Detecting AI-Synthesized Speech Using Bispectral Analysis

Ehab A. AlBadawy and Siwei Lyu
University at Albany, SUNY
Albany NY, USA
{ealbadawy, slyu}@albany.edu

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

Abstract

From speech to images, and videos, advances in machine learning have led to dramatic improvements in the quality and realism of so-called AI-synthesized content. While there are many exciting and interesting applications, this type of content can also be used to create convincing and dangerous fakes. We seek to develop forensic techniques that can distinguish a real human voice from synthesized voice. We observe that deep neural networks used to synthesize speech introduce specific and unusual spectral correlations not typically found in human speech. Although not necessarily audible, these correlations can be measured using tools from bispectral analysis and used to distinguish human from synthesized speech.

1. Introduction

Recent advances in AI-synthesized content-generation are leading to the creation of highly realistic audio [11, 4], image [6, 5], and video [10, 7, 14, 13, 1]. While there are many interesting and artistic applications for this type of synthesized content, these same techniques can also be weaponized to, for example, create a video of a world leader threatening another nation leading to an international crisis, or a video of a presidential candidate saying something inappropriate which, if released 24 hours before an election, could lead to interference with a democratic election, or a video of a CEO privately claiming that her company's profits are down leading to global stock manipulation. Advances in deep learning have led to the development of synthesis tools for creating the video and audio that can create these types of fakes.

As these synthesis tools become more powerful and readily available, there is a growing need to develop forensic techniques to detect the resulting synthesized content. We describe a technique for distinguishing human speech from synthesized speech that leverages higher-order spectral correlations revealed by bispectral analysis. We show that these correlations are not present in a wide variety of recorded human speech, but are present in speech synthesized with several state of the art AI systems. We also show that these correlations are likely the result of fundamental properties of the synthesis process, which would be difficult to eliminate as a counter measure.

In the general area of audio forensics, there are a number of techniques for detecting various forms of audio spoofing [15]. These techniques, however, do not explicitly address the detection of synthesized speech. Previous work [3] showed that certain forms of audio tampering can introduce the same type of higher-order artifacts that we exploit here. This previous work, however, did not address the issue of synthesized content.

In comparing different features and techniques for synthetic-speech detection, the authors in [12] found that features based on high-frequency spectral magnitudes and phases are most effective for distinguishing human from synthesized speech. These features are based on first-order Fourier coefficients or their second-order power spectrum correlations. In contrast to these first- and second-order spectral features – which might be easy to adjust to match human speech – we explore higher-order polyspectral features which are both discriminating and should prove to be more difficult to adjust by the synthesizer.

2. Methods

We begin by describing the data set of human and synthesized content that we recorded and created. We then describe the polyspectral analysis tools that underlie our technique followed by a qualitative assessment of the differences in the bispectral properties of human and synthesized content. We conclude this section with a description of a simple classifier that characterizes these differences for the purposes of automatically distinguishing between human and synthesized speech.

2.1. Data set

We collected a data set consisting of $1,845$ human and synthesized speech recordings. The human speech are obtained from nine people (five male and four female). These recordings were extracted from various high-quality podcasts. Each recording averaged 10.5 seconds in length.

The same texts spoken by the human subjects (transcribed from the recordings) were used to synthesize audio samples using various automatic text-to-speech synthesis methods including Amazon Polly, Apple text-to-speech, Baidu DeepVoice, and Google WaveNet[1]. We also include samples generated using the Lyrebird.ai API, which, unlike other synthesis methods, generates personalized speech styles (because of limited access to this API, the texts spoken were not matched to the human and other synthesized speech). In synthesizing these recordings, a range of speaker profiles was selected to increase the diversity of the synthesized voices.

2.2. Bispectral Analysis

In this section, we describe the basic statistical tools used to analyze audio recordings. The bispectrum of a signal represents higher-order correlations in the Fourier domain.

An audio signal $y(k)$ is first decomposed according to the Fourier transform:

$$Y(\omega) \;=\; \sum_{k=-\infty}^{\infty} y(k)e^{-ik\omega}, \qquad (1)$$

with $\omega \in [-\pi, \pi]$. It is common practice to use the power spectrum of the signal $P(\omega)$ to detect the presence of second-order correlations, which is defined as:

$$P(\omega) \;=\; Y(\omega)Y^*(\omega), \qquad (2)$$

where $*$ denotes complex conjugate. The power spectrum is blind to higher-order correlations, which are of primary interest to us. These correlations can, however, be detected by turning to higher-order spectral analysis [9]. The bispectrum, for example, is used to detect the presence of third-order correlations:

$$B(\omega_1,\omega_2) \;=\; Y(\omega_1)Y(\omega_2)Y^*(\omega_1+\omega_2). \qquad (3)$$

Unlike the power spectrum, the bispectral response reveals correlations between the triple of harmonics $[\omega_1, \omega_1, \omega_1 + \omega_1]$, $[\omega_2, \omega_2, \omega_2+\omega_2]$, $[\omega_1, \omega_2, \omega_1+\omega_2]$, and $[\omega_1, -\omega_2, \omega_1 - \omega_2]$. Note that, unlike the power spectrum, the bispectrum in Equation (3) is a complex-valued quantity. From an interpretive stance it will be convenient to express the complex bispectrum with respect to its magnitude:

$$|B(\omega_1,\omega_2)| \;=\; |Y(\omega_1)|\cdot|Y(\omega_2)|\cdot|Y(\omega_1+\omega_2)|, \qquad (4)$$

and phase:

$$\angle B(\omega_1,\omega_2) \;=\; \angle Y(\omega_1) + \angle Y(\omega_2) - \angle Y(\omega_1+\omega_2). \qquad (5)$$

Also from an interpretive stance it is helpful to work with the normalized bispectrum [2], the bicoherence:

$$B_c(\omega_1,\omega_2) \;=\; \frac{Y(\omega_1)Y(\omega_2)Y^*(\omega_1+\omega_2)}{\sqrt{|Y(\omega_1)Y(\omega_2)|^2|Y(\omega_1+\omega_2)|^2}}. \qquad (6)$$

This normalized bispectrum yields magnitudes in the range $[0, 1]$. Throughout, we compute the bicoherence with a segment length of $N = 64$ with an overlap of 32 samples.

In the absence of noise, the bicoherence can be estimated from a single realization as in Equation (6). However in the presence of noise some form of averaging is required to ensure stable estimates. A common form of averaging is to divide the signal into multiple segments. For example the signal $y(n)$ with $n \in [1, N]$ can be divided into K segments of length $M = N/K$, or K overlapping segments with $M > N/K$. The bicoherence is then estimated from the average of each segment's bicoherence spectrum:

$$\hat{B}_c(\omega_1,\omega_2) = \frac{\frac{1}{K}\sum_k Y_k(\omega_1)Y_k(\omega_2)Y_k^*(\omega_1+\omega_2)}{\sqrt{\frac{1}{K}\sum_k |Y_k(\omega_1)Y_k(\omega_2)|^2 \frac{1}{K}\sum_k |Y_k(\omega_1+\omega_2)|^2}}. \qquad (7)$$

2.3. Bispectral Artifacts

Shown in Figure 1 is the bicoherent magnitude and phase for three different human speakers. Shown in the second to the sixth rows are the bicoherent magnitude and phase for five different synthesized voices, as described in Section 2.1. Each bicoherent magnitude and phase panel are displayed on the same intensity scale. At first glance, there are some glaring differences in the bicoherent magnitude (with the exception of Apple) between the human and synthesized speech. There are also strong differences in the bicoherent phases across all synthesized speech.

As most of the synthesis methods use certain types of deep neural networks as underlying model, we hypothesize that these bicoherence differences are due to the underlying speech-synthesis network architecture and, in particular, that long-range temporal connections give rise to the unusual spectral correlations. To determine if this might be the case, we created three "clipped" WaveNet network architectures in which the network connectivity was effectively reduced. This was done by first noticing that WaveNet employs 3-tap filters in its convolutional layers. We, therefore, truncate the full WaveNet models in which the left-most value of the convolution filter in one of three layers was fixed at a value of zero[2]. With a total of 24 convolutional

[1] Sources: Amazon Polly aws.amazon.com/polly/, Apple text-to-speech API developer.apple.com/documentation/appkit/nsspeechsynthesizer, Baidu DeepVoice r9y9.github.io/deepvoice3_pytorch/, and Google WaveNet r9y9.github.io/wavenet_vocoder/.

[2] A more direct approach is to use simply use a 2-tap filter. This, however, would require retraining the entire model and so we adopted the simpler approach of zeroing out one of the filter values.

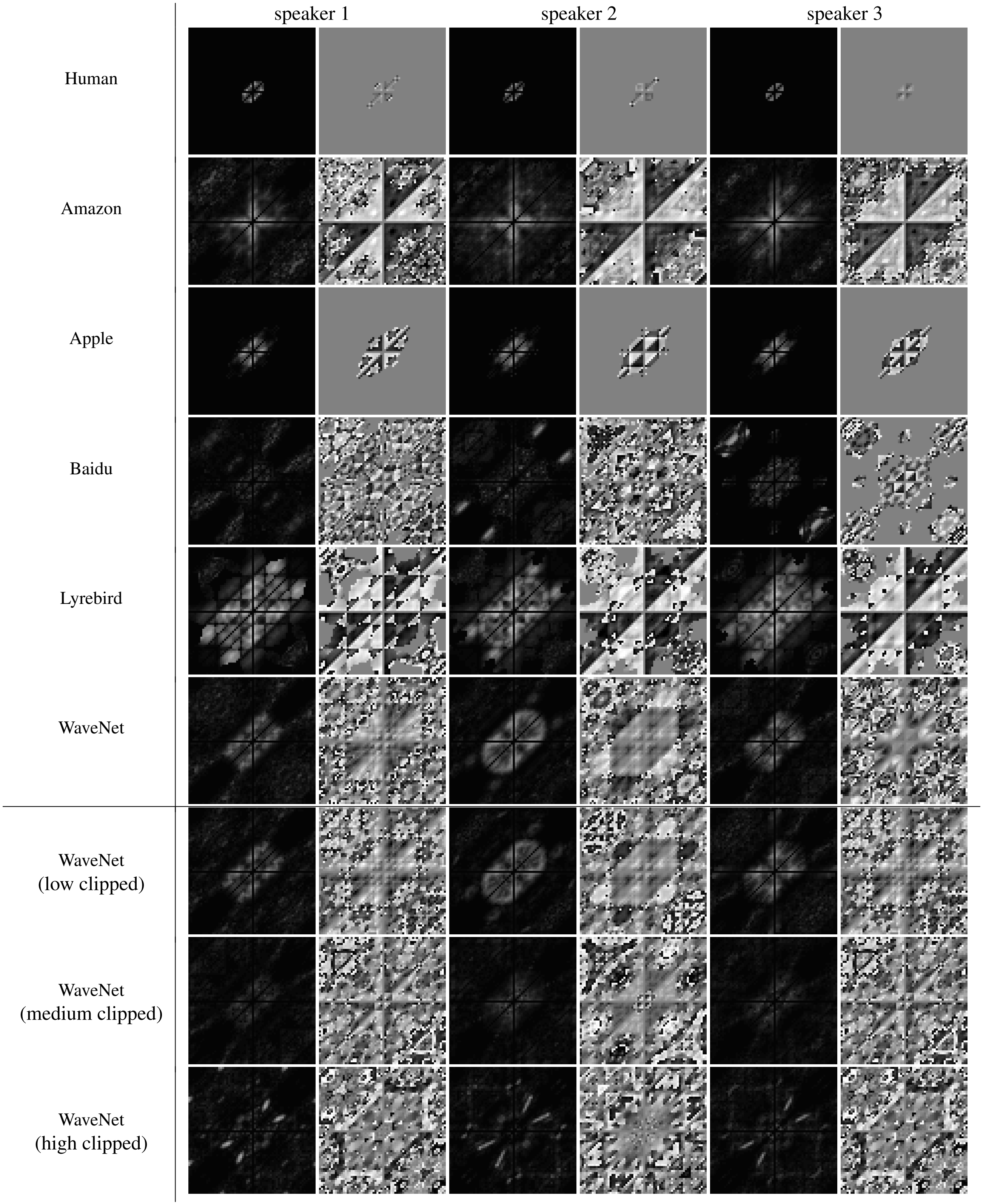

Figure 1. Bicoherent magnitude and phase for three human speakers and five synthesized voices. Shown in the lower three rows are the results for three different clipped versions of the WaveNet architecture. The magnitude plots are displayed on an intensity scale of $[0, 1]$ and the phase plots are displayed on a scale of $[-\pi, \pi]$. Note the generally larger magnitudes and the stronger phase correlations in the synthesized speech as compared to the human speech, and the reduction in magnitude for the clipped WaveNet architectures.

layers we performed this manipulation at level 24 (closest to the output level), 12, or 1 (closest to the input level). The effective network clipping was more pronounced for the lowest level manipulations.

Shown in the last three rows of Figure 3 are the resulting bicoherence magnitudes and phases for three recordings synthesized with these three networks with increasing amounts of "clipping". As can be clearly seen, the bicoherence magnitude reduces with an increasing reduction in network connectivity, and begins to appear more like the human speakers in the first row of Figure 3. At the same time, there is little impact on the bicoherence phase, most likely because our network manipulation did not remove all of the long-range connections. Although this does not prove that the network architecture is solely responsible for the increased bicoherence properties, it provides preliminary evidence to suggest that this is the case. We note that the artifacts from Apple are more subdued than others. This may be related to the fact that the quality of speech is significantly less realistic than Google and Amazon, possibly because the underlying technique is not based on the same type of network architecture that we believe is introducing the polyspectral correlations.

Regardless of precisely why these correlations are introduced, we next show that the bicoherence differences can be used to automatically distinguish between human and synthesized speeches.

2.4. Bispectral Classification

The bicohernece, Equation (7), is computed for each human and synthesized speeches, from which the bicoherence magnitude and phase are computed. These two-dimensional quantities are normalized such that the magnitude and phase for each frequency ω_1 are normalized into the range $[0, 1]$ by subtracting the minimum value and dividing by the resulting maximum value.

The normalized magnitude and phase are each characterized using the first four statistical moments. Let the random variable M and P denote the underlying distribution for the bicoherence magnitude and phase. The first four statistical moments are given by:

- mean, $\mu_X = E_X[X]$
- variance, $\sigma_X = E_X[(X - \mu_X)^2]$
- skewness, $\gamma_X = E_X\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^3\right]$
- kurtosis, $\kappa_X = E_X\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^4\right]$

where $E_X[\cdot]$ is the expected-value operator with regards to random variable X. From the magnitude $X = M$ and phase $X = P$, these four moments are estimated by replacing the expected-value operator with an average. With this statistical characterization, each recording is reduced to an 8-D feature vector.

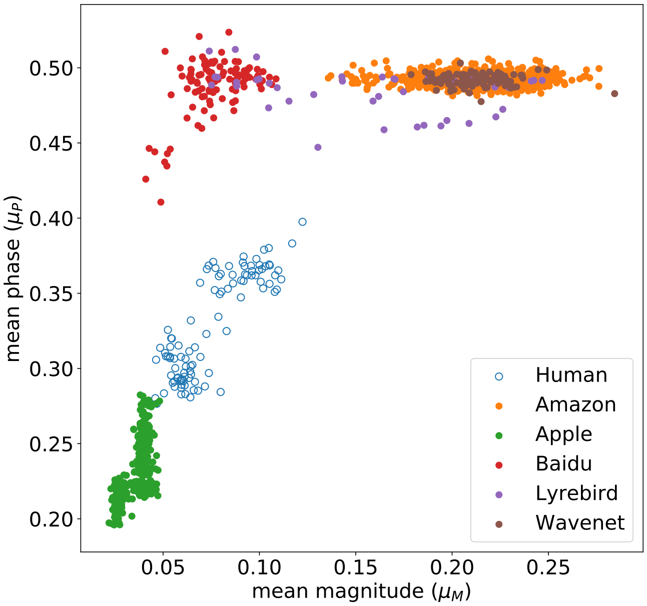

Figure 2. A 2-D slice of the full 8-D statistical characterization of the bicoherence magnitude and phase. The open blue circles correspond to human speech and the remaining filled colored circles correspond to synthesized speech. Even in this reduced dimensional space, the human speech is clearly distinct from the synthesized speech.

Shown in Figure 2 is a scatter plot of the mean bicoherence magnitude versus the mean bicoherence phase for the human speech and each type of synthesized speech. This figure illustrates some interesting aspects of the bicoherence statistics of the human and synthesized recordings. Even in this reduced-dimensional space that does not account for variance, skewness, or kurtosis, each type of signal is well clustered and (with the exception of Amazon and WaveNet) distinct from the other types. This suggests that it will be relatively straight-forward to distinguish between these different recordings.

Also shown in Figure 2 are six speech samples synthesized with a more recent generative adversary network (GAN) based model [8][3]. Although the GAN-based model has a different synthesis mechanism, the synthesized contents still exhibit distinct bispectral statistics.

The scatter plot in Figure 2 suggests two possible approaches to building a classifier. A one-class non-linear support vector machine (SVM) or a collection of linear classifiers. We, primarily for simplicity, choose the latter. In particular, we train a linear classifier to distinguish each category of recording – human, Amazon, Apple, Baidu,

[3] There is no code publicly available and the six samples were downloaded from fangfm.github.io/crosslingualvc.html.

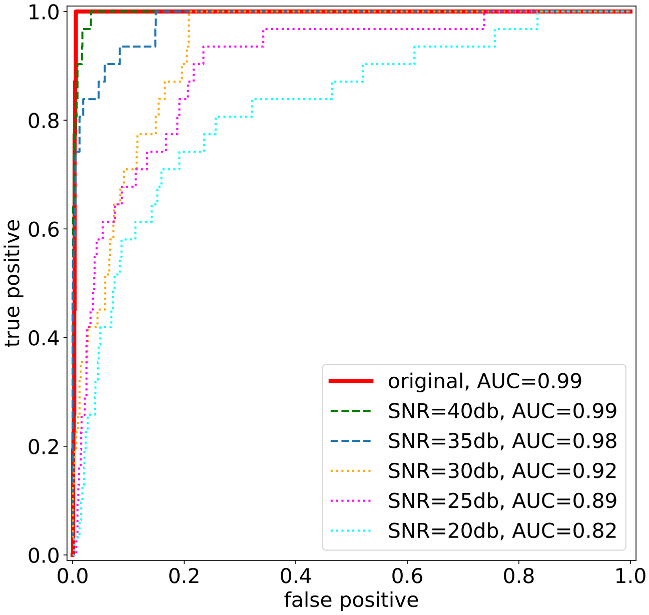

Figure 3. ROC curve for binary classification of human versus synthetic speech (solid red line). The dashed and dotted lines correspond to the accuracies for these same recordings with varying amounts of additive noise. See also Figure 4.

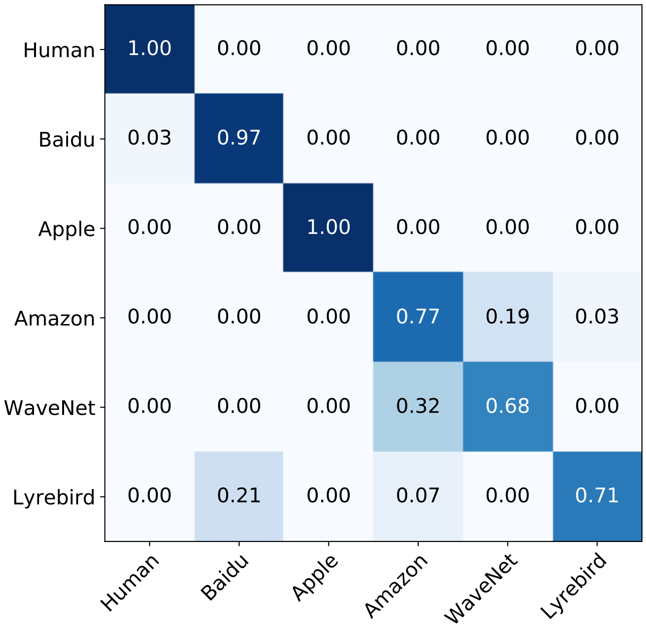

Figure 4. Confusion matrix for classifying a recording as human or as synthesized by one of five techniques. See also Figure 3.

Google, and Lyrebird – from all other recordings. Following this strategy, five separate logistic regression classifiers are trained to distinguish each synthesized audio from all other categories. For example, the first classifier is trained to distinguish Amazon recordings from Apple, Baidu, Google, Lyrebird, and human recordings. Our full data set consists of 100 human recordings, and 800 Amazon (8 speaker profiles), 400 Apple (4 speaker profiles), 100 Baidu (1 speaker profile), 400 Google (4 speaker profiles), and 45 Lyrebird recordings (5 recordings for each of 9 speaker profiles). Because of the across class imbalance, the training data set consisted of 70% of these samples with a maximum of 90 samples per category, with the remaining data used for testing.

The logistic regression classifier is implemented using `scikit-learn`[4]. At testing, a speech sample is classified by each classifier (Amazon, Apple, Baidu, Google, and Lyrebird). If the maximum classification score across all five classifiers is above a specified threshold, then the recording is classified as synthesized, otherwise it is classified as human.

3. Results

We test the performance of distinguishing human speech from synthesized speech based on the 8-D summary bicoherence statistics. Shown in Figure 3 are the receiver operator characteristic (ROC) curves for this binary classification. The solid curve with an area under the curve (AUC) of 0.99 corresponds to the original quality recordings. The remaining dashed/dotted colored curves correspond to the recordings that were laundered with varying amounts of additive noise (with a signal-to-noise ratio (SNR) between 20 and 40 db) followed by re-compression at a quality of 128 kilobits per second (kbit/s). At high SNR, the AUC remains above 0.98, and the AUC decreases with increasing amounts of additive noise.

When the original recordings are recompressed at a lower quality of 64 kbit/s, the overall AUC remains high at 0.99 suggesting that the bispectral statistics are robust to recompression.

Shown in Figure 4 is the confusion matrix for the multi-class classification showing that the differences in bicoherence statistics are sufficient not only to distinguish human from synthesized speeches but also, with a reasonable degree of accuracy, to distinguish between different types of synthesized speech.

4. Discussion

We have developed a forensic technique that can distinguish human from synthesized speech. This technique is based on the observation that current speech-synthesis algorithms introduce specific and unusual higher-order bispectral correlations that are not typically found in human speech. We have provided preliminary evidence that these correlations are the result of the long-range correlations introduced by the underlying network architectures used to

[4] `scikit-learn.org`

synthesize speech. This bodes well for us in the forensic community as it appears that these network architectures are also what is giving rise to more realistic sounding speech (despite the unusual bispectral correlations). More work, however, remains to be done to more precisely understand the specific source of the unusual bispectral correlations.

As with any forensic technique, thought must be given to counter-measures that our adversary might adopt. While it would be straight-forward to match first-order spectral correlations between human and synthesized speech, the higher-order spectral correlations are not so easily matched. In particular, we know of no closed-form solution for inverting the bispectrum or bicoherence. It remains to be seen if other techniques like generative adversarial networks can synthesize audio while matching the bispectral artifacts that currently can be used to distinguish human from synthesized speech.

Acknowledgment

This research was developed with funding from Microsoft, a Google Faculty Research Award, and the Defense Advanced Research Projects Agency (DARPA FA8750-16-C-0166). The views, opinions, and findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

[1] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. *arXiv preprint arXiv:1808.07371*, 2018. 1

[2] J.W.A. Fackrell and Stephen McLaughlin. Detecting nonlinearities in speech sounds using the bicoherence. *Proceedings of the Institute of Acoustics*, 18(9):123–130, 1996. 2

[3] Hany Farid. Detecting digital forgeries using bispectral analysis. Technical Report AI Memo 1657, MIT, June 1999. 1

[4] Yu Gu and Yongguo Kang. Multi-task WaveNet: A multi-task generative model for statistical parametric speech synthesis without fundamental frequency conditions. In *Interspeech*, Hyderabad, India, 2018. 1

[5] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 1

[6] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2018. 1

[7] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep Video Portraits. *ACM Transactions on Graphics*, (4):163, 2018. 1

[8] Jaime Lorenzo-Trueba, Fuming Fang, Xin Wang, Isao Echizen, Junichi Yamagishi, and Tomi Kinnunen. Can we steal your vocal identity from the internet?: Initial investigation of cloning Obama's voice using GAN, WaveNet and low-quality found data. In *The Speaker and Language Recognition Workshop (Odyssey)*, Les Sables d'Olonne, France, 2018. 4

[9] Jerry M. Mendel. Tutorial on higher order statistics (spectra) in signal processing and system theory: theoretical results and some applications. *Proceedings of the IEEE*, 79:278–305, 1996. 2

[10] Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, Hao Li, Richard Roberts, et al. paGAN: real-time avatars using dynamic textures. In *SIGGRAPH Asia 2018 Technical Papers*, page 258. ACM, 2018. 1

[11] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: 2000-speaker neural text-to-speech. *arXiv preprint arXiv:1710.07654*, 2017. 1

[12] Md Sahidullah, Tomo Kinnunen, and Cemal Hanilci. A comparison of features for synthetic speech detection. In *Interspeech*, Dresden, Germany, 2015. 1

[13] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing Obama: learning lip sync from audio. *ACM Transactions on Graphics*, 36(4):95, 2017. 1

[14] Justus Thies, Michael Zollhöfer, Christian Theobalt, Marc Stamminger, and Matthias Nießner. Headon: Real-time reenactment of human portrait videos. *ACM Transactions on Graphics*, 36(4):95, 2018. 1

[15] Mohammed Zakariah, Muhammad Khurram Khan, and Hafiz Malik. Digital multimedia audio forensics: past, present and future. *Multimedia Tools and Applications*, 77(1):1009–1040, 2018. 1