It was performed in the style of a “news flash,” an unprecedented style for the time. The broadcast depicted aliens landing in New Jersey, destroying a small town and ravaging its population. Of course, nothing of the sort actually occurred. The broadcast was the legendary and well-orchestrated hoax directed by a young Orson Welles. He hired a team of professional actors to perform a dramatized version of H.G. Welles science fiction work “War Of The Worlds” [1].

It’s useful to draw parallels between those events and the era of disinformation and fake news we live in today [2]. One huge difference is that, now, you don’t need to hire a team of professional actors or rent an expensive sound stage or special effects studio. One can craft high quality realistic fake content with openly available tools for the purposes of creating fake news and spreading disinformation. These tools are constantly evolving, employing techniques such as face swapping, puppet master, lip-sync, and voice cloning [3].

The sophistication and availability of these tools have many officials concerned. A parade of politicians and tech leaders have recently gathered on capital one to warn all of us of the dangers [4]. The concern is that one well-timed and convincing “fake” could launch a sequence of events leading to a catastrophic event - such as the rapid destabilization of a brittle financial market or could even ignite a powder keg military or civic conflict somewhere in the world.

Such an outcome may or may not happen. But, right now, it’s useful to take a step back and ask ourselves- who is fooled by these fakes?

Fake Speech Study

We did a small user study. We asked participants to differentiate between shorts clips of real speech vs fake ones. We used content from the AVSSpoof Challenge for 2019 [5], a bi-yearly competition that aims to fairly evaluate anti-spoofing countermeasures for speech and voice recognition systems. We found that our participants performed fairly well, the median accuracy

was 88%. Compare this to the median accuracy (equal error rate) of 92% for the state of the art algorithms evaluated for that challenge. We will publish the final study results and analysis at the site https://blackhat.deepfakequiz.com after the BlackHat 2019 conference.

So, it appears that both humans and machines hover around 90% accuracy or more. This seems pretty good. But if you consider the millions of items of media and content uploaded to Facebook and Youtube and other services every day, even a low error rate equates to many items getting past human reviewers and automated systems (false negatives) on a daily basis. And possibly worse, many items will be flagged as fake while actually being genuine (false positive.)

In the remainder of this white paper, we will dive deeper into some of these topics. In “Machines”, we will discuss how fakes are generated and discuss some advanced techniques to detect them. In “Biology”, we will explore the neuro-science around fakes. There, we propose a novel detection technique that is neither machine or human.

Machines

The act of forging video and audio of real people to supply a biased narrative is not a new initiative of malicious parties [15][16] and has been practiced for decades [17]. In the past, creating this type of fake content required the use of expensive, specialized equipment, and domain expertise. However, in recent years advances in machine learning have both substantially increased the quality of AI-synthesized fakes and decreased the amount of expertise needed to produce them at scale [18]. As these AI systems become easier to use and more refined, there is a pressing need to develop detection methods.

Spoof Detection via Bispectral Analysis

Detecting audio-based fakes is no easy task. An accurate analytical procedure is needed for automatic detection of audio-based fakes. Through the use of bispectral analysis we can find higher order correlations in audio waveforms and use them as fundamental features to differentiate between synthesized and real speech.

Starting with a raw audio signal, we can calculate the triple autocorrelation (third order cumulant) [27] to get the bispectrum of a signal. After normalizing the bispectrum we receive the bicoherence which includes higher order correlations related to phase coupling which lower order correlations like the power spectrum cannot take into account. Simple decomposable signals without the presence of noise would produce glaring artifacts in the bicoherence which would clearly fingerprint a signal. Since we are working with human and AI synthesized speech we assume that our input data coincides with varying levels of noise. We employ averaging of he bicoherence across multiple segments of the waveforms explicitly in chunks of 32 with overlap of 16 to represent steady estimates of the signal as a whole [13]. The final averaged

bicoherence artifact of a signal is a complex valued matrix which we can compute high level statistical attributes from. Explicitly we use as statistical based features (mean magnitude and mean phase) to use in a classification model to differentiate between real and AI synthesized speech. The specific model we chose to use was an SVM.

Classification

For the scope of this paper we place an SVM in a logistic regression based framework with the goal to differentiate between human and AI synthesized speech. Our dataset included collectively over 1800 samples from the LJ speech dataset including two different AI speech synthesizers (DC-TTS [29], Tacotron2 [22]) uttering the same phrases as human speakers. We employed a grid search hyperparameter optimization scheme (C=[1,10], gamma=[0.1,1]) to produce a robust model with high accuracy.

Using only two features from the bicoherence (mean magnitude and mean phase) we achieve 95% accuracy and 0.94 f1 score. This evidence shows AI speech synthesis systems are susceptible to detection via the features produced by latent bicoherence artifacts differing in human speech.

This technique is not immune to potential countermeasures that can incorporate similarities in bicoherence artifacts produced by humans. We do show however that through forensic analysis we can detect AI synthesized speech with high accuracy.

To reproduce the results we found in using machines to detect AI synthesized speech, the reader can run the analysis code at
https://github.com/cmrfrd/detecting-deep-fakes-blackhat2019/

Biology

Speech Perception and Spoof Detection

Current methods for detecting spoofed or faked speech are agnostic to the biological bases of production and perception of speech, instead relying on a general search for features and algorithmic architectures that render the greatest empirical spoof detection (19, 20). In the case of the most naive, data-only search, detection solutions are guaranteed to be particular to the design of the spoof generating algorithms that the training dataset comprises. For example - techniques that exploit the unnatural patterns of phase or high-frequency distortions in spoofed speech are likely successful only because most generation algorithms are conditioned only on (real valued) mel-frequency cepstrum coefficients (MFCCs) [13, 21, 22, 23].

While there is no *a priori* reason to believe that detecting spoofs will always be possible (ie. there is no reason to believe perfect fakes are impossible), more general spoof-detection algorithms could make better use of the a) acoustic constraints on real speech imposed by the human articulatory system, and b) perceptual heuristics used by the auditory system that make spoofed speech successful in the first place.

Coarticulation as a potential detection target

Speech is not just extraordinarily variable, but rather there is a fundamental *lack of invariance* in the relationship between acoustic cues and perceived phoneme[24]. Sources of variation include speaker anatomy, prosodic content, rate of speech, accent, environmental noise, among many others. One crucial source of lack of invariance in the acoustic signal of speech is coarticulation[25]. Because the articulators cannot move instantaneously, at normal rates of speech there is always some residual influence of the prior phoneme, as well as anticipatory movement to the following phoneme, on the articulatory/acoustical structure of a given phoneme.

We perceive coarticulated phonemes with different contexts as nearly-identical despite the lack of similarity in their acoustic structure. As a result, spoofed speech that is produced phoneme-by-phoneme can produce realistic sounding speech despite not respecting the coarticulatory constraints of real speech production. Because these faked 'out of order' phonemes could appear spectrally identical to real phonemes, they would not be detected by a spoof-detection algorithm that does not explicitly account for the physical nature of the articulatory system. Because the number of even 3-phoneme transitions are much higher than the number of individual phonemes, they are accordingly much more difficult to model -- and thus easier to detect, particularly in the limited data, transfer learning based spoof scenario.

Mice as a model for studying Phonetic Perception

For further detail on our work training mice to discriminate between phonetic categories, see our paper and repository (which also contains a freely accessible version of our paper as well as all data and analysis code):

- **Paper:** Saunders and Wehr [2019] [26]
- **Repository:** https://github.com/wehr-lab/SaundersWehr-JASA2019

Briefly, we believe that mice are a promising model to study complex sound processing. In this instance, because of our observation that mice appear to be able to learn complex acoustic distinctions that are reflective of their training sets, we believe that studying the computational mechanisms by which the mammalian auditory system detects fake audio could inform next-

generation, generalizable algorithms for spoof detection. Such experiments are impossible in humans because of a) a lifetime of exposure to natural speech and the according phonetic processing heuristics, and b) a lack of experimental instrumentation with the millisecond and micrometer spatiotemporal resolution.

Our mechanistic understanding of how the human auditory system is capable of normalizing to novel speakers, rates, contexts, etc. is limited, but should be the primary point of biological emulation for artificial neural networks designed to detect spoofed speech. By learning how the auditory system rapidly adapts to acoustic properties of a particular speaker, we may be able to design spoof-detection algorithms that are much more finely tuned to acoustic violations of a particular speaker's voice.

References

1. The War of the Worlds radio drama, https://en.wikipedia.org/wiki/The_War_of_the_Worlds_(1938_radio_drama)
2. "Orson Welles and the Birth of Fake News," https://www.nytimes.com/2018/10/30/opinion/orson-welles-war-of-the-worlds-fake-news.html
3. "Deep Fake," https://en.wikipedia.org/wiki/Deepfake
4. House Intelligence Committee on Deep Fakes, "https://www.c-span.org/video/?461679-1/house-intelligence-committee-hearing-deepfake-videos"
5. AVSSpoof 2019, https://www.asvspoof.org/
6. Suwajanakorn, Supasorn, et al. "Synthesizing Obama." *ACM Transactions on Graphics*, vol. 36, no. 4, 2017, pp. 1–13., doi:10.1145/3072959.3073640.
7. Do Nhu, Tai & Na, In & Kim, S.H.. (2018). Forensics Face Detection From GANs Using Convolutional Neural Network.
8. Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio" arXiv:1609.03499 [cs], Sep. 2016.
9. Chris Donahue, Julian McAuley, Miller Puckette, "Adversarial Audio Synthesis" arXiv:1802.04208v3 [cs] Feb. 2019
10. Shan Yang, Lei Xie, Xiao Chen, Xiao Lou, Xuan Zhu, Dongyan Huang, Haizhou Li, "Statistical Parametric Speech Using Generative Adversarial Networks Under A Multi-Task Learning Framework" arXiv:1707.01670v2 [cs] Jul. 2017
11. Generative Adversarial Networks "Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio" 1406.2661 [cs] Jun. 2014
12. Marc Schröder. Interpolating Expessions in Unit Selection. In *Proc. 2nd ACII*, Lisbon, Portugal, 2007
13. Albadawy, Ehab & Lyu, Siwei & Farid, Hany. (2019). Detecting AI-Synthesized Speech Using Bispectral Analysis.

14. Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. arXiv:1808.07371 2018
15. Take one part Kerry, one part Fonda https://www.theguardian.com/media/2004/feb/18/newmedia.uselections2004
16. Donald Trumps parents wore Ku Klux Klan attire? https://www.snopes.com/fact-check/donald-trumps-parents-kkk/
17. Photo tampering throughout history https://www.cc.gatech.edu/~beki/cs4001/history.pdf
18. DeepFaceLab https://github.com/iperov/DeepFaceLab
19. Massimiliano Todisco, Héctor Delgado, and Nicholas Evans. Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. Computer Speech & Language, 45:516–535, September 2017. ISSN 08852308. doi: 10.1016/j.csl.2017.01.001. URL https: //linkinghub.elsevier.com/retrieve/pii/S0885230816303114.
20. Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. Spoofing and countermeasures for speaker verification: A survey. Speech Communication, 66:130–153, February 2015. ISSN 0167-6393. doi: 10.1016/j.specom.2014.10.005. URL http://www.sciencedirect.com/science/article/pii/S0167639314000788.
21. Zhizheng Wu, Eng Siong Chng, and Haizhou Li. Detecting Converted Speech and Natural Speech for anti-Spoofing Attack in Speaker Recognition. page 4, 2012
22. Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgian- nakis, and Yonghui Wu. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. arXiv:1712.05884 [cs], December 2017. URL http://arxiv.org/abs/1712.05884.
23. Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. arXiv:1806.04558 [cs, eess], June 2018. URL http://arxiv.org/abs/1806.04558.
24. Lori L. Holt and Andrew J. Lotto. Speech perception as categorization. Attention, Perception, & Psychophysics, 72(5):1218–1227, July 2010. ISSN 1943-393X. doi: 10.3758/APP.72.5.1218. URLhttps://doi.org/10.3758/APP.72.5.1218
25. E. Farnetani. V-C-V Lingual Coarticulation and Its Spatiotemporal Domain. In William J. Hardcastle and Alain Marchal, editors, Speech Production and Speech Modelling, pages 93–130. Springer Netherlands, Dordrecht, 1990. ISBN 978-94-009-2037-8. doi: 10.1007/978-94-009-2037-8_5. URLhttps://doi.org/10.1007/978-94-009-2037-8_5.
26. Jonny L. Saunders and Michael Wehr. Mice can learn phonetic categories. The Journal of the Acoustical Society of America, 145(3):1168–1177, March 2019. ISSN 0001-4966. doi: 10.1121/1.5091776. URLhttps://asa.scitation.org/doi/abs/10.1121/1.5091776.
27. T. S. Rao, M. M. Gabr, An Introduction to Bispectral Analysis and Bilinear TimeSeries Models, Lecture Notes in Statistics, Volume 24, D. Brillinger, S. Fienberg, J. Gani, J. Hartigan, K. Krickeberg, Editors, Springer-Verlag, New York, NY, 1984.
28. LJ dataset https://keithito.com/LJ-Speech-Dataset/
29. Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention Hideyuki Tachibana, Katsuya Uenoyama, Shunsuke Aihara

arXiv:1712.05884 [cs], October 2017. URL https://arxiv.org/abs/1710.08969