International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 4, pp: 860-865 www.ijaem.net ISSN: 2395-5252

Detection of Fake Voice Generated By GAN Using CVAE

Nisarga J N, Poornima H N, Kavya S N, PavitraBai K
Dr.Hareesh K
Associate Professor
Department of Computer Science and Engineering
Government Engineering College KR Pet-571 426, Mandya, Karnataka
Corresponding Author: Nisarga J N,

Date of Submission: 02-08-2020 Date of Acceptance: 20-08-2020

ABSTRACT— The advent of deep learning generative models enables realistic generation from known data distribution, such as images, videos and sounds. Voice samples generated by such models can used for malicious purposes, i.e. fraud and impersonation if one fails to detect and report them. This poses challenges on the state-of-the-art voice verification systems to identify generated fake voices in order to prevent misuse of fake information. To test established verification systems against fake voices, we obtained a dataset of fake voices by CycleGAN-VC and used it to investigate two verification systems, 1) convolutional VAE, to see if they can detect generated fake voices.

Keywords-Cloned Audio; Generative Adversarial Network (GAN); Mel-Frequency CepstralCoeffients (MFFCs); Convolutional VariationalAutoencoder (CVAE); Voice Verification

I. INTRODUCTION

In recent years, deep learning has been applied to generating realistic media, like images [1] and voices. In the field of voice conversion, a model based on generative adversarial network (GAN)[2] achieved comparable performance as one of the most successful model[3]. At the same time, it can pose problems of malicious use for fake news and scams[4] on online media. As the technology matures, it may also pose issues for the authentication systems that are based on voice recognition. A reflection has to be done: is out speaker verification system strong enough to detect fake media generated by GAN?

To identify fake voice or performance voice authentication, voice verification system is often built and applied. Voice verification is a task where a model is hired to recognize the voice of a specific persons from others in terms of unique individual characteristics. Voice verification systems help to validate if a person matches the claimed identity and can be used as a means for identification security. Such verification systems are therefore trained to distinguish voices of its known individuals from those from unknown sources. The unknown sources can be from one who falsely impersonate others. To identify voices of a specific individual, several voice verification systems have been proposed to this end. Automatic voice verification, also referred to as, automatic speaker verification (ASV) are based on two main approaches, text-dependent and text-independent approaches.

Specifically, a text-dependent ASV requires inputs of fixed phrases for verification, indicating that the system considers not only the audio characteristics but also the language content. Works in text-dependent verification systems often use i-vector-Probabilistic Linear Discriminant Analysis (PLDA) paradigm, such as in [5] and [6].

On the other hand, text-independent systems do not assume such pre-defined phrases and can also work across inputs of different languages. Text-independent systems often take in extracted features such as Mel-frequency cepstral coefficients (MFCCs) and use models such as Gaussian Mixture Models (GMMs) [7]. This has been a classical model as a text-independent ASV. Specifically, the model uses maximum likelihood to estimate the speaker-independent UGB. With the model, one can calculate the likelihood of "background speakers" and also the likelihood of the particular speaker by maximum a posteriori (MAP). A likelihood ratio between them can be used as a criterion to determine the authenticity of specific speakers. Besides UGB, [11] proposed to combine GMMs with support vector machines (SVMs). Recently, deep neural network approaches, such as DNN embeddings[12], [13] or CNNs[14] have been applied to speaker verification systems, revealing comparable performance with the classical Gaussian Mixture Method (GMM)-based models [15], which

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 4, pp: 860-865 www.ijaem.net ISSN: 2395-5252

are supervised approaches and require datasets with labeled fake voice samples for training.

To evaluate a verification system, previous studies made use of datasets including audio samples by humans, such as [16]. While the successful models are able to map voices to their corresponding individuals, it's unlike that two voices from the training data are intentionally made to sound alike. Hence one cannot easily estimate the model performance for such scenarios as such data samples are difficult to obtain. In recent years, deep learning based generative models have been increasingly proposed to generating realistic data. In the scenario of voice conversion, [17] has been proposed using similar structures as CycleGAN[18] and achieved comparable performance to the state-of-the-art model[3]. Using those models, one can easily generate almost realistic voices to intentionally imitate other individuals.

The verification system can be vulnerable to specific types of fake voices. Building a verification system involves training on datasets collected offline and online evaluation on upcoming speakers. In the case where impersonated or synthesized voices come in, the system may be misled and accept such fake voices as the authentic ones.

To evaluate the potential danger of such attacks using deep learninig based voice conversion model, we simulated the attack by 1) generating fake voice with CycleGAN voice conversion model 2) testing the ability of the fake voice to penetrate the conventional verification system. We use the CVAE model as the baseline model to estimate vulnerability to deep-fake-voice attacks.

II. METHODOLOGY

In this section, we shall describe models used for 1) fake voice generation, 2) fake voice detection. For both tasks, we use Mel-frequency cepstral coefficients (MFCCs) as input and/or output of the models, instead of the original voice samples.

A. MFCCs

MFCCs are features of voices based on Fourier transformation, which represents results of a cosine transform of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale. They are widely used in speech and speaker recognition as well as voice conversion. One advantage of using MFCCs is that it can be used to reconstruct the voice as well. Because MFCCs take the format as a two-dimensional image, we can simply apply deep learning techniques commonly used in vision (e.g CycleGAN) on the extracted MFCC features.

The pipeline of extracting MFCC features includes the following steps: [19]:

1) Use discrete Fourier transform (DFT) to turn the windowed speech segment into the frequency domain. Short term frequency spectrum P(f) is obtained

2) Convert the probability measure. Calculate the spectrum P(M) where:

$$M = 2595\log(1+ f /700) \quad (1)$$

3) Convolve P(M) into θ(Mk)(k = 1,2,..K) with a triangular low-pass filter.

$$\theta(M_k) = \sum_z P(M - Mk)\psi(M) \quad (2)$$

4) Finally we use the cosine transform to further compress the representation:

$$MFCC(d) = \sum_k Xk \cos(d(k - 0.5)\pi/K) \quad (3)$$

where $X_k = \ln(\theta(M_k))$ and d = 1,2,..,D (D is often much smaller than K).

B. Fake Voice generation

For the voice conversion, we describe a recently proposed model CycleGAN-VC as the voice conversion system [17]. Denote the voice sample of target speaker as S_{target} and that of source speaker as S_{source}. Instead of converting raw voice samples directly, CycleGAN-VC convert the MFCCs of S_{source} to the MFCCs of S_{target}, and then convert the MFCCs to human voices by WORLD systems [20]. Since the MFCCs can be computed at each time point of the voice, we can construct two-dimensional features using MFCCs, thus the original CycleGAN can be directly applied in the the feature space of MFCCs. In the following sections, we will denote MFCCs as X. The network architecture of the voice conversion system is mainly based on the Gated-CNN layers.

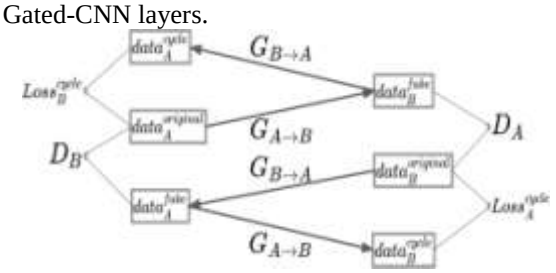

Fig. 1: overview of CycleGAN structure

1) CycleGAN: CycleGAN-VC leverages CycleGAN [18], which has been successfully applied in the image style transformation. CycleGAN is, in short, composed of two GAN networks for style conversion and is optimized with cycle loss force the network to preserve image content during image transformation. An illustration CycleGAN is described Figure 1.

CycleGAN network enables bi-directional style transform between domains A and B. It has two generators, namely $G_{A\to B}$ and $G_{B\to A}$, and two

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 4, pp: 860-865 www.ijaem.net ISSN: 2395-5252

discriminators, namely D_A and D_B. The generator $G_{A\to B}$ take the data of A as input and convert it generate fake data $G_{A\to B}(X_A)$ in the style of B. Likewise for $G_{B\to A}$. The discriminator D_B (and D_A), on the other hand, is optimized to distinguish fake data generated by the $G_{A\to B}$ (and $G_{B\to A}$ respectively) by maximizing the loss,

$$Loss^{adv}(G_{A\to B},D_B)=E_{X\sim P}\ [\log(D_B(X_B))] + E_{XA}\sim P_{data,A}\ [\log(1-D_B(G_{A\to B}(X_A)))].(4)$$

Meanwhile, the generator tries to deceive the discriminator better by minimizing this loss $Loss^{adv}$. At the same time, generator also needs to keep the faithful content of the generated data while converting the style of the data. To keep the content consistent, the following regularizationterm isaddedtothegeneratorlossLoss^{adv}.

$$Loss^{cycle}(GA\to B, GB\to A) = E_{XA}\sim P_{data,A}\ [\|X_A - G_{B\to A}(G_{A\to B}(X_A))\|_1] + E_{XB}\sim P_{data,B}\ [\|X_B - G_{A\to B}(G_{B\to A}(X_B))\|_1]. \quad (5)$$

This regularizer computes a L1 distance between the original data and the data whose style is converted twice as, source style→ target style→ source style. By integrating objectives above, the discriminator is trained by maximizing the following Losstotal, while generator is trained by minimizing Losstotal where,

$$Loss^{total} = Loss^{adv}{}_A + Loss^{adv}{}_B + \lambda^{cycle} Loss^{cycle}. \quad (6)$$

2) Gated CNN: One of the characteristics of speech is that it has sequential and hierarchical time dependencies. Gated CNNs [21] is an effective way to represent such dependency, which not only allows parallel propagation over sequential data but also achieves state-of-the-art in language modeling [21] and speech modeling [22]. A GLU may be a data-driven activation function, and therefore the (l+1)-th layer output Hl+1 is calculated using the l-th layer output Hl and model parameters Wl ,Vl , bl , and cl

$$Hl+1 = (Hl * Wl + bl) \otimes \sigma(Hl * Vl + cl), \quad (7)$$

where$\otimes$ is that the element-wise product and σ is that the sigmoid function. This gated mechanism allows the knowledge to be selectively propagated counting on the previous layer states.

C. Voice verification

1) Convolutional VariationalAutoencoder (CVAE): Variationalautoencoder is another unsupervised deep learning technique used in voice recognition. Simply speaking, we have a probabilistic models for encoder and decoder and encoder extract latent features z from data X and decoder decodes data X from the latent space of z. Theencoder distribution P(z X) is modeled by Gaussian distribution N(μz(X),σ 2 z (X)) where μz(X),σ 2 z (X) are functions represented by neural networks. z is sampled from N(μz(X), σ 2(X)). The decoding distribution P(X z) is also represented by neural network.

If the representative power of encoder network is strong enough, minimizing the negative of right hand side as a loss function actually equivalents to maximizing the probability P(X) for X. In CVAE, the encoding and decoding functions μz(X),σz(X),μx(z) are represented by CNN. To calculate the probability of P(X), we do importance sampling for z. We first sample z from encoder network Q(z|X), then we calculate importance weight N(z|0,I)/N(z|μ(X),σ(X)). Finally, we calculate P(X) as:

$$P(X) = 1/N \sum_z P(X|z) (N(z|0,I))/(N(z|\mu z(X),\sigma 2z(X)))$$

To distinguish generated voice against speaker voice, we can train CVAE on the target speaker voice dataset and uniform background dataset respectively. Then for each testing voice, we can calculate the estimated probability the target speaker as Pspeaker(X) and the background as PUGB(X). Finally, we compute the log-likelihood ratio log (Pspeaker(X))/(PUGB(X). Performing the same hypothesis testing as in GMM-UGB model, one can decide whether the voice is from speaker or fake source.

III. EXPERIMENTAL SETTINGS

• Datasets.
• Training details, e.g. xxx model is trained for xxx iterations, taking xxx hours.
•GPU, e.g. using a single GPU Titan X.

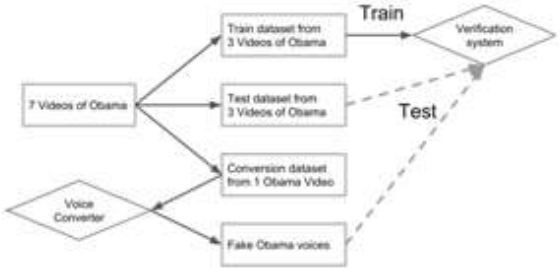

Fig. 2: The overview of the dataset and the experiment.

A. Datasets

For the experiments, we created a dataset of Obama from 7 videos available on YouTube, in which his speech was very clear. And we split the dataset into three datasets for voice conversion, verification and testing as shown in Figure 2. For

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 4, pp: 860-865 www.ijaem.net ISSN: 2395-5252

the voice conversion, the voice data of the source speaker were taken from the dataset of VCC challenge 2018 [24].

For training the verification system, we used a part of the Voxceleb dataset [16], which was obtained by sub-sampling. To train the GMM for target speaker and background model given the hypotheses $P(Y|H_0)$ and $P(Y|H_1)$, we used the MFCC features extracted from the voice datasets of Obama for $P(Y|H_0)$ and used the MFCCs for other arbitrarily selected speakers from [16] for $P(Y|H_1)$.Since manipulating sound data of a large length is inefficient, all datasets are split into a collection of small voice segments of 5-10 seconds.

B. Training of models

The CycleGAN-VC model is computationally heavy. It took 19 hours for training 1000 epochs with a GPU, GTX 1070. The training of the The GMM UBG model consumes large memory and it took 40 GB memory and 4 hours to train on the eular (of HPC Cluster of ETH Zurich) using 8 processors.

C. The network architecture for convolutional VAE

We take similar network architecture as in the work[25]. The decoding network and encoding network are symmetric. Tanh activation function is used for every layer. No layer normalization or batch normalization is applied. The network is trained for 30 epoches using rmsprop optimizer. The step in each epoch is set as 40.

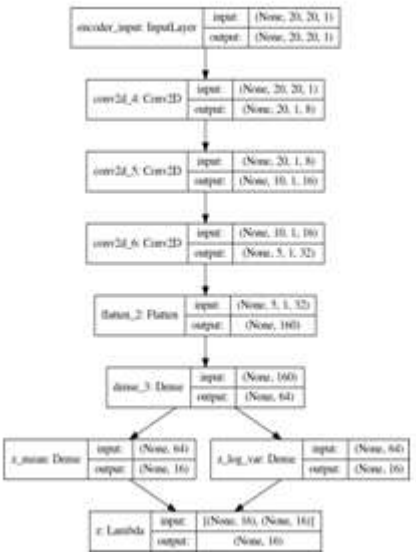

Fig. 3: Encoder Network architecture for the Convolutional Variational Autoencoder. For decoder network, we take the symmetric architecture respect to the encoder network.

IV. RESULTS

We obtained converted fake voices and ensured the generated voices are highly realistic. To analyze the performance of CVAE under deep-fake-voice attacks, we first presented the histograms of log likelihood ratios of the attacks along with those of our target speaker, Barack Obama and background speakers. Then we investigate the performance of models at different levels of realism of voices to reveal possible vulnerability in attacks.

• Generated voices

• Detection performance at different realism levels, (scores and figures)

A. Results of voice verification by CVAE

The estimated log likelihood ratio log Pspeaker(X) Pubg(X) is given in figure 6. The mean of the log likelihood ratio for test dataset is positive while the mean of log likelihood ratio for test dataset is negative. Specifically, the histograms of log likelihood ratio of the target speaker and the background cannot be separated by any threshold. This is inferior to the performance of GMM-UBG verification system which almost perfectly separated the target speaker from the backgroond. This indicates that CVAE fails to capture the characteristics of the voice and cannot be used for conventional voice verification task let alone for fake voice detection.

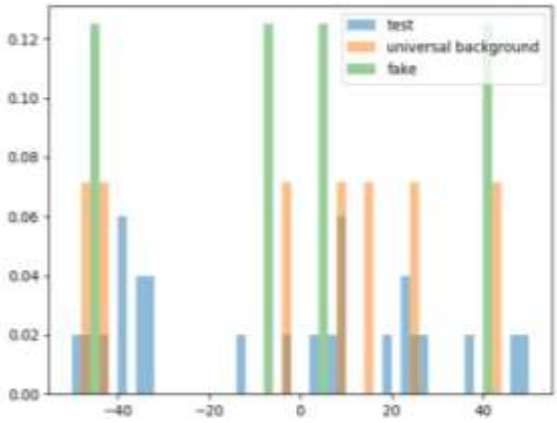

Fig. 4: Log likelihood ratio for CVAE speaker model vs CVAE UBG model for three types of voice segments. Xaxis represents the log likelihood; y-axis represents the corresponding probability density.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 4, pp: 860-865 www.ijaem.net ISSN: 2395-5252

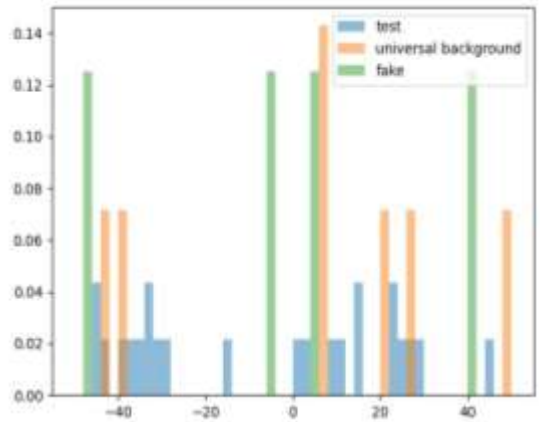

Fig. 5: Log likelihood ratio for CVAE speaker model vs CVAE UBG model for three types of voice segments. Xaxis represents the log likelihood, y-axis represents the corresponding probability density.

V. CONCLUSION

To summarize, we simulated possible attacks by generative models at commonly used automatic voice verification system and reported the performance of verification systems under such attacks. With the obtained results, we identified the vulnerability of conventional verification system and experimentally raised the concerns that generated fake voices by deep learning based model can deceive the verification system. Thus our contribution is to point out the critical danger of the deep fake voice, which was not previously taken seriously or even noticed. We raise alarm over the deep fakes, which can degrade the trustability of the online media and can potentially plunge our society into confusion.

REFERENCES

[1]. T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," arXiv preprint arXiv:1710.10196,2017.

[2]. A.VanDenOord,S.Dieleman,H.Zen,K.Simonyan,O.Vinyals,

a. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet:Agenerativemodelforrawaudio."inSSW,2016,p.125.

[3]. T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," IEEE Transactions on Audio, Speech, and Language Processing, vol.15,no.8,pp.2222–2235,2007.

[4]. S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Syn- thesizing obama: learning lip sync from audio," ACM Transactions on Graphics(TOG),vol.36,no.4,p.95,2017.

[5]. A. Larcher, K. A. Lee, B. Ma, and H. Li, "Phonetically-constrained plda modeling for text-dependent speaker verification with multi- ple short utterances," in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013, pp.7673–7677.

[6]. T. Stafylakis, P. Kenny, P. Ouellet, J. Perez, M. Kockmann,and

[7]. P. Dumouchel, "Text-dependent speaker recognition using plda with uncertainty propagation," matrix, vol. 500, p. 1, 2013.

[8]. D. A. Reynolds, R. C. Rose et al., "Robust text-independent speaker identification using gaussian mixture speaker models," IEEE transac-tionsonspeechandaudioprocessing,vol.3,no.1, pp.72–83,1995.

[9]. N.Dehak,P.J.Kenny,R.Dehak,P.Dumouchel,andP.Ouellet,"Front- end factor analysis for speaker verification," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788– 798,2011.

[10]. S. Prince, P. Li, Y. Fu, U. Mohammed, and J. Elder, "Probabilistic models for inference about identity," IEEE Transactions on Pattern AnalysisandMachineIntelligence,vol.34,no.1, pp.144–157,2012.

[11]. D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," Digital signal processing, vol.10,no.1-3,pp.19–41,2000.

[12]. W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using gmm supervectors for speaker verification," IEEE SignalProcess.Lett.,vol.13,no.5,pp.308–311,2006.

[13]. D. Snyder, D. Garcia-Romero, and D. Povey, "Time delay deep neural network-based universal background models for speaker recognition," in Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on. IEEE, 2015, pp.92–97.

[14]. D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," inProc.Interspeech,2017,pp.999–1003.

[15]. T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N.Evans,

[16]. J.Yamagishi,andK.A.Lee,"Theasvspoof2017challenge:Assessing thelimitsofreplayspoofingattackdetection,"2017.
[17]. F. Tom, M. Jain, and P. Dey, "End-to-end audio replay attack detection using deep convolutional networks with attention," Proc. Interspeech 2018, pp. 681–685,2018.
[18]. A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speakeridentificationdataset,"arXivpreprintarXiv:1706.08612,2017.
[19]. T.Kaneko and H. Kameoka, "Parallel-data-free voice conver- sion using cycle-consistent adversarial networks," arXiv preprint arXiv:1711.11293,2017.
[20]. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to- image translation using cycle-consistent adversarial networks," arXiv preprint,2017.
[21]. Kei Ishikawa, Jingqiu Ding, Xiaoran Chen "Can We Detect Fake Voice Generated by GANs?",2019
[22]. F. Zheng, G. Zhang, and Z. Song, "Comparison of different imple- mentations of MFCC," Journal of Computer Science and Technology, 2001.
[23]. M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," IE- ICE TRANSACTIONS on Information and Systems, vol. 99, no. 7, pp. 1877–1884,2016.
[24]. Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language mod- eling with gated convolutional networks," in Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017, pp.933–941.
[25]. T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, "Sequence- to-sequence voice conversion with similarity metric learned using generative adversarial networks," in Proc. Interspeech, 2017, pp. 1283–1287.
[26]. F.Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin- Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "Atutorial on text-independent speaker verification," EURASIP Journal on Advances in Signal Pro- cessing,vol.2004,no.4,p.101962,2004.
[27]. T. Toda, D. Saito, Z. Ling, F. Villavicencio, J. Yamagishi, J. Lorenzo- Trueba, T. Kinnunen et al., "The voice conversion challenge 2018: database and results,"2018