The Sound of Silence: Efficiency of First Digit Features in Synthetic Audio Detection

Daniele Mari, Federica Latora, Simone Milani
Department of Information Engineering, *University of Padova*, Padova, Italy
{daniele.mari,federica.latora,simone.milani}@dei.unipd.it

***Abstract*—The recent integration of generative neural strategies and audio processing techniques have fostered the widespread of synthetic speech synthesis or transformation algorithms. This capability proves to be harmful in many legal and informative processes (news, biometric authentication, audio evidence in courts, etc.). Thus, the development of efficient detection algorithms is both crucial and challenging due to the heterogeneity of forgery techniques.**

This work investigates the discriminative role of silenced parts in synthetic speech detection and shows how first digit statistics extracted from MFCC coefficients can efficiently enable a robust detection. The proposed procedure is computationally-lightweight and effective on many different algorithms since it does not rely on large neural detection architecture and obtains an accuracy above 90% in most of the classes of the ASVSpoof dataset.

***Index Terms*—synthetic speech detection, first digit statistics, fake audio detection, silenced signal analysis, Random Forest**

I. Introduction

Fake human speech recordings have recently proved significantly harmful with respect to misinformation, fake news widespreading, frauds, and ID replacement [1]. Such results are the outcome of the recent evolution of computing facilities and deepfake technologies, which has allowed the generation of more and more credible synthetic images, videos, and speech audio signals. As a matter of fact, this development has urged the need for accurate fake audio detection strategies that help human listeners in discriminating fraudulent audio samples from bonafide ones.

Several fake audio detection strategies were proposed in literature targeting different types of acoustic features that are present in a real signal and, at the same time, are difficult to synthesize.

Traditional methods rely on estimating fake audio peculiarities from audio transform coefficients like MFCC or LPC [2]. More recently, such coefficients have been replaced by learned feature representations generated with CNN or RNN architectures [3]. Some other strategies rely on the effects of the physical acquisition environment on the signal (e.g., reverberation, noise, etc.) [4]–[6] or on prosodic and emotional characteristics [7]. Other solutions rely on statistics and symmetry properties of speech signals [8]. Among these, it is worth mentioning the First Digits (FD) statistics computed on signal transform coefficients [9]–[11], whose applications have been widely exploited in other multimedia contents [12].

Although these solutions aim at detecting the peculiar characteristics of a fake audio signal, more recent works [13] have highlighted how synthetic speech algorithms prove to be effective in spoken parts but fail in generating realistic silence. The work by Muller *et al.* shows that the length of trailing silenced parts[1] in synthetic speech samples from ASVSpoof dataset [14] prove to have different statistics with respect to bonafide samples. Indeed, removing such parts dramatically reduces the detection efficiency of most algorithms.

The current paper aims at investigating this eventuality more in depth by analyzing the discriminative potentialities of silenced parts in ASVSpoof dataset. More precisely, we show that FD statistics prove to be effective in discriminating fake audio samples since they allow catching irregularities in silenced parts between the different words of the speech. Tests were run both on the full audio sequence, on the silenced parts, and on the voiced segments (regardless of their lengths). Experimental results show that the statistical characteristics of silence prove to be a discriminative feature since the performance on silent sections matches the detection performance on the full sequence (while this result is not verified for voiced sections). This implies that silence extraction is no longer needed, allowing to avoid parameter tuning and arbitrary set-ups. The final performance has proved to be higher than previous state-of-the-art approaches with a limited computational effort.

It is possible to summarize the novel contributions of the current paper as follows.

- We analyzed the role of silent parts in detection showing that most of the classification accuracy derives from the difficulty in synthesizing statistically-realistic silence intervals.
- We evaluated the efficiency of MFCC FD statistics in detecting audio fake samples generated by a set of different heterogeneous algorithms. Such features have proved to be extremely useful in highlighting the statistics of silenced parts.
- We designed a lightweight classifier whose efficiency can cope with more complex detectors.

The code developed to produce the results presented in this work can be found at https://github.com/Dan8991/The-Sound-Of-Silence.

In the following, the paper is organized as follows. Section II overviews some audio forgery detection algorithms that have been proposed in the literature. Section III describes the

[1]Silent intervals at the end and at the beginnning of the audio sequence.

proposed approach, Section IV illustrates the dataset and the experimental setup, while Section V reports the final accuracy on different types of datasets. Final conclusions are drawn in Section VI.

II. Related Works

Generative audio speech approaches can mainly be divided in two branches i.e. text to speech (TTS) and voice conversion (VC) algorithms. The former starts from a textual representation and aims at producing the corresponding waveform, while the latter modifies the signal to change the perceived identity of the speaker in the audio.

Early TTS approaches were based on waveform concatenation [15], [16] where diphones from large datasets are concatenated seamlessly. More recently researchers have started to design techniques that produce audio features from text representations using an acoustic model (usually a hidden markov model) [17], [18] that are then processed with a vocoder synthesizer such as STRAIGHT [19], WORLD [20] or VOCAINE [21] to produce the corresponding waveform. To improve upon this, neural networks have also been used to substitute either the acoustic model [22] or the vocoder [23], [24] later leading to the first end to end TTS generation algorithms [25], [26].

On the other hand, VCs pipelines usually extract an intermediate representation of the audio signal (feature extraction step), this is then mapped to a representation that matches the target characteristics (feature mapping step) which is finally used to obtain the final waveform (reconstruction step).

Most feature extraction techniques are usually based on pitch synchronous overlap and add (PSOLA) [27] that represents the input as the parameters required by a vocoder synthesizer to reproduce it. This is a useful intermediate characterization of the signal because it allows performing reconstruction with a vocoder, which is convenient since these algorithms are well tested and efficient. On the other hand, the mapping function is usually implemented with parallel training methods by using a gaussian mixture model [28] or neural networks [29], [30]. The mapping can also be performed by means of Generative Adversarial Networks (GANs) since the task is similar to image to image translation allowing similar techniques to be adopted [31], [32].

Classic audio forgery detection algorithms usually perform classification by relying on hand crafted features such as Constant-Q Cepstral Coefficients [33], Log Magnitude Spectrum or phase- based features like Group Delay [34] and Linear Frequency Cepstral Coefficients (LFCC) or MFCCs [35]. More discriminative representations have been recently proposed by exploiting the bicoherence matrix [36], long-short term features computed in an autoregressive manner [37], environmental cues [6], and even emotions [7].

Also in this case neural network based techniques have proven very effective. Some examples are [5], where the frequency representations of the signals are fed to simple convolutional neural networks (CNNs), and in [3] where the convolutional filters are just used for feature extraction while a recurrent neural network is exploited for classification. Some approaches have also been directly applied to the raw input signal (i.e. in the time domain) [38]. In particular, Rawnet2 [38] has achieved impressive results both for synthetic speech detection and user identification. For this reason, it has been proposed as the baseline for the ASVSpoof 2021 challenge [14] i.e. where the dataset considered in this paper for training and testing was proposed.

III. First digit features for synthetic audio tracks.

First digit law has proved very effective in the detection of multiple compressed data [39]–[41]. More recently, it has also been shown its effectiveness in detecting GAN generated images [12]. Following this trend, it is possible to verify that any synthetic signal generated by a set of FIR filters with limited support fits Benford's law with a different accuracy with respect to a natural signal.

Audio waveforms $x(t)$ are represented in the frequency domain by computing the MFCC coefficients $m_w(f)$, where f is the considered frequency and w is the index of the frame. This representation has already proved very effective in highlighting the more meaningful frequency elements in audio signals and in detecting forged waveforms [35]. Since the original samples in the considered dataset sometimes contain long sequences of zeros (which result in zero-valued MFCCs coeffiecients) and since computing FD statistics requires processing non-zero signals, zero values were removed from the input data. This operation does not compromise the final results because this eventuality was verified on both training and test sets, as well as on both natural and synthetic audio.

In order to obtain rich features that can highlight irregularities in the data, MFCC coefficients were quantized with different step values Δ as

$$m_{w,\Delta}(f) = \frac{m_w(f)}{\Delta}. \tag{1}$$

At this point, first digits were computed on $m_{w,\Delta}(f)$ as

$$d_{w,\Delta}(f) = \left\lfloor \frac{|m_{w,\Delta}(f)|}{b^{\lfloor log_b |m_{w,\Delta}(f)| \rfloor}} \right\rfloor \tag{2}$$

where b is the considered integer representation base (e.g. 10 for decimal).

For each distinct cepstral coefficient and for each quantization step, we computed the probability mass function

$$p_{f,\Delta}(d) = \sum_{w=1}^{n_w} \frac{\mathbb{1}_d(d_{w,\Delta}(f))}{n_w} \tag{3}$$

where $\mathbb{1}_d(d_{w,\Delta}(f))$ is the indicator function for digit d, and n_w is the number of windows in the signal whose value depends on the duration of the audio and on the window overlap.

Several previous studies show that this p.m.f. can be approximated by the generalized Benford's law, i.e.,

$$\hat{p}_{f,\Delta}(d) = \beta log_b\left(1 + \frac{1}{\gamma + d^\delta}\right) \tag{4}$$

and the approximation accuracy highly varies if we are considering bonafide w.r.t. forged data [10], [11]. As a matter

of fact, such accuracy was measured using different distance and divergence measures to quantify the proximity of $p_{f,\Delta}(d)$ w.r.t. $\hat{p}_{f,\Delta}(d)$. In the rest of the paper, we will omit indexes Δ and f for the sake of simplicity although in the creation of the final set of features multiple values of f and Δ were considered.

A first traditional divergence metric is the Shannon divergence

$$D^{JS}(p|\hat{p}) = D^{KL}(p|\hat{p}) + D^{KL}(\hat{p}|p). \tag{5}$$

which can be seen as a symmetrized version of the Kullbak-Leibler divergence $D^{KL}(p|\hat{p})$. Additionally, since such metric proves to be unstable for biased pmfs, so we computed Reny $D^R_\alpha(p|\hat{p})$ and Tsallis $D^T_\alpha(p|\hat{p})$ ($\alpha \in [0,1]$) divergences as well

$$D^R_\alpha(p|\hat{p}) = \frac{1}{1-\alpha}\big(logS_\alpha(p,\hat{p}) + logS_\alpha(\hat{p},p)\big) \tag{6}$$

$$D^T_\alpha(p|\hat{p}) = \frac{1}{1-\alpha}\big(2 - S_\alpha(p,\hat{p}) - S_\alpha(\hat{p},p)\big) \tag{7}$$

where

$$S_\alpha(p,q) = \sum_{d=1}^{b-1} \frac{p(d)^\alpha}{q(d)^{\alpha-1}} \tag{8}$$

Additionally, since Reny, Tsallis, and Shannon divergences can be highly correlated for certain values of α (in this work $\alpha = 0.3$ is used), we also added the mean square error

$$D^{MSE}(p,\hat{p}) = \frac{1}{b-1}\sum_{d=1}^{b-1}(p(d) - \hat{p}(d))^2 \tag{9}$$

This addition was supported by some preliminary tests where the divergences of original and voice converted audios were compared: it was possible to deduce that Reny Tsallis and Shannon divergences often agree, meaning that the three divergences in the original sample are always smaller than those in the forged sample or vice-versa. This statement does not always hold for MSE.

In the end, the total number of features n_f is equal to $n_f = n_d n_c n_b n_q$ where n_d is the number of divergences, n_c is the number of chosen cepstral coefficients, n_b is the number of basis for the first digit extraction and n_q is the number of different Δ parameters.

A. A FIR-oriented interpretation of FD statistics for synthetic speech

In the past literature, several works have provided different explanations for the effectiveness of FD statistics in detecting forgeries (on images, audio files, etc.) [11]. Most of the proposed works were focusing on the original data statistics on which FDs were computed. Indeed, Benford's law and its generalized version can be verified for any set of data m such that their probability mass function (pmf) has an exponentially-decreasing behavior (this has been largely verified on images, where coefficients can be modeled with a Laplacian or a generalized Gaussian distribution) [10]. Whenever the image or the set of data are altered, the property is not verified anymore

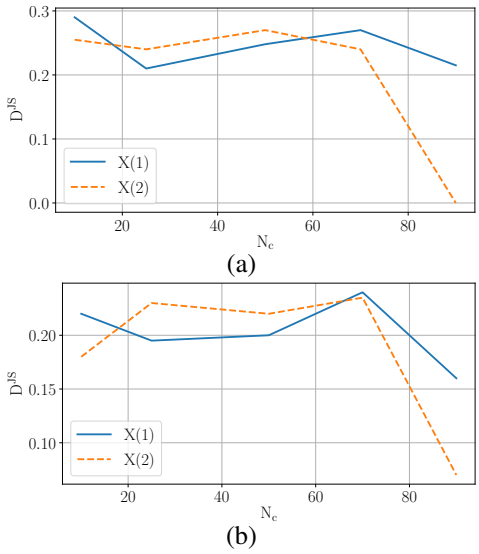

Fig. 1. Jensen-Shannon divergence values between computed and fitted *fd* statistics. Data were computed with (a) $\Delta = 0.008$ and (b) $\Delta = 0.01$.

since the modification redistributes data among the bins of the quantizer. Indeed, the final pmf presents some oscillating probability values that deviate from the ideal distribution.

Anyway, such oscillations can be also related to the ripples in the frequency response of small FIR filters (like those used in GAN-based or VOCODER-based speech synthesizers) that propagates to the statistics of MFCC coefficients. After these multiplications, the values $|m_{w,\Delta}(f)|$ belonging to the same quantization bin are re-distributed unevenly among the other bins. Instead, whenever the gain is perfectly flat, all the MFCC coefficients are rescaled with the same factor, and as a matter of fact, the whole statistics is simply stretched. From these considerations, it is possible to conclude that the flatness of filter gain is deeply connected to the verification of Benford's law. Fig. 1a, 1b report the values of the Jensen-Shannon divergence $D^{JS}\left(\hat{p}_{f,\Delta}(d)|p_{f,\Delta}(d)\right)$ between $\hat{p}_{f,\Delta}(d)$ and $p_{f,\Delta}(d)$ obtained on a 1-D Gaussian i.i.d. signal $x(t)$ filtered by an FIR filter $h(t)$. The filter $h(t)$ is a standard low-pass FIR filter with normalize cut-off frequency 0.2, stop-band frequency 0.7, and N_c coefficients. Filter coefficients were computed using the Parse-McLellan algorithm.

Divergences have been computed from the statistics of quantized MFCC coefficients computed at frequencies 2 and 3. It is possible to notice that $D^{JS}\left(\hat{p}(d)|p(d)\right)$ decreases as N_c increases; however, such behavior can be more or less enhanced depending on coefficient frequency and the quantization parameter Δ. As a matter of fact, it is necessary to use different frequency values and quantization set-ups.

IV. Dataset preparation and experimental setup

The dataset used to validate the proposed approach is ASVSpoof [14] since it provides a great variety of synthetic speech samples generated by heterogeneous algorithms (see Tables I,II). In fact, ASVSpoof covers both text to speech (TTS) and voice conversion (VC) scenarios including samples generated by algorithms based on waveform concatenation

(WC), transfer function (TF), non parallel voice conversion systems (NP), and neural networks (NN).

Data can be divided in three main parts: training, development, and evaluation datasets. The former was used in training and selecting the final classification model. Samples were randomly decimated to ensure that every generated class has the same number of audio traces and that the total amounts of bonafide and synthetic data are balanced (the adopted classifier is a random forest which has lower overfitting problems so the data used for training is plenty).

Development and evaluation datasets were used for closed-set and open-set testing, respectively. More precisely, the former includes newly generated samples (not seen before) generated by the same set of algorithms of the training set, while the latter includes samples generated by different strategies not included in the training set.

The work by Muller et. al. [13] shows that a bias can be found in the distribution of the lengths of leading and trailing silences in bonafide and synthetic speeches. Authors argue that most detectors are just probably discriminating between forged and bonafide samples by using this information. In order to bypass this problem, silent parts were removed from the signal, as suggested in [13] but this led to a big loss in performance.

Therefore, we have decided to analyze the effectiveness of FD features on the silent (without considering leading and ending silences) and voiced parts of the signals, independently. This allowed us to understand which speech elements proved to be the most discriminative and whether the proposed approach was reliably effective. For this purpose, we selected signal windows of 101 samples with energy $E(s,t)$ higher than -40 dB (assuming energy is normalized).

From this filtering, only a few samples (less than 1%) were then removed since the number of silent values was not enough to obtain meaningful statistics. Arguably, this is not an issue since as shown in [13], the very low amount of silence in the audio track allows an easy detection of synthetic audio samples. Moreover, computing FD statistics on a limited amount of signal windows would lead to highly irregular statistics: this implies strong divergences/distances with respect to Benford's law (and therefore, a correct classification).

Starting from the original samples, three datasets have been generated, one called *Full* containing the whole waveform, one referred as *Silence* made with the silent parts of the signals, and one called *Voiced* with the remaining samples.

On these samples, cepstral analysis was carried on in order to generate a feature array for each sample. In this process, the following parameter values were selected after an extensive set of optimizations.

- In the computation of MFCCs, a filter bank of 26 filters was adopted: only coefficients from the second to the fourteenth frequency were considered. Computation was carried out on window sizes of 1024 samples with an overlap of 512 in the case of *Full* and *Voiced*. Overlap was set to 128 in the case of *Silence* to have a sufficient number of signal windows (and therefore stable FD statistics).

Dataset	**Development**					
Algorithm	**A01**	**A02**	**A03**	**A04**	**A05**	**A06**
Type	TTS	TTS	TTS	TTS	VC	VC
Approach	NN	NN	NN	WC	NN	TF
Silence Δ=1	0.944	0.962	0.961	0.819	0.949	0.471
Silence Δ=1-2	**0.953**	0.972	0.970	0.829	0.961	0.472
Silence Δ=1-3	0.951	0.972	0.972	0.836	**0.964**	0.466
Silence Δ=1-4	0.952	0.973	0.972	0.838	0.963	0.456
Silence b=10	0.945	0.959	0.961	0.830	0.924	0.468
Silence b=20	0.866	0.973	0.881	0.796	0.957	0.434
Full Δ=1-4	0.951	**0.982**	**0.949**	**0.871**	0.956	0.424
Voiced Δ=1-3	0.755	0.708	0.713	0.548	0.574	**0.532**

TABLE I
ON-SET RESULTS AND ABLATION STUDIES FOR THE PROPOSED ALGORITHM

- The base for the first digit was chosen as $b \in \{10, 20\}$ since higher values would imply only a few samples (or no samples at all) for many FD values.
- The quantization factor Δ varied in the set $\{1, 2, 3, 4\}$.

At the end of the generation process, feature arrays were made of $n_f = 420$ features.

Given the number and the statistical independence of features (as well as the need for a low complexity classifier), we avoided the adoption of complex neural network architectures. For this reason, a simple random forest classifier was selected as it proved well suited for tabular data processing and highly robust w.r.t. overfitting problems and unbalancing.

The best configuration was selected by running a grid search over the number of trees in the random forest ($n_{trees} \in \{10, 100, 500, 1000\}$) and the criterion for the split quality ($criterion \in \{gini, entropy\}$).

V. RESULTS

Given the three testing scenarios (*Full*, *Voiced*, *Silence*), various ablation studies were carried out to verify the efficiency of the classification in the different set-ups and identify the most discriminative elements in the FD divergences. In order to guarantee a fair comparison between the various configurations, we run an independent grid search for each features configuration.

Table I reports the one-vs-one on-set results obtained by performing binary classification between bonafide and samples generated with a single algorithm. Table II reports the one-vs-one off-set accuracy.

Full ablation studies are reported only for *Silence* for the sake of conciseness, while for *Full* and *Voiced* only the best results are reported. Considering the impact of base selection, keeping the features generated by both $b = 10$ and $b = 20$ (with all the selected values of Δ) turned out to be more effective than choosing only one base value. Indeed, we verified that the statistics generated for $b = 10$ are not very correlated with those obtained for $b = 20$, and therefore, merging them provides additional information to the system.

With respect to the quantization parameter Δ, in Table I and II the features related to different Δs were incrementally concatenated one at a time in order to measure how much they affected the final performance. It is possible to see that having only one Δ value is usually not enough to maximize performance. Experiments show that in general 3 or 4 different quantization values allow to achieve the best performance.

Dataset	Evaluation												
Algorithm	**A07**	**A08**	**A09**	**A10**	**A11**	**A12**	**A13**	**A14**	**A15**	**A16**	**A17**	**A18**	**A19**
Type	TTS	TTS	TTS	TTS	TTS	TTS	TTS+VC	TTS+VC	TTS+VC	TTS	VC	VC	VC
Approach	NN	NN	NN	NN	NN	NN	NN	NN	NN	WC	NN	NP	TF
Silence Δ=1	0.946	0.948	0.955	0.947	0.947	0.952	0.953	0.931	0.876	0.860	0.597	0.615	0.592
Silence Δ=1-2	0.953	0.953	0.965	0.954	0.956	0.959	0.960	0.939	**0.888**	0.861	0.598	0.626	0.597
Silence Δ=1-3	**0.955**	**0.957**	**0.968**	**0.956**	**0.957**	**0.962**	**0.962**	0.941	**0.888**	0.864	0.600	0.629	**0.599**
Silence Δ=1-4	0.951	0.955	0.965	0.952	0.956	0.960	0.959	**0.942**	0.887	0.864	**0.601**	0.625	0.598
Silence b=10	0.925	0.933	0.944	0.927	0.929	0.936	0.928	0.912	0.860	0.846	0.598	**0.642**	**0.599**
Silence b=20	0.919	0.897	0.929	0.924	0.924	0.903	0.945	0.889	0.842	0.820	0.590	0.579	0.593
Full Δ=1-4	0.941	0.942	0.952	0.939	0.940	0.915	0.951	0.896	0.853	**0.866**	0.597	0.581	0.596
Voiced Δ=1-3	0.656	0.796	0.798	0.629	0.648	0.628	0.687	0.720	0.709	0.640	0.526	0.533	0.580

TABLE II
OFF-SET RESULTS AND ABLATION STUDIES FOR THE PROPOSED ALGORITHM

Algorithm	Development	Evaluation
Silence	0.869	0.819
Full	0.871	**0.820**
STLT + Bicoherence 128	**0.942**	0.735
STLT + Bicoherence 512	0.907	0.741

TABLE III
CLASSIFICATION ACCURACY OF THE PROPOSED ALGORITHM AND OF SOME STATE OF THE ART APPROACHES

In both on-set and off-set accuracies, it is possible to see that performing classification over the features computed on *Silent* leads to performance that is comparable to or even better w.r.t. the one obtained for *Full*. In particular, when considering off-set tests, silences provide a higher detection accuracy for almost all the algorithms.

On the other hand, removing silences from the signals leads to very poor performance (see results on *Voiced* sections). This might suggest that algorithms reconstruct realistic voices more easily (low-pass regular signal), while the noise present in silent sequences can not be easily modeled. The slightly lower performance achieved on *Full* w.r.t. *Silence* might be explained by the presence of spoken parts that might be skewing the FD statistic towards the ideal FD distribution.

It is possible to see that this approach has a lower performance when referred to algorithms A06, A17, A18, A19. It is worth noticing that all these approaches perform a voice conversion task starting from real audio samples as input and converting them into voiced samples for a desired speaker. On the contrary algorithm A05 is also a VC algorithm but the task is carried out by a neural network that processes the full sequence (silence included) leading to FD statistics that are detected by our approach. Additionally, voice converted samples generated starting from TTS outputs (see results referred to algorithms A13, A14, and A15) are also easily classified with high accuracy. A possible explanation for this evidence is that VC algorithms do not change significantly silenced sections as they are not relevant in characterizing speaker ID and present a completely different statistics w.r.t. voiced parts. In these cases, the statistics of the original silenced intervals are not altered leading to a higher misclassification probability. Note that this outcome is not verified whenever VC is applied after TTS since in that case also the generated nature of silence leads to non-conventional FD statistics (thus leading to higher divergences/distances).

On top of that, it is worth spending a few comments on A16 and A04 approaches. These are based on waveform concatenation, i.e., signals are obtained by concatenating real samples from big databases of diphones (realistic ones). Such composite nature makes the synthetic speech FD statistics closer to that of bonafide samples (and further from the generated audio) leading to a higher misclassification probability.

In the end, we compared our approach with a similar state-of-the-art algorithm (i.e. with separate feature extraction and classification steps). The work by Borrelli et. al. [37] exploits short-term and long-term (STLT) cues and the bicoherence matrix to extract a discriminative representation between forged and bonafide samples. In Table III the results of the two best configurations proposed in the aforementioned work, i.e. using both STLT and bicoherence features computed with window sizes 128 and 512, are compared with the results obtained with the best configurations in the *Silence* and *Full* datasets. It is possible to see that while the proposed approach is less accurate in the detection of forged data in the development dataset, it achieves better performance in off-set evaluation proving more robust when presented with unseen algorithms.

VI. CONCLUSIONS

In this paper, we analyzed the impact of voiced and silenced parts in synthetic speech detection. Following some preliminary studies on trailing silences, we showed that silenced parts within the speech contain most of the discriminative information. From these results, we proposed a method for forged audio detection based on first digit statistics that achieves good detection performance against a variety of algorithms and that has very low computational complexity. Empirical results showed that most audio forging algorithms are able to produce statistically meaningful voice signals but (especially neural networks) often fail at creating realistic silences. Future works should try to tackle the problem of detection in a voice conversion scenario (possibly by integrating this with other well working state of the art approaches) since the transformation of a naturally acquired signal could retain most of the statistics for the silenced parts thus leading to a higher misclassification probability.

REFERENCES

[1] "A voice deepfake was used to scam a CEO out of $243,000.," https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000, Accessed: 2022-06-23.

[2] Madhu R. Kamble, Hardik B. Sailor, Hemant A. Patil, and Haizhou Li, "Advances in anti-spoofing: from the perspective of asvspoof challenges," *APSIPA Transactions on Signal and Information Processing*, vol. 9, pp. e2, 2020.

[3] Chunlei Zhang, Chengzhu Yu, and John HL Hansen, "An investigation of deep-learning frameworks for speaker verification antispoofing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 684–694, 2017.

[4] Simone Milani, Pier Francesco Piazza, Paolo Bestagini, and Stefano Tubaro, "Audio tampering detection using multimodal features," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4563–4567.

[5] Alessandro Lieto, Daniele Moro, Francesco Devoti, Claudia Parera, Vincenzo Lipari, Paolo Bestagini, and Stefano Tubaro, "" hello? who am i talking to?" a shallow cnn approach for human vs. bot speech classification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2577–2581.

[6] Davide Capoferri, Clara Borrelli, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, and Stefano Tubaro, "Speech audio splicing detection and localization exploiting reverberation cues," in *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2020, pp. 1–6.

[7] Emanuele Conti, Davide Salvi, Clara Borrelli, Brian Hosler, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, Matthew C Stamm, and Stefano Tubaro, "Deepfake speech detection through emotion recognition: A semantic approach," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8962–8966.

[8] Arun Kumar Singh and Priyanka Singh, "Detection of ai-synthesized speech using cepstral & bispectral statistics," in *2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2021, pp. 412–417.

[9] Tiziano Bianchi, Alessia De Rosa, Marco Fontani, Giovanni Rocciolo, and Alessandro Piva, "Detection and classification of double compressed mp3 audio tracks," in *Proceedings of the First ACM Workshop on Information Hiding and Multimedia Security*, 2013, p. 159–164.

[10] Tomas Pevny and Jessica Fridrich, "Detection of double-compression in jpeg images for applications in steganography," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 2, pp. 247–258, 2008.

[11] Fernando Perez-Gonzalez, Greg L. Heileman, and Chaouki T. Abdallah, "Benford's lawin image processing," in *2007 IEEE International Conference on Image Processing*, 2007, vol. 1, pp. I – 405–I – 408.

[12] Nicolo Bonettini, Paolo Bestagini, Simone Milani, and Stefano Tubaro, "On the use of benford's law to detect gan-generated images," in *2020 25th international conference on pattern recognition (ICPR)*. IEEE, 2021, pp. 5495–5502.

[13] Nicolas Müller, Franziska Dieckmann, Pavel Czempin, Roman Canals, Jennifer Williams, and Konstantin Böttinger, "Speech is silver, silence is golden: What do asvspoof-trained models really learn?," in *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 06 2021.

[14] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, et al., "Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," *arXiv preprint arXiv:2109.00537*, 2021.

[15] Alan W Black and Nick Campbell, "Optimising selection of units from speech databases for concatenative synthesis.," 1995.

[16] Soumya Priyadarsini Panda and Ajit Kumar Nayak, "A waveform concatenation technique for text-to-speech synthesis," *International Journal of Speech Technology*, vol. 20, no. 4, pp. 959–976, 2017.

[17] M Kiran Reddy and K Sreenivasa Rao, "Robust pitch extraction method for the hmm-based speech synthesis system," *IEEE signal processing letters*, vol. 24, no. 8, pp. 1133–1137, 2017.

[18] Keiichi Tokuda, Heiga Zen, and Alan W Black, "An hmm-based speech synthesis system applied to english," in *IEEE speech synthesis workshop*. IEEE Santa Monica, 2002, pp. 227–230.

[19] Hideki Kawahara, "Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, vol. 27, no. 6, pp. 349–353, 2006.

[20] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[21] Yannis Agiomyrgiannakis, "Vocaine the vocoder and applications in speech synthesis," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 4230–4234.

[22] Wenfu Wang, Shuang Xu, Bo Xu, et al., "First step towards end-to-end parametric tts synthesis: Generating spectral parameters with neural attention.," in *Interspeech*, 2016, pp. 2243–2247.

[23] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[24] Jean-Marc Valin and Jan Skoglund, "Lpcnet: Improving neural speech synthesis through linear prediction," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5891–5895.

[25] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al., "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[26] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan Ömer Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller, "Deep voice 3: 2000-speaker neural text-to-speech.," 2017.

[27] Levent M Arslan, "Speaker transformation algorithm using segmental codebooks (stasc)," *Speech Communication*, vol. 28, no. 3, pp. 211–226, 1999.

[28] Yannis Stylianou, Olivier Cappé, and Eric Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on speech and audio processing*, vol. 6, no. 2, pp. 131–142, 1998.

[29] Huaiping Ming, Dong-Yan Huang, Lei Xie, Jie Wu, Minghui Dong, and Haizhou Li, "Deep bidirectional lstm modeling of timbre and prosody for emotional voice conversion.," in *Interspeech*, 2016, pp. 2453–2457.

[30] Kou Tanaka, Hirokazu Kameoka, Takuhiro Kaneko, and Nobukatsu Hojo, "Atts2s-vc: Sequence-to-sequence voice conversion with attention and context preservation mechanisms," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6805–6809.

[31] Takuhiro Kaneko and Hirokazu Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2100–2104.

[32] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, "Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6820–6824.

[33] Massimiliano Todisco, Héctor Delgado, and Nicholas Evans, "Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[34] Xiong Xiao, Xiaohai Tian, Steven Du, Haihua Xu, Engsiong Chng, and Haizhou Li, "Spoofing speech detection using high dimensional magnitude and phase features: the ntu approach for asvspoof 2015 challenge.," in *Interspeech*, 2015, pp. 2052–2056.

[35] Md Sahidullah, Tomi Kinnunen, and Cemal Hanilçi, "A comparison of features for synthetic speech detection," 2015.

[36] Ehab A AlBadawy, Siwei Lyu, and Hany Farid, "Detecting ai-synthesized speech using bispectral analysis.," in *CVPR workshops*, 2019, pp. 104–109.

[37] Clara Borrelli, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, and Stefano Tubaro, "Synthetic speech detection through short-term and long-term prediction traces," *EURASIP Journal on Information Security*, vol. 2021, no. 1, pp. 1–14, 2021.

[38] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher, "End-to-end anti-spoofing with rawnet2," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6369–6373.

[39] Cecilia Pasquini, Giulia Boato, and Fernando Perez-Gonzalez, "Multiple jpeg compression detection by means of benford-fourier coefficients," in *2014 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2014, pp. 113–118.

[40] Fabian Zach, Christian Riess, and Elli Angelopoulou, "Automated image forgery detection through classification of jpeg ghosts," in *Pattern Recognition. DAGM/OAGM 2012. Lecture Notes in Computer Science, vol 7476*, 08 2012, vol. 7476, pp. 185–194.

[41] Simone Milani, Marco Tagliasacchi, and Stefano Tubaro, "Discriminating multiple jpeg compressions using first digit features," *APSIPA Transactions on Signal and Information Processing*, vol. 3, pp. e19, 2014.