Audio Engineering Society

Conference Paper

Presented at the Conference on
Audio Forensics
2019 June 18 – 20, Porto, Portugal

# Fighting AI with AI: *Fake Speech Detection using Deep Learning*

Hafiz Malik[1] and Raghavendar Changalvala[1]

[1] *Information Systems, Security, and Forensics Lab,, Department of Electrical and Computer Engineering,, University of Michigan - Dearborn*

Correspondence should be addressed to Hafiz Malik (hafiz@umich.edu)

## ABSTRACT

Voice cloning technologies have found applications in a variety of areas ranging from personalized speech interfaces to advertisement, video gaming, and so on. Existing voice cloning systems are capable of learning speaker characteristics from few samples and generating perceptually indistinguishable speech. These advances pose new security and privacy threats to voice-driven interfaces. This paper presents a deep learning-based framework for learning cloned speech synthesis models and the bona-fide speech production processes. To this end, a convolutional neural network is trained and tested on spectrogram estimated from input audio recordings. Performance of the proposed method is evaluated on cloned and bona-fide audios. Experimental results indicate that the proposed method is capable of detecting bona-fide and cloned audios with a close to perfect accuracy.

## 1 Introduction

Artificial human speech synthesis from text, also known as text-to-speech (TTS), is an essential feature in many applications including voice-driven interfaces, humanoid robots, navigation systems, video games, chatbots, and accessibility for the visually-impaired. Modern TTS systems are based on complex, multi-stage processing pipelines, each of which may rely on hand-engineered features and heuristics. Recent generative model-based methods have been successfully used for cloned image, audio, and video generation [1, 2, 3, 4]. Generative models based on deep neural networks have been successfully applied to many domains such as image generation [1], speech synthesis [2, 3], and language modeling [4].

Advances in artificial intelligence (AI), speech synthesis, image and video generation technologies pose new security and privacy threats to biometric-based access control systems and voice-driven interfaces. For instance, applications of voice-driven interfaces and services including Amazon Alexa [5], Google Home [6], Apple Siri [7], Microsoft Cortana [8], and so on, are on the rise. The private banking division of Barclays is the first financial services firm to deploy voice biometrics as the primary means to authenticate customers to their call centers [9, 10]. Since then, many voice-biometric-based solutions have been deployed across several financial institutions, including Banco Santander, Royal Bank of Canada, Tangerine Bank, Chase Bank, Citi Bank, Bank of America, and Manulife, HSBC Bank [10, 9, 11, 12, 13, 14, 15, 16, 17], and this list keep growing.

The existing voice-driven interfaces are vulnerable to replay and cloned voice attacks, that is, injection of recorded or cloned voice of an authentic user [18]. Recently, other researchers have investigated vulnerabilities of smart speakers like Google Home [6] and Amazon Alexa[5]. For instance, Lei et al. [19] demonstrated vulnerabilities of smart speakers like Google Home [6] and Amazon Alexa [5]. Likewise, Malik et al. [20] proposed higher-order spectral analysis based framework work to attack voice-interfaces including Good Home and Amazon Alexa through replay attack. Recently, in [21] we proposed a higher-order spectral-based framework for cloned speech detection. The existing state-of-the-art in cloned voice and replay attack detection is still in its infancy. There is an urgent need to develop tools and techniques detect and prevent such attacks.

This paper presents a framework to secure voice-driven interfaces and services against cloned speech attacks. The proposed method exploits generative model artifacts for cloned speech detection. Specifically, trained generative models rely on linear operation of excitation source and learned weights for cloned speech generation process which differs from natural speech generation (bona-fide speech) process. We claim that cloned speech leave characteristics distortions in the synthesize speech. To investigate this claim, both cloned and bona-fide speech recordings are transformed in the spectral domain. Specifically, spectrogram estimated from input audio recording is used to artifacts due to study cloned speech synthesis process (please see Figure 2). It can be observed from Figure 2 that cloned speech synthesis process leaves characteristics distortion in the spectrogram estimated from cloned speech recording. Deep learning is used is used to learn the underlying distortion models for both the cloned speech synthesis and the bona-fide speech production processes. To this end, convolutional neural network (CNN) is trained and tested on spectrogram estimated from input audio recording to the proposed system. Performance of the proposed method is evaluated on cloned audios generated using speaker adaptation- and speaker encoding-based approaches. Performance of the proposed method is evaluated on a dataset consisting of 124 cloned and 124 bona-fide speech samples. Experimental results indicate that the proposed method is capable of detecting bona-fide and cloned audios with a perfect detection rate.

### 1.1 Contributions

In this paper, we propose a framework for cloned speech detection. Contributions of our work are:

1. We have demonstrated that generative models leave characteristics artifacts in the resulting cloned audios.
2. We proposed cloned voice attack model on voice driven-interfaces and ASR systems.
3. We proposed a new method for cloned audio detection. Deep learning based on convolution neural network is used to capture traces of generative models for both speech synthesis and human speech generation models.
4. Effectiveness of the proposed method is on Baidu cloned audio data set available via [22]. Performance is evaluated on cloned audios synthesized using (i) **speaker adaptation** and (ii) **speaker encoding**. The performance of the proposed method is also evaluated on voice impersonation using voice morphing via embedding manipulations.

## 2 Cloned Voice Generation - *An Overview*

Recent advances in AI and deep learning has enabled researchers to clone human speech (e.g., Deep Voice [3, 23, 24]) from few audio samples [22], images, and video. Cloned speech is an extension of traditional text-to-speech pipelines. Deep learning based methods [22] train generative models that adopt the same structure of T2S, but differ in replacing all components with neural networks and rely on relatively simpler features to represent generative model. These generative models can be conditioned on text and speaker identity [23] for speech synthesis. For these generative models, text provides linguistic information and controls the content of the generated speech, whereas, speaker identity carries speaker specific characteristics, e.g., pitch, accent, etc. The multi-speaker speech synthesis systems jointly train a generative model and speaker embeddings on text, audio and speaker identity [24]. These systems share majority of the model parameters across all speakers and use low-dimensional embeddings to encode the speaker-specific information. These methods are capable of generating speech for speakers observed during training phase.

Recently, Arık et al. [22] proposed few-shot generative modeling of speech conditioned on speaker identity. Their system is capable of voice cloning from few speech samples of an unseen speaker. One of the salient features of the few-shot generative modeling system presented in [22] is that it is capable of voice cloning of a new speaker characteristics from a very limited training data, e.g., just few seconds of audio data.

### 2.1 Multi-speaker generative modeling for voice cloning

Consider a multi-speaker generative model $\phi(t_{i,j}, s_i; \pi, e_{s_i})$, which takes a text $t_{i,j}$ and a speaker identity, $s_i$, trainable parameters, $\pi$, and trainable speaker embedding corresponding to speaker i, $e_{s_i}$. The trainable parameters,$\pi$, and $e_{s_i}$, are optimized by minimizing a loss function $\mathscr{L}(.)$, expressed as,

$$\min_{\pi,e} \mathop{\mathbb{E}}_{\substack{s_i \sim \mathscr{S} \\ (t_{i,j},\alpha_{i,j}) \sim \mathscr{T}_{s_i}}} \{\mathscr{L}(\phi(t_{i,j}, s_i; \pi, e_{s_i}), \alpha_{i,j})\} \quad (1)$$

where $\mathscr{S}$ is a set of speakers, $\mathscr{T}_{s_i}$ is a training set of text-audio pairs for speaker $s_i$, and $\alpha_{i,j}$ is the ground-truth audio for $t_{i,j}$ of speaker $s_i$. The expectation is estimated over text-audio pairs of all training speakers.

Estimates of $\pi$ , and $e$, $\hat{\pi}$ , and $\hat{e}$ denote the trained parameters and embeddings. Arık et al. [22] proposed two approaches:

1. **Speaker adaptation** which is based on fine-tuning a multi-speaker generative model. Fine-tuning can be applied to either the speaker embedding or the whole model.
2. **Speaker encoding** which directly estimate the speaker embedding from audio samples of an unseen speaker. Such a model does not require any fine-tuning during voice cloning process. More details on details of these systems can found in [22] and references therein.

## 3 Attack Model

Cloned voice can be used to attack both automatic speaker recognition (ASR) systems and voice-driven interfaces. Shown in Fig. 1 are two possible attack vectors for both systems.

**A1:** *Impersonation attack* where attacker play cloned speech using either a smart-speaker or a humanoid robot in front of the target system, e.g, an ASR system or a voice-driven device.

**A2:** *Injection attack* where attacker directly injects cloned audio into to the target system, e.g, an ASR system or a voice-driven device.

It is important to highlight that injection attack lacks speaker-microphone processing chain. Likewise, a replay attack includes a microphone-speaker-microphone [25] processing chain. Both attack vectors are expected to introduce two different types of distortions. For instance, Injection attack is expected to be more linear than impersonation attack, this is due the fact that microphone/speaker are nonlinear devices [26, 25]. Injection attack, on the other hand, is expected to be relatively more linear when compared with bona-fide speech. This is mainly because cloned voice is a generation process that is relatively more linear than the bona-fide speech generation process, which consists four nonlinear sub-processes respiration, phonation, resonance, and articulation. Next section outlines a framework for cloned voice detection.

## 4 Cloned Audio Detection using Deep Learning

We claim that *generative models leave characteristic artifacts in the cloned speech signals which can be used to distinguish between cloned and bona-fide audios*. To verify this claim, spectrograms of two audios (a ground-truth (bona-fide) speech and a cloned speech generated using speaker adaptation with whole model presented in [22]) are computed using same set of parmeters. Shown in Fig. 2 are the spectrogram plots estimated from a bona-fide speech (top panel) and from a cloned speech generated using speaker adaptation with whole model approach discussed in [22] (bottom panel).

It can be observed from Fig. 2 that cloned speech exhibit vertical lines across time axis (highlighted using yellow ellipses). The spectrogram for bona-fide speech, on the other hand, is smoother across time axis and lacks such vertical line spectra. We have observed, through extensive experimentation, that these artifacts are consistent for all cloned speech signals irrespective of speech cloning method used. Various approaches can be developed to capture such artifacts. In this paper, we propose to use a deep learning based framework to learn

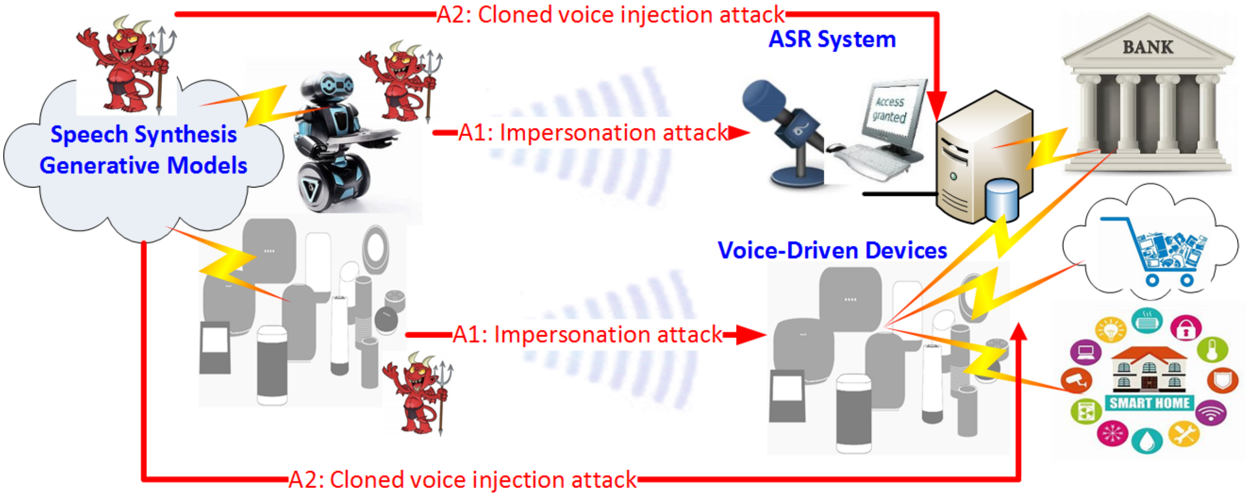

**Fig. 1:** Cloned voice attack model for automated speaker recognition systems and voice-driven interfaces.

underlying models for these artifacts and used them for cloned speech detection. Specifically, convolution neural network (CNN) is considered for learning the underlying model of the cloned audio and used trained model for cloned audio detection. Traditional machine learning- and parametric modeling-based approaches can also be used to achieve this goal. For example, In [21], we proposed a framework based on higher-order statistics to capture artifacts due to nonlinear processing of microphone processing block in the bona-fide audio production and acquisition process, and used it for cloned audio detection. The motivation behind selecting deep learning based model learning is due to its ability to effectively, efficiently, and reliably learn the hidden models from the input data with minimal user intervention in the feature selection and model learning phases.

### 4.1 Convolutional Neural Network (CNN) - *An Overview*

The CNN is one of the most commonly used deep learning techniques for image classification and other AI applications. A simple CNN is a sequence of steps where each step transforms a set of activations from differentiable functions. Similar to any neural network, the CNNs are made up of neurons with learnable weights and biases [27]. Though the weight vector optimization is like the conventional neural networks, the CNNs are designed to deal specifically with 2D or 3D image data. A variety of combinations of linear and non-linear differentiable steps could be used to realize a deep CNN and that determines the complexity of the system. Generally, an image classification CNN model takes a 3D input image $x_i\, i \in \{1..n\}$ and transforms it into a prediction probability vector $\overline{y_i}$ for $n$ different classes. The model is trained using $N$ number of labeled images $\{x_i, y_i\}$ where the label $y_i$ is a class function. A brief overview of the CNN architecture used for the proposed cloned audio detection is provided next.

### 4.2 The CNN Architecture for Cloned Speech Classification

For this study, we designed and developed a non-parametric and fully supervised CNN model to perform speech classification. Our deep learning model consists of four stacks of basic CNN building blocks like *convolution*, *pooling* and *activation* layers. Each layer implements a *2D convolution function* for the convolution operations, *max pooling function* for data size reduction and a *rectified linear unit (ReLu) function* for non-linear activation. The stack of four layers is connected at the end by a fully connected layer of predefined length. A *soft-max activation function* is used at the end to classify input audios into respective cloned audio and bona-fade class. Shown in the Fig. 3 is the symbolic block diagram of the CNN architecture used. A brief overview of each CNN layer is provided next.

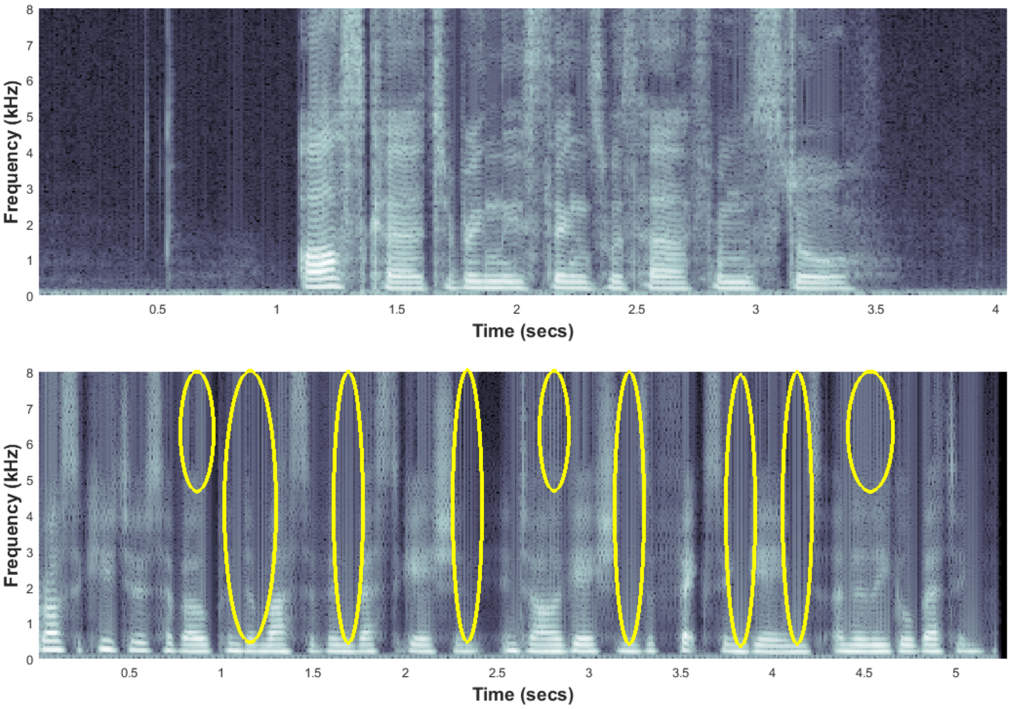

**Fig. 2:** Shown in the top panel is the spectrogram estimated from a bona-fide speech and in the bottom panel is the spectrogram estimated from a cloned speech generated using speaker adaptation with whole model discussed in [22].

#### 4.2.1 The Convolution Layer

A 2D input such as a 3-channel image (e.g., RGB image) of size $R \times C$ is represented as a third-order tensor $R \times C \times D$ in a CNN. The channel depth $D$ is equal to number of image channels, i.e., $D = 3$ for an RGB images. In any given convolution layer of a deep CNN multiple convolution kernels are used. For a kernel count of $K$, a fourth order tensor $W \in \{R \times C \times D \times K\}$ represents all kernels. Convolution layer could be used to alter the input tensor size with variable strides and padding. A simple convolution layer with a stride of unity and zero padding, the output of any convolution layer $y_{i,j,k}$ for $i \in \{0,R\}, j \in \{0,C\}, k \in \{0,K\}$ can be represented as

$$y_{i,j,k} = \sum_{i'=0}^{R} \sum_{j'=0}^{C} \sum_{d'=0}^{D} W_{i' j' d' k} \times x_{i+i',j+j',d'} \qquad (2)$$

where $x_{i+i',j+j',d'}$ in in 2 represents the $(i+i', j+j', d')$ indexed element of the input tensor $x$.

#### 4.2.2 The Pooling Layer

The pooling layer is the basic building block of the CNN architecture. It performs feature space reduction that is feature vector size reduction. It achieves feature reduction by combining related values in the feature vector. In the implemented model, pooling layer is used to reduce the input sample size along the width and height dimensions of the input spectrogram image. The size of pooling window is an adjustable hyperparameter.

#### 4.2.3 The Non-linear Activation Layer

The activation layer represented by the Rectified Linear Unit (ReLU) function that introduces required nonlinearity when applied to the feature vector. The ReLU layer does not modify the size of input feature vector. A simple implementation of ReLU can be expressed as,

$$y_{i,j,k} = \max(0, x_{i,j,k}) \qquad (3)$$

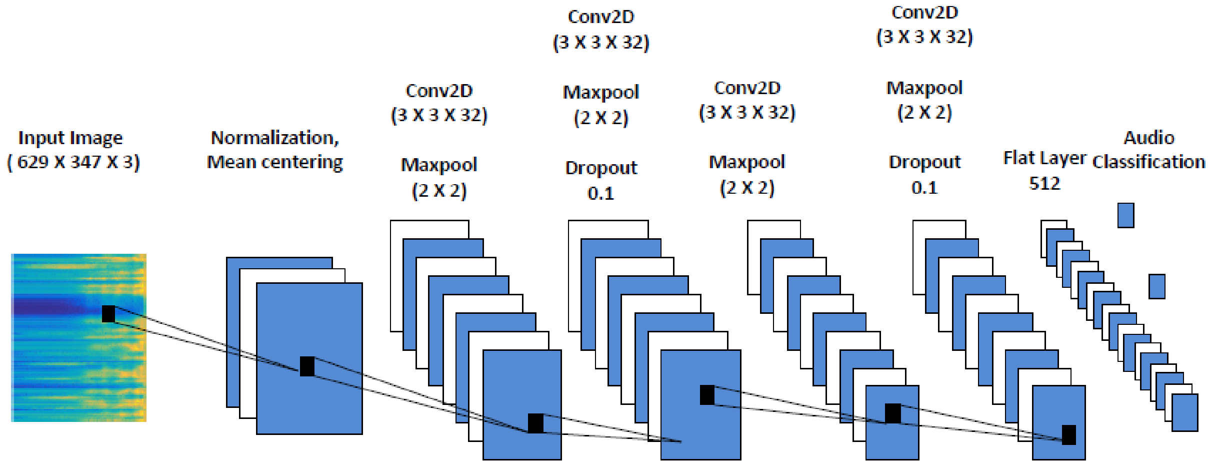

**Fig. 3:** Block diagram of the CNN architecture used for model learning/testing.

where $x$ and $y$ in 3 are input and output tensors of each layer.

### 4.3 The Proposed System

The proposed CNN-based cloned audio detection system can be divided into three processing steps: (i) **audio collection**, (ii) **audio processing and transformation**, and (iii) **model learning and classification**. Shown in Fig. 4 is the block diagram of the proposed system. Brief description of each processing stage is provided in the following:

1. **The data collection stage** collects data either from a cloned speech synthesis algorithm or from a recording system capturing speech of a bona-fide speaker.
2. **The data processing stage transforms** the input audio recording from 1D data sequence into spectrogram, a 2D spectral-temporal plot, using short-time Fourier Transform (STFT).
3. **The model learning and classification stage** learns the underlying models for cloned and bona-fide speech generation processes from spectrogram. Specifically, the 2D representation of an audio recording, e.g., spectrogram, is applied to the model learning (or CNN) stage which captures characteristic features of audio production/synthesis process via deep learning. The learned models are used for binary classification.

**Table 1:** Summary of the Cloned Audio Dataset

| no of | $VCTK_t \to VCTK_g$ | | $LS_t \to VCTK_g$ | |
|---|---|---|---|---|
| samples | SEA | WMA | SEA | WMA |
| 1 | 4 | 4 | 4 | 4 |
| 5 | 4 | 4 | 4 | 4 |
| 10 | 4 | 4 | 4 | 4 |
| 20 | 4 | 4 | 4 | 4 |
| 50 | 4 | 4 | 4 | 4 |
| 100 | 4 | 4 | 4 | 4 |
| Speaker encoder ($LS_t \to VCTK_g$) | | | | |
| | Without fine tuning | | With fine tuning | |
| 1 | 4 | | 4 | |
| 5 | 4 | | 4 | |
| 10 | 4 | | 4 | |

## 5 Experimental Results

### 5.1 Dataset

Baidu Silicon Valley AI Lab cloned audio dataset is used for performance evaluation. It is downloaded from https://audiodemos.github.io. This dataset consists of 10 ground truth audio samples, 120 cloned recordings, and four morphed speech recordings. Summary of the dataset used is provided in Table I. More details about the dataset can be found in [22].

Here, SEA – Speaker Embedding Adaption
WMA – Whole Model Adaption

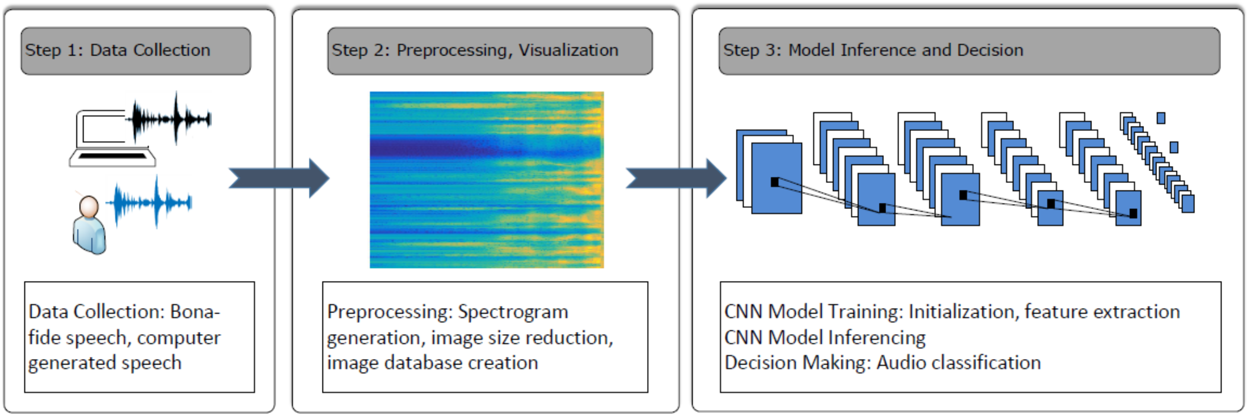
Step 1: Data Collection
Step 2: Preprocessing, Visualization
Step 3: Model Inference and Decision
Data Collection: Bona-fide speech, computer generated speech
Preprocessing: Spectrogram generation, image size reduction, image database creation
CNN Model Training: Initialization, feature extraction
CNN Model Inferencing
Decision Making: Audio classification

**Fig. 4:** Proposed CNN-based system for cloned speech detection

$LS_t$ – training on Libri speakers
$VCTK_t$ – training on VCTK speakers
$VCTK_g$ – using VCTK speakers for cloned speech generation

For this dataset generation, the multi-speaker and speaker encoder models are trained on 84 VCTK speakers (48 KHz sampling rate), other VCTK speakers (48 KHz sampling rate) are used for voice cloning. The multi-speaker and speaker encoder models are trained on LibriSpeech (LS) speakers (16 KHz sampling rate), VCTK speakers (down-sampled to 16 KHz samples/sec.) are used for voice cloning.

For bona-fide speech dataset, 124 speech recordings are randomly selected from TIMIT dataset [28].

For the sake of consistency, each cloned speech recording is downsampled to 16K samples/sec. The bona-fide speech dataset has sampling rate of 16K therefore downsampling is not needed. Each speech recording is then made 5 seconds long through either truncation or concatenation operation. The spectrogram is then estimated from every recording using '*spectrogram()*' - a MATLAB built-in function for spectrogram calculation, with default parameter settings except *segment size* = $0.02 \times 16000 = 320 samples$, $overlapping$ = 50%.

### 5.2 Experimental Setup

The input spectrogram images are normalized and mean centered to nullify the range variation in pixel values. Dropout layers are added to randomly ignore the activation and to prevent over-fitting. For convolution layer a $3 \times 3$ kernel, with a kernel count of 32 is used. A symmetric window of size $2 \times 2$ is used for the pooling layer.

Adadelta optimizer that optimizes the gradient based on the moving window of gradient updates is used. Learning rate $\eta$ is set to adaptive model and it is chosen by the optimizer. The model is trained for ten epochs using a batch-size of sixteen.

The deep learning model for audio classification is built using TensorFlow in Python environment. The audio dataset had 248 images each of size $625 \times 469 \times 3$ pixels. A stack of four layers is used to reduce each input image to a fully connected layer of size 512. A softmax activation function is applied on a fully connected layer to get the probability distribution vector of class labels over categorical cross-entropy loss function.

Due to limited training dataset, a five-fold cross validation approach implemented to remove any training sample selection bias. Each iteration of the five-fold validation divided the training dataset in a $1:4$ ratio with 198 training samples and 50 validation samples. After each iteration the model is evaluated on a separate sample set that is not used in training or validation. The performance is evaluated in terms of prediction precision and accuracy. The results presented here are averaged over five iterations.

## 6 Experimental Results

**Experiment 1:** The goal of this experiment is to evaluate performance of the proposed system on spectrograms as a input vector to the CNN model. To this end,

**Table 2:** Model precision for spectrogram input

| Validation (Trial #) | Original Audio | Fake Audio |
|---|---|---|
| 1 | 100% | 100% |
| 2 | 100% | 100% |
| 3 | 100% | 100% |
| 4 | 100% | 100% |
| 5 | 100% | 100% |
| mean | 100% | 100% |

**Table 3:** Model prediction accuracy on independent samples

| Validation (Trial #) | Accuracy |
|---|---|
| 1 | 100% |
| 2 | 100% |
| 3 | 100% |
| 4 | 100% |
| 5 | 100% |
| mean | 100% |

the model is trained and tested on the dataset as per description provided in Section 5.2. Shown in Table 2 are the model precision for both class labels for each input method. It can be observed from Table 2 that the implemented model is able to successfully differentiate between the cloned and bond-fide speech recordings with a perfect detection accuracy.

The trained model is then evaluated on the test samples. The classification accuracy of the model on the validation dataset for the spectrogram feature input shown in Tables 3 and 4. It can be observed from Tables 3 and 4 that the proposed method achieved 100% accuracy for the predictions on test dataset.

## 7 Conclusion

This paper presents a framework for cloned speech detection using deep learning. We have demonstrated that cloned speech exhibit characteristic artifacts which are used for cloned speech detection. Spectrogram, a spectro-temporal representation of an audio recording, is used to capture traces of cloned speech synthesis process. A deep learning based framework is proposed to learn differentiating characteristics of speech production and cloned speech synthesis processes. A CNN-based architecture is used to train the deep models on cloned and bona-fide speech recordings. Effectiveness of the proposed method is evaluated on cloned audios generated using speaker adaptation- and speaker encoding-based approaches proposed in [22], and bona-fide speech recordings randomly selected from TIMIT dataset [28]. Performance of the proposed for a dataset of 248 speech samples indicate that the proposed method is capable of detecting bona-fide and cloned audios with a perfect detection rate.

**Table 4:** Model accuracy for all labels

| Validation (Trial #) | Accuracy |
|---|---|
| 1 | 100% |
| 2 | 100% |
| 3 | 100% |
| 4 | 100% |
| 5 | 100% |
| Mean | 100% |

Our future work aims to investigate performance of our system on other spectral features such as MFCC, log-MFCC, higher order spectral features, etc.

## 8 Acknowledgement

This research is supported by the National Science Foundation (NSF) under the awards No. 1816019.

## References

[1] Karras, T., Aila, T., Laine, S., and Lehtinen, J., “Progressive Growing of GANs for Improved Quality, Stability, and Variation,” *CoRR*, abs/1710.10196, 2017.

[2] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W., and Kavukcuoglu, K., “WaveNet: A Generative Model for Raw Audio,” *CoRR*, abs/1609.03499, 2016.

[3] Arik, S. Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Raiman, J., Sengupta, S., and Shoeybi, M., “Deep Voice: Real-time Neural Text-to-Speech,” *CoRR*, abs/1702.07825, 2017.

[4] Józefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y., “Exploring the Limits of Language Modeling,” *CoRR*, abs/1602.02410, 2016.

[5] “Anazom Alexa,” 2019.

[6] “Google Home,” 2019.

[7] “Apple Siri,” 2019.

[8] “Microsoft Cortana,” 2019.

[9] Matt Warman, “Say goodbye to the pin: voice recognition takes over at Barclays Wealth,” Retrieved June 5, 2013.

[10] International Banking, “Voice Biometric Technology in Banking | Barclays,” Retrieved February 21, 2016.

[11] Chase-Voice ID:, “Chase Introduces ‘Voice ID’ to Credit Card Customers,” April 19, 2018.

[12] Citi Voice Biometrics:, “Citi Tops 1 Million Mark for Voice Biometrics Authentication for Asia Pacific Consumer Banking Clients,” March 20, 2017.

[13] Bank of American Voice Authetication:, “How do I obtain a voice authorization?” February 12, 2019.

[14] Nuance, “Voice Biometrics for fast, secure authentication in your IVR and mobile apps,” Retrieved February 21, 2016.

[15] Sarita Harbour, “The Voice: Four Ways Voice Command Technology Can Simplify Your Life Right Now,” Retrieved February 12, 2019.

[16] Ewen MacAskill, “Did ’Jihadi John’ kill Steven Sotloff? | Media,” Retrieved February 12, 2019.

[17] Julia Kollewe, “HSBC rolls out voice and touch ID security for bank customers | Business,” Retrieved February 12, 2019.

[18] Phone Losers, “PLA Radio Episode #17 – Voice Authentication,” February 12, 2019.

[19] Lei, X., Tu, G.-H., Liu, A. X., Li, C.-Y., and Xie, T., “The Insecurity of Home Digital Voice Assistants – Vulnerabilities, Attacks and Countermeasures,” in *IEEE Conference on Communications and Network Security*, 2018.

[20] Malik, K. M., Malik, H., and Baumann, R., “Towards Vulnerability Analysis of Voice-Driven Interfaces and Countermeasures for Replay Attacks,” in *The 2nd IEEE International Workshop on “Fake MultiMedia” (FakeMM’19)*, San Jose, CA, USA, 2019.

[21] Malik, H., “Securing Voice-driven Interfaces against Fake (Cloned) Audio Attacks,” in *The 2nd IEEE International Workshop on “Fake MultiMedia” (FakeMM’19)*, San Jose, CA, USA, 2019.

[22] Arik, S., Chen, J., Peng, K., Ping, W., and Zhou, Y., “Neural Voice Cloning with a Few Samples,” in S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pp. 10040–10050, Curran Associates, Inc., 2018.

[23] Arik, S. Ö., Diamos, G. F., Gibiansky, A., Miller, J., Peng, K., Ping, W., Raiman, J., and Zhou, Y., “Deep Voice 2: Multi-Speaker Neural Text-to-Speech,” *CoRR*, abs/1705.08947, 2017.

[24] Ping, W., Peng, K., Gibiansky, A., Arik, S. Ö., Kannan, A., Narang, S., Raiman, J., and Miller, J., “Deep Voice 3: 2000-Speaker Neural Text-to-Speech,” *CoRR*, abs/1710.07654, 2017.

[25] Malik, H. and Miller, J., “Microphone Identification using Higher-Order Statistics,” in *46th AES Conference on Audio Forensics*, Denver, CO, USA, 2012.

[26] Malik, H., “Securing Speaker Verification System Against Replay Attack,” in *46th AES Conference on Audio Forensics*, Denver, CO, USA, 2012.

[27] Andrej Karpathy, “CS231n convolutional neural networks for visual recognition,” 2016.

[28] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., and Zue, V., “TIMIT Acoustic-Phonetic Continuous Speech Corpus,” 1993.