UNIVERSAL ADVERSARIAL PERTURBATIONS GENERATIVE NETWORK FOR SPEAKER RECOGNITION

Jiguo Li[1,2,4], *Xinfeng Zhang*[4], *Chuanmin Jia*[2], *Jizheng Xu*[3], *Li Zhang*[3], *Yue Wang*[3], *Siwei Ma*[2†], *Wen Gao*[2]

[1]Institute of Computer Technology, Chinese Academic of Sciences, jiguo.li@vipl.ict.ac.cn
[2]Peking University, {cmjia, swma, wgao}@pku.edu.cn
[3]Bytedance.Inc, {xujizheng, lizhang.idm, wangyue.v}@bytedance.com
[4]University of Chinese Academic of Sciences, xfzhang@ucas.ac.cn

ABSTRACT

Attacking deep learning based biometric systems has drawn more and more attention with the wide deployment of fingerprint/face/speaker recognition systems, given the fact that the neural networks are vulnerable to the adversarial examples, which have been intentionally perturbed to remain almost imperceptible for human. In this paper, we demonstrated the existence of the universal adversarial perturbations (UAPs) for the speaker recognition systems. We proposed a generative network to learn the mapping from the low-dimensional normal distribution to the UAPs subspace, then synthesize the UAPs to perturbe any input signals to spoof the well-trained speaker recognition model with high probability. Experimental results on TIMIT and LibriSpeech datasets demonstrate the effectiveness of our model.

***Index Terms*—** Universal Adversarial Perturbations, Adversarial Examples Generation, Deep Learning Attack, Speaker Recognition

1. INTRODUCTION

With the success of deep neural networks (DNNs) since Krizhevsky *et al.* [1] won the ImageNet challenge [2] in 2012, more and more deep-based models for biometric systems, such as fingerprint/face/speaker recognition, have been deployed in our daily life. However, these systems are facing the risk of being attacked since deep models are vulnerable to adversarial examples [3], which have been intentionally perturbed. Meanwhile, attacking the deep models and finding the weaknesses of the models can help us avoid the potential risk and design corresponding methods to defense against these attacks. In these widely deployed biometric systems, previous works mainly focus on the vision-based systems, the audio-based systems, such as speaker recognition, have not been well-studied, although the speaker recognition systems have been widely deployed. In this paper, we focus on the attack for speaker recognition models by generating the universal adversarial perturbations (UAPs), which are independent of the input samples and can be applied to the whole dataset.

Before UAPs have been found by Moosavi-Dezfooli *et al.* [4], generating the adversarial examples and spoofing the well-trained deep models have become an emerging topic since Szegedy *et al.* [5] found DNNs are vulnerable to the adversarial examples with intentional imperceptible perturbations. Following [5], some other optimization methods, such as Adam [6], Fast Gradient Sign Method (FGSM) [3] or the genetic algorithm [7] are used to find the perturbations for the input image. Recently, Moosavi-Dezfooli *et al.* [4] demonstrated that there exists a universal and small perturbation that can spoof the well-trained DNN image classifier with high probability. Subsequently, Hayes *et al.* [8] crafted the UAPs by leveraging a generative network to synthesize the perturbation from the input noise which samples from the normal distribution, and improved the attack success rate as well as showed the transferability cross different models for the same dataset. Motivated by the existence of UAPs in the image classification, in this paper, we attempt to find the UAPs of the speaker recognition systems by designing a generative model [8].

In addition to attacking the vision-based systems, the attack for speaker recognition systems has also been addressed for a long time. Before the DNNs have been used in the speaker recognition, the replay and synthesis attacks had been studied to avoid the risk in the voice verification systems [9]. In recent years, with the wide deployment of DNN-based systems, attacking the DNN-based speaker recognition models has drawn more and more attention. Gong *et al.* [10] crafted the adversarial examples using FGSM to attack the well-trained speech verification model and showed the deep models are vulnerable to the adversarial attack. However, the evidence is missing on large-scale datasets [10]. Using the same

The work was done when Jiguo Li interned in Bytedance.Inc. This work was supported by the National Natural Science Foundation of China and Royal Society (61961130392), National Natural Science Foundation of China (61632001), MoE-China Mobile Research Fund Project (MCM20180702), and High-performance Computing Platform of Peking University, which are gratefully acknowledged.

[†]Siwei Ma (swma@pku.edu.cn) is the corresponding author.

978-1-7281-1331-9/20/$31.00 © 2020 IEEE

optimization method, Kreuk *et al.* [11] presented white box attacks for text-dependent speaker verification on the deep end-to-end network on NTIMIT [12] and YOHO [13].

In this paper, we attempt to generate the UAPs by learning the mapping from the low-dimensional normal distribution to the universal perturbation subspace via a generative model, given the fact that the UAPs are not unique [4]. We demonstrate the effectiveness of our proposed method by attacking the state-of-the-art speaker recognition model [14] under non-targeted and targeted settings on TIMIT [15] and LibriSpeech [16] datasets. Our contributions can be summarized as follows:

- We demonstrate the existence of the UAPs for the well-trained speaker recognition model, which are the potential risks for the widely deployed speaker recognition systems in our daily life.
- We can synthesize different UAPs efficiently by mapping the normal distribution into the UAPs subspace using the generative model. The experimental results show that our model can achieve an SER of 97.0 with an SNR of 49.87 and a PESQ of 3.00 in the non-targeted attack on TIMIT dataset, indicating the effectiveness of our proposed model.
- The ablation study for the UAPs shows that our proposed model can learn useful *universal patterns*, map the low-dimensional normal distribution into the UAPs subspace, and generate UAPs that perform much better than the random perturbations.

2. RELATED WORKS

2.1. UAPs Generation

The existence of UAPs have been demonstrated in many areas [17, 18] , since Moosavi-Dezfooli *et al.* [4] found the UAPs in the image classification. Here we mainly review some UAPs generation models on image classification and audio-based systems that are related to our work. Different from the iterative optimization method used in [4], Hayes *et al.* [8] crafted the UAPs by leveraging a generative network to synthesize the perturbation from the input noise which samples from the normal distribution, and improved the attack success rate as well as showed the transferability cross different models for the same dataset. In addition to the works in image classification, some works about UAPs generation for audio-based systems are also proposed recently. Neekhara *et al.* [18] iteratively searched the UAPs with minimal norm under the constraint of high attacking success rate, and only one UAP can be found in once optimization. Our work is different from the above two works in two aspects: (1) our work focuses on the unexplored task, speaker recognition, to study the potential risk of the widely deployed authentication systems; (2) our generative attacker can synthesize different UAPs efficiently once trained, which has been demonstrated more effective than the iterative methods in image classification attacks [8].

2.2. Speaker Recognition Attack

Attacking the speaker recognition models has drawn the researchers' attention because: (1) deep model attacking has become a hot topic in the machine learning community; (2) the speaker recognition/verification systems have been widely deployed in our daily life. Gong *et al.* crafted the adversarial examples iteratively to attack the speaker recognition model trained on a small dataset and demonstrated the existence of the adversarial example for the speaker recognition models. Subsequently, Kreuk *et al.* [11] attempted to fool the end-to-end speaker verification model which is trained on MFCC features by optimizing the perturbation using FGSM [3]. However, these white-box attacks need gradients in the testing phase. In this paper, we proposed a semi-white attack model to learn the UAPs, which is more practical than the white box methods in the real scenario because: (1) our generative attacker needs no gradient in the testing stage; (2) the adversarial perturbations are universal and they can spoof the well-trained speaker recognition model with any input speeches.

3. PROPOSED METHOD

As illustrated in Fig. 1, our generative attacker aims to map the input noise, which is sampled from the low-dimensional normal distribution $\mathcal{N}(0, 1)$, into a UAP, and the following well-trained speaker recognition model is spoofed by the input adversarial example, which is perturbed by the generated UAP. Given a speech s and its speaker label y, the non-targeted attack for the speaker recognition model with UAPs can be formulated as:

$$\arg\min_{s'} L(s, s+\delta) \quad \text{s.t.} \quad f(s+\delta) = y' \qquad (1)$$
$$\text{where } \delta = G_\theta(z) \text{ and } y' \neq y,$$

where G is the generative attack model to synthesize the UAP δ from the noise z, y' is the prediction of the adversarial example $s' = s + \delta$, f is a well-trained state-of-the-art speaker recognition model, L is a distance function to measure the distortion between the raw signal and the adversarial example. For the targeted attack, we modify the constraint for y' from $y' \neq y$ as $y' = y_t$, in which y_t is the target class.

3.1. The Victim Model

We use the state-of-the-art speaker recognition model SincNet [14] as our target victim model. SincNet achieved state-of-the-art performance on TIMIT [15] and LibriSpeech [16] datasets by replacing the first convolution layer as the learnable band pass filters. Given the frequency band $[f_1, f_2]$, the learnable band pass filter can be described as:

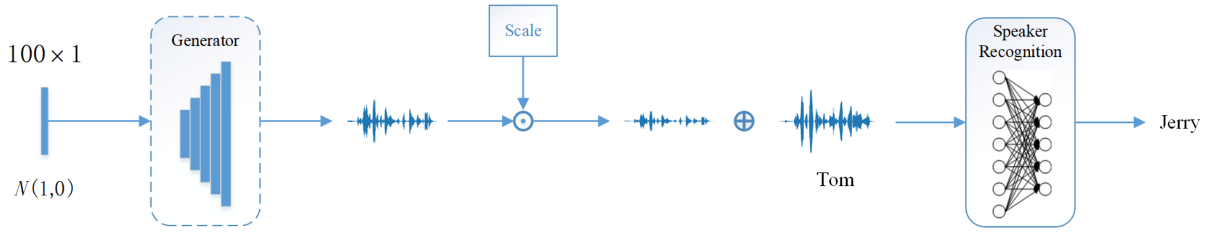

Fig. 1. The framework for our proposed universal adversarial perturbations generative network for speaker recognition.

$$h[n, f_1, f_2] = 2f_2\text{sinc}(2\pi f_2 n) - 2f_1\text{sinc}(2\pi f_1 n), \qquad (2)$$

where $\text{sinc}(x) = \sin x/x$, and f_1, f_2 are the learnable parameters. By using the band pass filters rather than the convolution filters in the first layer, the model is more interpretable and achieves better results [14].

3.2. The Framework

Our model aims to spoof the well-trained speaker recognition model with UAPs. As illustrated in Fig. 1, the *Generator* is a generative network with several upsampling blocks to synthesize the UAP from the input noise with 100 dimensions (following [8]), which samples from the standard normal distribution $\mathcal{N}(0, 1)$. Subsequently, the UAP is scaled to control the distortion for the real data before being added on the input raw speech data with the real label *Tom*. The *Speaker Recognition* model, which is fixed and well-trained, is spoofed by the adversarial examples, which are the input speech data with UAPs, and predicts the input as *Jerry* by mistake.

Since the UAPs are not unique [4], we use a *Generator* to learn the mapping from the normal distribution into the UAP subspace. We use several *UpBlock*s to synthesize the high-dimensional UAPs from the low-dimensional noise, and the convolution layer, batchnorm [19], and ReLU [20] are used in each *UpBlock*.

3.3. Optimization

The optimization objective of our model is to find the adversarial examples with the smallest distortion, and they can attack the well-trained speaker recognition model successfully. Given the input noise z, the raw speech data s and its class label y, the goal above can be formulated as follows:

$$g = R_{x,\delta} - \lambda D_{x,x+\delta}, \qquad (3)$$

where R denotes the attack success **R**ate, D denotes the **D**istortion, λ is the hyper-parameter to get a trade-off between R and D, $\delta = G(z)$ is the UAP. In the optimization this objective function will be maximized.

In non-targeted attacks, attacking successfully means the victim model predicts by mistake, so we can optimize the attack success rate by reducing the prediction probability for the true class, and increasing the prediction probability for any wrong class. To spoof the victim model with minimal cost, we increase the probability for the class which is the top-1 class except for the true class. So R can be formulated as follows:

$$R_{x,\delta} = \begin{cases} \max\limits_{j \neq y} p_j - p_y, & \text{if} \quad R < T \\ \qquad\qquad T, & \text{else}, \end{cases} \qquad (4)$$

where $p = f(x + \delta)$ is the output of pre-softmax layer (logit) with the adversarial example as input, T is a threshold to stop the optimize for this sample. In targeted attacks, the attack is successful as long as the prediction class is the target class. Given the target class t, R can be formulated as follows:

$$R_{x,\delta,t} = \begin{cases} p_t - \max\limits_{j \neq t} p_j, & \text{if} \quad R < T \\ \qquad\qquad T, & \text{else}. \end{cases} \qquad (5)$$

The distortion for UAPs can be measured in two aspects: the objective quality and the perceptual quality. We use Signal-Noise Ratio (SNR) and the Perceptual Evaluation of Speech Quality (PESQ) score [21] to evaluate the quality of the adversarial examples with perturbations in objective and perceptual, respectively. SNR is defined as:

$$\text{SNR}(x, x') = 10 \log_{10} \frac{\|x\|_2}{\|x - x'\|_2}, \qquad (6)$$

where $x' = x + \delta$ is the adversarial example with the perturbation δ. PESQ, as an ITU-T recommendation standard [22], is an integrated model to measure the distortion for the speech in telephony. It is a full-reference algorithm with range $[-0.5, 4.5]$ to measure the perceptual quality of the speech after a temporal alignment. It is worth mentioning that PESQ is not differentiable, so we only use it in the testing phase. In the training phase, D is only the SNR and we just optimize SNR by minimizing the L_2 norm of the perturbations.

3.4. Inference

The inference is not intuitive because the input data are with variable lengths. In the training phase, we can clip the data into slices with a fixed length, but in the testing phase, we can not just drop the data beyond the UAPs. In our implementation, we use a simple but effective method *repeat+clip* to repeat the UAP until it is longer than the input sample and then clip it to make them two matched. In our experiments, we will conduct comparison experiment to study the influence of different UAP lengths.

4. EXPERIMENTAL RESULTS

4.1. Datasets and Metric

Datasets: Following [14], which proposed our victim model, we conduct the experiments on TIMIT [15] (462 speakers totally) and LibriSpeech [16] (2484 speakers totally) datasets. The training/testing split follows the official implementation of [14], in which 2310/1386 samples are used for training/testing in TIMIT, and 14481/7452 samples are used for training/testing in LibriSpeech.

Metric: We use the sentence error rate (SER) to represent the attack success rate in the non-targeted attack, and the prediction target rate (PTR) is used in targeted attacks because the attack is successful as long as the prediction is the target class in the targeted attack. The distortion is measured by SNR (for objective quality) and PESQ (for perceptual quality), as introduced in subsection 3.3. We use the official open-source implementation [1] for PESQ in our experiments.

4.2. Implementation Details

Our *Generator* can only synthesize UAPs with the fixed length, so we randomly select a slice with a fixed length from the raw speech data in training phase. In our experiments, we synthesize UAPs for 200ms, which is 3200 dimensional because the data are with a sampling rate of 16000. We use the pretrained victim model which is released by the author of [14] [2]. The hyper-parameter λ will be finetuned in our experiments and the scale factor is fixed as 1 because the perturbations are constrained on a small scale by the distortion item in our optimization objective. The threshold T for non-targeted/targeted attack is set as 10/0 after being finetuned to get a good trade-off between R and D. Besides, we initialize the biases and weights of the last convolution layer as zero to ensure that no perturbation is added on the signals at the beginning of the training[3].

4.3. Non-Targeted Attack

We conduct the non-targeted attack on TIMIT and LibriSpeech datasets to demonstrate the effectiveness of our proposed model. The results of these two datasets are illustrated in Table. 1. We can observe from the results that:

Table 1. Non-targeted attack on TIMIT/LibriSpeech dataset

Dataset	λ	SER(%)↑	SNR(dB)↑	PESQ↑
TIMIT	-	1.52*	-	-
	500	97.5	44.13	2.13
	1000	94.9	46.76	2.40
	1500	97.0	49.87	3.00
	2000	93.9	49.77	2.92
	2500	92.4	50.49	2.79
Libri	-	0.30*	-	-
	1500	99.7	28.87	2.09
	2000	97.0	30.62	2.22
	2500	96.3	31.15	2.33
	3000	93.7	32.72	2.45
	3500	94.7	33.68	2.54

* Error rate without attack on TIMIT/LibriSpeech.

- For non-targeted attacks, the UAPs exist and our model manages to map the normal distribution into the UAPs subspace because our model can synthesize the UAPs which can attack the well-trained speaker recognition model with high success rate.
- On the TIMIT dataset, with $\lambda = 1500$, the UAPs generated by our model can attack the well-trained speaker recognition model with an SNR of 49.87dB and a PESQ of 3.00, which means that the noise is noticeable but not intrusive.
- On the LibriSpeech dataset, with $\lambda = 2500$, the UAPs generated by our model can attack the well-trained speaker recognition model with an SNR of 31.15dB and a PESQ of 2.33, which means that the noise is noticeable and a little intrusive.
- On both TIMIT and LibriSpeech datasets, by tuning λ, we can control the trade-off between the attack success rate and the adversarial example quality.

It is worth mentioning that *the random perturbations in non-targeted attack can also achieve a high SER as long as the perturbations are intense enough*, so a high SER here cannot provide enough evidence that our model has learned some useful *universal patterns*. In Section 5, we will compare the UAPs generated by our model with the random perturbations to show that our model has learned the *universal patterns*.

4.4. Targeted Attack

In this subsection, we show our model's effectiveness on the targeted attack. In the targeted attack, we fix λ as 3000/2000 for TIMIT/LibriSpeech dataset, respectively. We randomly select 5 speakers from TIMIT/LibriSpeech dataset as the targets, and attack the victim model to misclassify any input sample as the target class. The attack results are illustrated in Table. 2. Some conclusions can be drawn from the results:

[1] https://github.com/dennisguse/ITU-T_pesq

[2] https://github.com/mravanelli/SincNet

[3] The code, data, and pretrained models will be released soon.

Table 2. Targeted attack on TIMIT/LibriSpeech dataset

Dataset	Target	PTR(%)↑	SNR(dB)↑	PESQ↑
TIMIT	0	99.1	48.09	2.49
	100	98.4	48.86	2.41
	200	98.0	48.55	2.52
	300	93.9	48.93	2.42
	400	96.6	48.20	2.55
	avg	97.2	48.53	2.48
Libri	0	64.5	29.73	2.10
	500	40.8	30.73	2.14
	1000	64.9	30.13	2.08
	1500	54.9	29.75	2.07
	2000	95.5	29.38	2.18
	avg	64.1	29.94	2.11

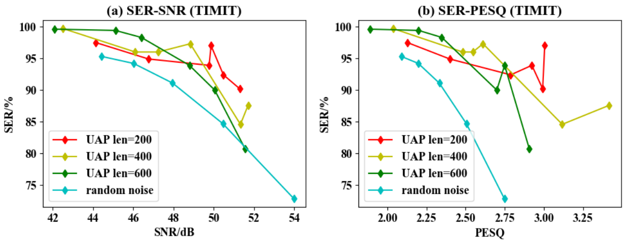

Fig. 2. The influence of the UAPs' length.

- For the targeted attack, the UAPs exist and our model is successful to synthesize UAPs for the targeted attack on both TIMIT and LibriSpeech datasets.
- On the TIMIT dataset, we can achieve a PTR of 97.2% on average with an SNR of 48.53dB and a PESQ of 2.48, which means that the noise is noticeable, and a little intrusive, given the fact that the targeted attack is more challenging than the non-targeted attack.
- On the LibriSpeech dataset, we can achieve a PTR of 64.1% on average with an SNR 29.94dB and a PESQ of 2.11. This is not as good as that on the TIMIT dataset, because the speaker number in LibriSpeech (2484) is much more than that in TIMIT dataset (462).

Besides, *the high success rate here can demonstrate that our model has learned the universal patterns for the targeted attack*. Although random perturbations are able to achieve a high SER in non-targeted attack, they will fail to achieve a high PTR in targeted attack because they are random. Thus a high PTR in targeted attack can demonstrate that our model has learned the useful *universal patterns*.

5. ABLATION STUDY

5.1. The Length of UAPs

We can only generate UAPs with a fixed length, but the input signals are with variable lengths. So we use *repeat+clip* method to make them two matched. The length of the UAPs may affect the performance of our model, so in this subsection, we conduct experiments to study how the UAP length influences the attacking performance. We generate the UAPs with duration 200ms, 400ms, and 600ms on TIMIT dataset, and we plot the SER-SNR and SER-PESQ curves to take both the adversarial examples quality and the attack success rate into account. Besides, we compare our generated UAPs with the *random noise*, which are the random perturbations sampled from the normal distribution $\mathcal{N}(0, \sigma^2)$ (σ is tuned to get five results with different trade-offs between SNR/PESQ with SER).

As illustrated in Fig. 2, a curve is higher than another means that this model can achieve a higher SER with the same SNR/PESQ, indicating this model performs better than the other. From Fig. 2 (a), we can observe that: (1) the UAPs generated by our model perform better than the random perturbations, indicating that our model has learned the useful *universal patterns*; (2) with SNR below 50dB, UAPs with different lengths achieve comparable performance, but with SNR higher 50dB, the shorter the UAPs are, the better performance they can achieve; (3) with UAPs length as 600ms, UAPs generated by our model may perform not as good as the random perturbations when SNR is higher than 52dB, the reason may be *the models for longer UAPs are more difficult to train but we train all models for different UAPs lengths for the same epochs to make a fair comparison*. From Fig. 2 (b), similar conclusions can be drawn except that the UAPs generated by our model performs much better with PESQ higher than 2.5 because the random perturbations struggle to achieve good perceptual quality (PESQ). On both the objective quality (SNR) and perceptual quality (PESQ), the UAPs generated by our model perform better than the random perturbations, demonstrating that our model has learned the useful *universal patterns* to attack the well-trained speaker recognition model.

5.2. Noise Interpolation

To demonstrate our model is able to map the noise into the UAP subspace, we synthesize UAPs from the noise which is interpolated from two random noises and evaluate these UAPs on attacking the well-trained speaker recognition model. With an interpolation parameter β, the interpolation noise z' can be obtained by: $z' = \beta z_1 + (1 - \beta) z_2$, where z_1 and z_2 are two low-dimensional noise vectors sampled from the normal distribution $\mathcal{N}(0, 1)$. As illustrated in Table 3, with different β, the UAPs generated by our model can achieve similar SER, SNR, and PESQ, validating our model can map $\mathcal{N}(0, 1)$ into the UAPs subspace indirectly.

6. CONCLUSION

In this paper, we attempted to demonstrate the existence of the UAPs for the speaker recognition, and we proposed a genera-

Table 3. Noise interpolation results (SER) on TIMIT datasets

β	0	0.2	0.4	0.6	0.8	1.0
SER(%)↑	98.3	97.2	96.9	94.5	97.5	98.0
SNR(dB)↑	49.6	50.0	50.1	50.5	49.8	49.5
PESQ↑	2.99	3.02	3.04	3.05	3.01	2.99

tive network to map the low-dimensional noise space into the UAPs subspace to synthesize the UAPs efficiently. Experimental results showed that our model can generate UAPs and fool the state-of-the-art speaker recognition model with high success rate. The ablation study provided enough evidence to show that our model had learned useful *universal patterns* for attacking the well-trained speaker recognition model. We envision our work to provide a benchmark for universal attacks for speaker recognition.

7. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems(NeurIPS)*, 2012, pp. 1097–1105.

[2] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., “Imagenet large scale visual recognition challenge,” *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[3] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy, “Explaining and harnessing adversarial examples,” in *ICLR*, 2015.

[4] Seyed Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard, “Universal adversarial perturbations,” in *CVPR*, 2017.

[5] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, “Intriguing properties of neural networks,” in *ICLR*, 2014.

[6] Nicholas Carlini and David Wagner, “Towards evaluating the robustness of neural networks,” in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 39–57.

[7] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai, “One pixel attack for fooling deep neural networks,” *IEEE Transactions on Evolutionary Computation*, 2019.

[8] Jamie Hayes and George Danezis, “Learning universal adversarial perturbations with generative models,” in *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2018, pp. 43–49.

[9] Pavel Korshunov, Sébastien Marcel, Hannah Muckenhirn, André R Gonçalves, AG Souza Mello, RP Velloso Violato, Flávio O Simoes, M Uliani Neto, Marcus de Assis Angeloni, José Augusto Stuchi, et al., “Overview of btas 2016 speaker anti-spoofing competition,” in *BTAS*. IEEE, 2016, pp. 1–6.

[10] Yuan Gong and Christian Poellabauer, “Crafting adversarial examples for speech paralinguistics applications,” *DYNAMICS Workshop*, 2018.

[11] Felix Kreuk, Yossi Adi, Moustapha Cisse, and Joseph Keshet, “Fooling end-to-end speaker verification with adversarial examples,” in *ICASSP*. IEEE, 2018, pp. 1962–1966.

[12] Charles Jankowski, Ashok Kalyanswamy, Sara Basson, and Judith Spitz, “Ntimit: A phonetically balanced, continuous speech, telephone bandwidth speech database,” in *ICASSP*. IEEE, 1990, pp. 109–112.

[13] Joseph P Campbell, “Testing with the yoho cd-rom voice verification corpus,” in *ICASSP*. IEEE, 1995, vol. 1, pp. 341–344.

[14] Mirco Ravanelli and Yoshua Bengio, “Speaker recognition from raw waveform with sincnet,” in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[15] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett, “Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1,” *NASA STI/Recon technical report n*, vol. 93, 1993.

[16] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *ICASSP*. IEEE, 2015, pp. 5206–5210.

[17] Xing Wu, Lifeng Huang, and Chengying Gao, “G-uap: Generic universal adversarial perturbation that fools rpn-based detectors,” *Proceedings of Machine Learning Research*, vol. 101, pp. 1204–1217, 2019.

[18] Paarth Neekhara, Shehzeen Hussain, Prakhar Pandey, Shlomo Dubnov, Julian McAuley, and Farinaz Koushanfar, “Universal adversarial perturbations for speech recognition systems,” *Proc. Interspeech 2019*, pp. 481–485, 2019.

[19] Sergey Ioffe and Christian Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *ICML*. 2015, ICML’15, p. 448–456, JMLR.org.

[20] Vinod Nair and Geoffrey E Hinton, “Rectified linear units improve restricted boltzmann machines,” in *ICML*, 2010, pp. 807–814.

[21] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, “Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs,” in *ICASSP*. IEEE, 2001, vol. 2, pp. 749–752.

[22] “Perceptual evaluation of speech quality (pesq), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs,” *TTU-T Draft Recommendation P.862*, May, 2000.