Can We Detect Fake Voice Generated by GANs?

Kei Ishikawa, Jingqiu Ding, Xiaoran Chen

***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) classical verification system, GMM-UBG model, 2) convolutional VAE, to see if they can detect generated fake voices.**

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], probabilistic models with i-vectors [8][9]. With GMMs being an intensively studied method for this task, [10] proposed to combine GMMs with a universal background model (UGB) as GMM-UGB model. 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 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 classical GMM-UGB 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:

$$M = 2595\log(1+f/700) \tag{1}$$

3) Convolve $P(M)$ into $\theta(M_k)(k=1,2,..K)$ with a triangular low-pass filter.

$$\theta(M_k) = \sum_M P(M-M_k)\psi(M) \tag{2}$$

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

$$MFCC(d) = \sum_k X_k \cos(d(k-0.5)\pi/K) \tag{3}$$

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.

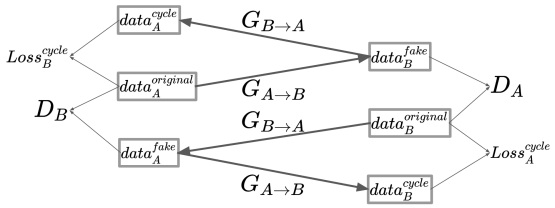

Fig. 1: overview of CycleGAN structure

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 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,

$$\begin{aligned} Loss_B^{adv}(G_{A\to B}, D_B) = \mathbb{E}_{X_B\sim P_{data,B}}[log(D_B(X_B))] \\ + \mathbb{E}_{X_A\sim P_{data,A}}[log(1-D_B(G_{A\to B}(X_A)))]. \end{aligned} \tag{4}$$

Meanwhile, the generator tries to deceive the discriminator better by minimizing this loss $Loss_B^{adv}$. 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 regularization term is added to the generator loss $Loss_B^{adv}$.

$$\begin{aligned} Loss^{cycle}(G_{A\to B}, G_{B\to A}) = \\ \mathbb{E}_{X_A\sim P_{data,A}}[||X_A - G_{B\to A}(G_{A\to B}(X_A))||_1] \\ + \mathbb{E}_{X_B\sim P_{data,B}}[||X_B - G_{A\to B}(G_{B\to A}(X_B))||_1]. \end{aligned} \tag{5}$$

This regularizer computes a L_1 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 $Loss^{total}$, while generator is trained by minimizing $Loss^{total}$ where,

$$Loss^{total} = Loss_A^{adv} + Loss_B^{adv} + \lambda^{cycle} Loss^{cycle}. \tag{6}$$

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]. In a gated CNN, gated linear units (GLUs) are used as an activation function. A GLU is a data-driven activation function, and the $(l+1)$-th layer output

H_{l+1} is calculated using the l-th layer output H_l and model parameters W_l, V_l, b_l, and c_l

$$H_{l+1} = (H_l * W_l + b_l) \otimes \sigma(H_l * V_l + c_l), \tag{7}$$

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

C. Voice verification

1) GMM-UGB: Gaussian mixture model-universal background model (GMM-UGB) and its variant verification systems represent a type of the most commonly used model for voice verification. The model consists of two different GMM models, one is the model for the target speaker and the other is the model for universal background.

By denote a segment of speech $Y \in \mathbb{R}^t$, with t being the temporal length of the speech, and a hypothesis speaker S. The task of determining whether y_i is spoken by S can be formulated as hypothesis testing:
$H_0 : Y$ *is spoken by* S vs $H_1 : Y$ *is not spoken by* S.
Using likelihood ratio test for composite hypothesis, we obtain a test:

$$\frac{P(Y|H_0)}{P(Y|H_1)} \geq \theta \ \textit{accept}\ H_0,\ \textit{reject otherwise}, \tag{8}$$

Here, $P(Y|H_0)$ stands for speaker model while $P(Y|H_1)$ represents universal background model. For UBG-model, an model trained with voice of varieties of people is used instead of comparing with a large collection of speaker model.

Then log likelihood ratio of each time point of the voice segment were averaged and used for deciding whether that voice clip is fake or not. The hyperparameters of the GMM models were chosen according to [23].

2) Convolutional Variational Autoencoder (CVAE): Variational autoencoder 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. The encoder distribution $P(z|X)$ is modeled by Gaussian distribution $N(\mu_z(X), \sigma_z^2(X))$ where $\mu_z(X), \sigma_z^2(X)$ are functions represented by neural networks.z is sampled from $N(\mu_z(X), \sigma_z^2(X))$. The decoding distribution $P(X|z)$ is also represented by neural network.

From the derivation of the evidence lower bound (ELBO), for a given input X, the probability density is lower bounded by

$$\begin{aligned}\log P(X) &\geq \log P(X) - KL(Q(z|x)||P(z|X)) \\ &= \mathbb{E}_{z\sim Q(z|x)} \log P(X|z) - KL(Q(z|X)||P(z)).\end{aligned}$$

Assuming $P(X|z) = N(X|\mu_x(z), \sigma_x^2)$ and $Q(z|x) = N(\mu_z(X), \sigma_z^2(X))$, the right hand side reduces to:

$$-KL(N(\mu_z(X), \sigma_z^2(X))||N(0,I)) - \frac{1}{2\sigma_x^2}\mathbb{E}_{z\sim Q(z|X)}||X - \mu_x(z)||^2 + \log \sigma_x + \text{constant}$$

. 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 $\mu_z(X), \sigma_z(X), \mu_x(z)$ 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 $N(z|0,I)/N(z|\mu(X), \sigma(X))$. Finally, we calculate $P(X)$ as:

$$P(X) = \frac{1}{N}\sum_z P(X|z)\frac{N(z|0,I)}{N(z|\mu_z(X), \sigma_z^2(X))}$$

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 $P_{speaker}(X)$ and the background as $P_{UGB}(X)$. Finally, we compute the log-likelihood ratio $\log \frac{P_{speaker}(X)}{P_{UGB}(X)}$. 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..

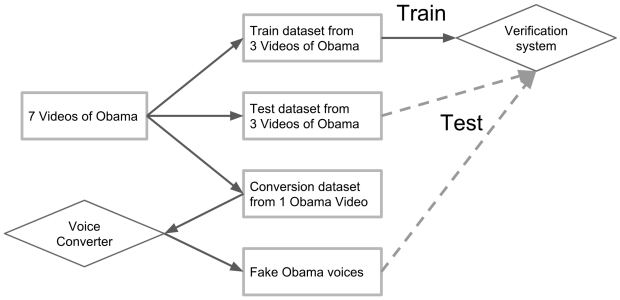

Fig. 2: The overview of the dataset and the experiment.

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 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.

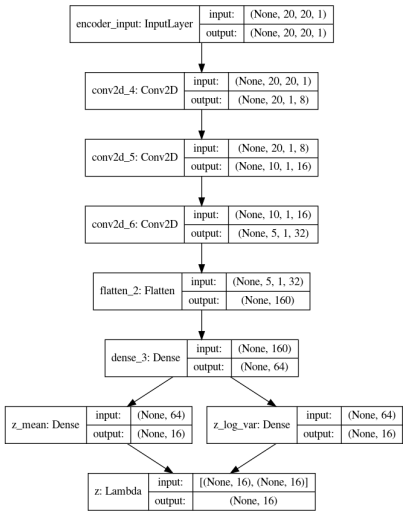

Fig. 3: Encoder Network architecture for the Convolutional Variational Autoencoder. For decoder network ,we take the symmetric architecture respect to the encoder network.

IV. RESULTS

We obtained converted fake voices and ensured the generated voices are highly realistic. To analyze the performance of GMM-UBG and 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 GMM-UBG model

In order to evaluate the ability of the verification system based on GMM-UGB model, we computed the log likelihood for fake voice and the real voice. Figure4 is a plot of mean log likelihood for GMM based verification model. We computed mean of the log likelihood of the MFCC at each time-point of small segment of voice. We obtained histograms for the original speaker, fake voices of this speaker and the universal background. As can be seen in the plot, there can exist a threshold to separate the histograms of the original speaker and the background. This indicates the verification system has been properly trained to distinguish different speaking voices. However, it can be difficult to set such a threshold as to distinguish the fake voices from the original voices of the speaker. Therefore, it is clear that even though GMM-UBG verification system almost perfectly separates the voice clips of Obama (test voice) from the universal background model, it does not separate the fake voice and the test voice. This implies that it is impossible to distinguish the fake voice using conventional GMM based verification system. The results reveal that GMM-UBM, as a common verification system, can be at risk under attacks by such generated samples.

To quantitatively evaluate the performance, we also calculated the area under curve (AUC) score of Precision-Recall (PR) curve of GMM-UBG verification system while the training voice conversion system. To further investigate the failure of detecting fake voice by GMM-UBM, we tested generated samples at different levels of realism. Since audio samples are difficult to visualize, we selected generated voices at different stages of training to represent realism. To obtain different levels of realism, we assume that voice conversion by CycleGAN-VC becomes more accurate as the training proceeds and therefore the realism of generated voices increases. We use voices obtained in an early training stage as the low level of realism, a later stage as the high level of realism, and stages in between as the medium level of realism. Figure 5 shows the AUC scores for fake voice generated by voice conversion system at each stage of training. The score started as high as around 0.90 in early stages and fell to around 0.5 in the middle stages. Since the middle stages, i.e. After 250 epochs and later, the score remains at around 0.50. This implies that GMM-UBG is resilient against deep-fake-voice attacks in the early stage but becomes unreliable in later training.

B. Results of voice verification by CVAE

The estimated log likelihood ratio $\log \frac{P_{speaker}(X)}{P_{ubg}(X)}$ 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 char-

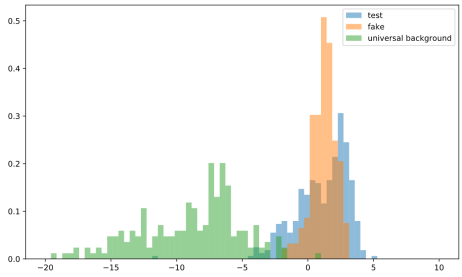

Fig. 4: Histogram of Log likelihood ratio for GMM speaker model vs GMM-UBG model for three types of voice segments. X-axis represents the log likelihood, y-axis represents the corresponding probability density.

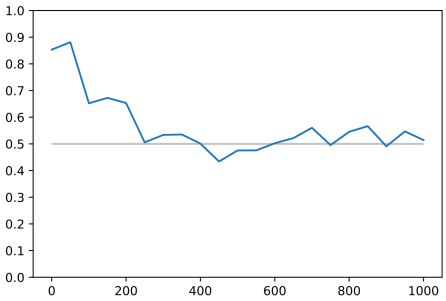

Fig. 5: AUC scores of GMM-UBG verification system. AUC score is computed for the fake voice generated by CycleGAN-VC at every 50 epochs of training from 0 epoch to 1000 epochs.

acteristics of the voice and cannot be used for conventional voice verification task let alone for fake voice detection.

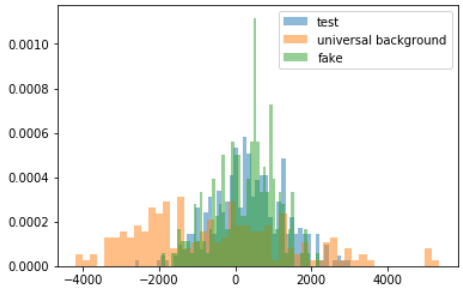

Fig. 6: Log likelihood ratio for CVAE speaker model vs CVAE UBG model for three types of voice segments. X-axis represents the log likelihood, y-axis represents the corresponding probability density.

V. DISCUSSION

A. Failure in detecting fake voice

From the above results of the GMM based verification system GMM-UBG, we can conclude that it is difficult for CycleGAN-VC to detect fake voices via traditional verification methods. This implies the potential of misuse of this voice conversion. For example, fake media can be spread via some Social Media like Facebook or twitter to propagate misleading messages without being detected. We ought to build more powerful verification system to protect the users against such fake information by detecting whether a speech voice comes from human beings or deep learning based conversion system.

B. Failure of CVAE

We tried to make alternative verification system by applying the CVAE. CVAE was explored because it can capture the time dependencies of the MFCC features while GMM-UBG verification model cannot capture the time dependencies of the features. We chose CVAE also for the reason that it is unsupervised and we can compute likelihood of the model like GMM model. However, from the experiment, we found that CVAE is not a successful way to model the voice verification. The reason why CVAE fails to capture the characteristics of the voice might be that it was hard to learn the mapping from low-dimensional latent variable to the high-dimensional MFCC feature space.

C. Possible solution for Fake Voice Detection

In the training process, the goal of the learning system is to distinguish speaker voice and uniform background voice. However, the distribution of fake voice generated by deep learning systems might be very different from the distribution of background voice. Therefore, the probability density of the fake voice samples for speaker and UBG model could be both small, which makes the likelihood ratio between two systems large. In this case, it is difficult for the verification system to rule out the possibility that the fake voice samples come from the speaker.

One possible solution is to add some fake voices generated by deep learning systems to the UBG dataset. We can train a GAN model similar to Cycle-GAN systems for generating fake voices and then generate voice samples from it and add these samples into the UBG dataset.

Another solution might be to train a GAN for generating fake voices and then use the discriminator network for transfer learning.

VI. SUMMARY

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. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio." in *SSW*, 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, "Synthesizing 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 multiple 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 P. Dumouchel, "Text-dependent speaker recognition using plda with uncertainty propagation," *matrix*, vol. 500, p. 1, 2013.

[7] D. A. Reynolds, R. C. Rose *et al.*, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE transactions on speech and audio processing*, vol. 3, no. 1, pp. 72–83, 1995.

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

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

[10] 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.

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

[12] 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.

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

[14] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," 2017.

[15] 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.

[16] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[17] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *arXiv preprint arXiv:1711.11293*, 2017.

[18] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint*, 2017.

[19] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of MFCC," *Journal of Computer Science and Technology*, 2001.

[20] M. Morise, F. Yokomori, and K. 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] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 933–941.

[22] 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.

[23] 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, "A tutorial on text-independent speaker verification," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 4, p. 101962, 2004.

[24] 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.

[25] W. N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2017-Augus, pp. 1273–1277, 2017.