INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Inaudible Adversarial Perturbations for Targeted Attack in Speaker Recognition

*Qing Wang, Pengcheng Guo, Lei Xie**

Audio, Speech and Language Processing Group (ASLP@NPU),
School of Computer Science, Northwestern Polytechnical University, Xi'an, China
{qingwang, pcguo, lxie}@nwpu-aslp.org

Abstract

Speaker recognition is a popular topic in biometric authentication and many deep learning approaches have achieved extraordinary performances. However, it has been shown in both image and speech applications that deep neural networks are vulnerable to adversarial examples. In this study, we aim to exploit this weakness to perform targeted adversarial attacks against the x-vector based speaker recognition system. We propose to generate inaudible adversarial perturbations based on the psychoacoustic principle of frequency masking, achieving targeted white-box attacks to speaker recognition system. Specifically, we constrict the perturbation under the masking threshold of original audio, instead of using a common l_p norm to measure the perturbations. Experiments on Aishell-1 corpus show that our approach yields up to 98.5% attack success rate to arbitrary gender speaker targets, while retaining indistinguishable attribute to listeners. Furthermore, we also achieve an effective speaker attack when applying the proposed approach to a completely irrelevant waveform, such as music.
Index Terms: targeted adversarial attack, inaudible, adversarial example, speaker recognition

1. Introduction

In recent years, attacks and defenses of speaker recognition systems have attracted more and more attention. As one of the most prominent biometric authentication methods, the security of speaker identification system is extremely important. Prior works have found speaker recognition systems are not only facing the spoofing attacks [1–3] including impersonation, replay, speech synthesis, as well as voice conversion, while adversarial attacks are also be able to affect speaker recognition systems. In [4], Das *et al.* gave an overview of the attacker's perspective on speaker verification.

Adversarial attacks are usually conducted by adversarial examples, which are designed by constructing imperceptible perturbations to lead a mis-classification. Adversarial examples were first proposed by Szegedy *et al.* [5] in computer vision tasks, which show that a certain network is vulnerable to a crafted small perturbation in the training set. Goodfellow *et al.* [6] proposed an effective approach, fast gradient-sign method (FGSM), to generate adversarial examples through the linearization of the loss function. Since then, various experimental results have shown that adversarial examples can successfully influence a variety of models [7, 8].

Apart from the applications in image tasks, speech-related tasks could also be affected by adversarial examples [9, 10]. There has been plenty of work focused on attacking automatic speech recognition (ASR) systems using adversarial examples. In [11], Carlini *et al.* demonstrated the effectiveness of targeted audio adversarial examples on a end-to-end ASR system. With optimization-based attacks, they were able to turn any audio waveform into any target transcription. Moreover, adversarial examples are able to worsen a keyword spotting system [12, 13]. Instead of using a l_p norm to measure the maximum perturbation introduced as above, Schönherr *et al.* [14] introduced a new type of adversarial examples based on psychoacoustic hiding and attacked the Kaldi ASR system [15] successfully. Next, Qin *et al.* [16] extended this idea and developed effectively imperceptible audio adversarial examples by leveraging the psychoacoustic principle of auditory masking.

In speaker recognition area, some researchers have applied adversarial learning [17–20], while adversarial examples could also be used to attack and to defend the system. In [21], Kreuk *et al.* used adversarial examples for fooling a speaker verification (SV) system by adding a peculiar noise to the original speaker examples. In our previous work [22], we added adversarial perturbations on feature-level to conduct a non-targeted attack to SV system. We also explored using adversarial examples for model regularization and improved the robustness of the SV system. Xie *et al.* [23] made the DNN based speaker recognition system can identify the speaker as any target label by adding audio-agnostic universal perturbations on speakers' voice input. In [24], Li *et al.* proposed to generate universal adversarial perturbations (UAPs) by learning the mapping from the low-dimensional normal distribution to the universal perturbation subspace via a generative model. However, the aforementioned adversarial examples are mostly restricted to make a slight change of original signal in form of audio sampling points, without considering the human perceptibility of sound.

In this study, we are inspired by the work in [14, 16] and propose to generate inaudible adversarial perturbations for targeted attacking speaker recognition directly on wave-level. We use the structure of x-vector speaker recognition system proposed in [25] as our baseline to conduct targeted white-box attacks. To generate inaudible adversarial perturbations, we adopt the frequency masking concept where one faint but audible sound becomes inaudible in the presence of another louder audible sound. Our experimental results based on Aishell-1 [26] corpus demonstrate that the inaudible adversarial perturbations can achieve better targeted attack performance than previous l_p norm based adversarial examples. To further compare the frequency masking based approach with previous ones, we also evaluate them from both subjective and objective metrics. Results show that the adversarial perturbations generated by proposed methods are more inaudible, even with larger absolute energy. Finally, we attempt to conduct targeted attacks using the music portion of the MUSAN corpus [27], which is a completely irrelevant non-speech dataset. Experiments show even

*corresponding author

http://dx.doi.org/10.21437/Interspeech.2020-1955

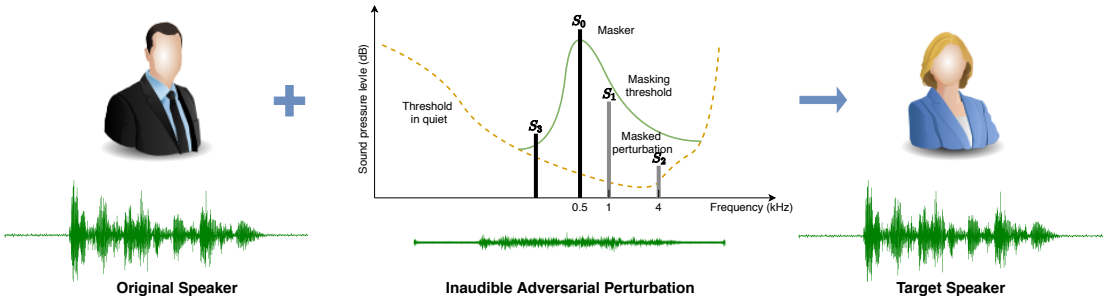

Figure 1: *An overview of the generation of adversarial examples based on frequency masking.*

non-speech can also achieve a high speaker attack success rate.

2. Inaudible adversarial perturbations

In this section, we introduce how we generate the inaudible perturbations that can conduct targeted speaker attacks. Figure 1 shows an overview of the generation of adversarial examples base on frequency masking.

2.1. Adversarial example generation

An adversarial example is defined as an instance with imperceptible, intentional perturbation that causes a well-trained model to make a false prediction. Conventional approaches to generate adversarial perturbations are typically by performing gradient descent w.r.t the input sample. Specifically, given an input speech x, its label speaker y, an arbitrary target label y' and a well-trained speaker recognition model $f(\cdot)$, the adversarial perturbation δ can be generated by

$$\begin{aligned} &\min L_{CE}(f(x+\delta), y'), \\ &s.t. \quad \|\delta\| < \epsilon, \end{aligned} \tag{1}$$

where $y' \neq y$ and $L(\cdot)$ is the loss function. The hyperparameter ϵ is used to control the maximum perturbation generated.

2.2. Frequency masking

Our goal is to generate indistinguishable adversarial perturbations in the human perceptibility of audio, instead of maintaining a slight noise to the clean speech sample points. In order to achieve that, we utilize the idea of *frequency masking*, which refers to the phenomenon that one faint but audible sound (the maskee) becomes inaudible in the presence of another louder audible sound (the masker) [28]. Therefore, we can modify adversarial perturbations to be inaudible, as long as the perturbation falls under the masking threshold of the original speech. In [28], Lin *et al.* investigated the algorithm of computing masking threshold, which consists of 3 steps.

STEP 1: Identifications of maskers

In order to obtain the frequency masking threshold of the original speech, raw audio signals from the time domain are first converted into time-frequency representations by short-time Fourier transform (STFT). The output of STFT $s_x(k)$ refers the k-th bin of the spectrum at frame x. Then, the power spectral density (PSD) of $s_x(k)$ can be computed as

$$P_x(k)/\text{dB} = 10\log_{10}|\frac{1}{N}s_x(k)|^2. \tag{2}$$

After that, the PSD estimate $P_x(k)$ is normalized to a sound pressure level (SPL) of 96 dB,

$$\overline{P}_x(k)/\text{dB} = 96 - max\{P_x(k)\} + P_x(k). \tag{3}$$

The normalized PSD estimate of reasonable maskers must satisfy three constraints. First is local maxima,

$$\overline{P}_x(k) \geq \overline{P}_x(k+1) \quad \text{and} \quad \overline{P}_x(k) \geq \overline{P}_x(k-1). \tag{4}$$

Secondly, they should be larger than the absolute threshold of hearing (ATH),

$$\overline{P}_x(k) \geq \text{ATH(k)}. \tag{5}$$

Finally, any group of maskers should keep a maximum amplitude within 0.5 Bark (a psychoacoustically-motivated frequency scale) and only the masker with the highest SPL is retained,

$$\overline{P}_{x_1,x_2}(k) = \underset{k_0\in[-0.5,0.5]}{\arg\max}\ \overline{P}_{x_1,x_2}(k+k_0). \tag{6}$$

Since the masking effect is additive in the logarithmic domain, the SPL of each masker can be further smoothed by

$$\overline{P}_x(\overline{k}) = 10\log_{10}[10^{\frac{\overline{P}_x(k-1)}{10}} + 10^{\frac{\overline{P}_x(k)}{10}} + 10^{\frac{\overline{P}_x(k+1)}{10}}]. \tag{7}$$

STEP 2: Calculation of individual masking thresholds

An individual masking threshold $T[b(j), b(i)]$ means that the masker at frequency index j contributes to the masking effect on the maskee at frequency index i, where $b(j)$ and $b(i)$ are the masker and maskee's frequencies in Bark scale. The individual masking thresholds can be calculated as:

$$T[b(j), b(i)]/\text{dB} = \overline{P}_x[b(j)] + \Delta[b(j)] + \text{SF}[b(j), b(i)], \tag{8}$$

where $\Delta[b(j)] = -6.025 - 0.275b(j)$ and $\text{SF}[b(j), b(i)]$ is a two-slop spread function.

STEP 3: Calculation of global masking threshold

After the individual masking thresholds are obtained, the global masking threshold can be calculated by combining them with the absolute threshold of hearing. The global masking threshold at frequency index i is calculated according to

$$T_G(i)/\text{dB} = 10\log_{10}[10^{\frac{\text{ATH(i)}}{10}} + \sum_{j=1}^{N_M} 10^{\frac{[b(j),b(i)]}{10}}], \tag{9}$$

where ATH(i) is the SPL of threshold in quiet at frequency index i, N_M is the number of maskers, and $T[b(j), b(i)]$ is corresponding individual masking threshold. Readers can get more detail about the calculation of masking threshold in [28].

2.3. Optimization procedure

Given an input speech x, its label speaker y, an arbitrary target speaker label y', where $y \neq y'$, and a well-trained x-vector speaker recognition model $f(\cdot)$, the additional loss function to modify the perturbation fall under the masking threshold can be defined as

$$L_{TH}(x, \delta) = \mathbb{E}_k \max\{\overline{P}_\delta(k) - T_G(k), 0\}, \quad (10)$$

where $\overline{P}_\delta(k)$ means the normalized PSD estimated of δ at the k-th frequency bin. The inaudible adversarial perturbation δ can be generated by

$$\min L(x, \delta, y') = L_{CE}(f(x+\delta), y') + \alpha \cdot L_{TH}(x, \delta), \quad (11)$$

where L_{CE} aims to make the adversarial examples fool the well-trained speaker recognition system into predicting an arbitrary target label and the L_{TH} constrains the normalized PSD estimate of perturbation to be inaudible. The α is a hyper-parameters to scale different losses.

The whole optimization procedure is separated into two stages. In **Attack Stage1**, we focus on finding a relative small perturbation using a common l_p norm based algorithm as defined in Eq. (1). The δ is initialized to a zero vector and ϵ is gradually reduced from a large value. For each iteration, δ is updated by

$$\delta \leftarrow clip_\epsilon(\delta - lr_1 \cdot sign(\nabla_\delta L_{CE}(f(x+\delta), y'))). \quad (12)$$

In **Attack Stage2**, we further optimize above perturbation by introducing frequency masking based loss as defined in Eq. (11). The α starts from 0.05 and adaptively updated based on the performance of attack. For each iteration, δ is updated to be inaudible through:

$$\delta \leftarrow \delta - lr_2 \cdot \nabla_\delta L(x, \delta, y'). \quad (13)$$

3. Experimental setup

3.1. Dataset

We use the Mandarin Aishell-1 corpus [26] as the evaluation data set. The entire corpus contains 400 speakers (214 female, 186 male), sampled at 16kHz, including training, development and test sets, without speaker overlapping. Training set is used in x-vector baseline training, while test set is used to evaluate the baseline system. For conducting inaudible adversarial targeted attacks, we randomly choose 10 female speakers (denoted as F) and 10 male speakers (denoted as M) from the training set, each with 100 utterances, as the original speaker set. Another 10 female speakers (denoted as F') and 10 male speakers (denoted as M') are selected as the attack targets. We assign these selected sets into 4 test modes. The first one is using 10 male original speakers to attack 10 male target speakers, denoted as M2M'. Similarly, the other three test modes are M2F', F2M' and F2F'.

Besides, we use the music portion of MUSAN [27] corpus as our non-speech dataset, which consists of western art music (e.g., Baroque, Romantic, and Classical) and popular genres (e.g., jazz, bluegrass, hip-hop, etc). We randomly choose 200 pieces of western art music and cut them into 1000 pieces of 6 seconds short segments. This subset is used as the original wave to attack the selected male target speakers.

3.2. Experimental setup

3.2.1. Baseline

We use x-vector system [25] as our baseline. The 30-dimensional Mel-frequency cepstral coefficients (MFCC) features are extracted as the input for all experiments. The configuration of x-vector network is exactly the same as in [25]: a 5-layer TDNN with ReLU followed by batch normalization is used for extracting frame-level hidden features. The number of hidden nodes is 512 and the dimension of frame-level hidden features for pooling is 1500. Each frame-level feature is generated from a 15-frame context of acoustic features. Pooling layer aggregates frame-level features, followed by 2 fully-connected layers with ReLU activation functions, batch normalization, and a Softmax output layer. The EER of the x-vector baseline system is 4.27%. Note that we use the whole sentence as input instead of using chunks as in [25], because we need to compute the gradient w.r.t the sentence-level perturbation.

After training the x-vector baseline system, we calculate the speaker prediction for the original utterances with their true labels. The accuracy for the M set is 95.9%, while the accuracy for the F set is 97.9%. We also calculate the prediction accuracy for the original utterances with assigned target speakers. All the results of the four test modes are 0.00%.

3.2.2. Inaudible adversarial perturbations

We first compute the STFT of original speech to get the time-frequency representations. The window type of STFT is the modified Hann window with a length of 2048 and a hop length of 512. In Attack Stage1, the learning rate lr_1 is set to be 100 and the δ will be updated 3000 times for each mini-batch. We use the l_∞ norm to measure the perturbation bound. The ϵ starts from 2000 and will multiply 0.8 when attacking successfully. In Attack Stage2, the learning rate lr_2 is 1 and the total training step for each mini-batch is 1000. The scale parameter α begins with 0.05 and will increase to 1.2α when attacking successfully or decrease to 0.8α when fails. All systems are implemented using PyTorch [29] and optimized by Adam optimizer [30].

3.2.3. Evaluation metrics

We use various metrics to measure the performance of proposed method. First, we compute the attack success rate to evaluate the performance of targeted attacks in speaker recognition. Formally, the accuracy is computed as:

$$Acc = N_s/N, \quad (14)$$

where N is the total number of utterances we used to test and N_s refers to the number of utterances attacking. Besides, perceptual evaluation of speech quality (PESQ) [31] and signal-to-noise ratio (SNR) are also computed to measure the distortion of generated adversarial examples. Finally, we also conduct a subjective evaluation to evaluate the adversarial examples from the human perceptibility of audio. successfully.

4. Experimental results and analysis

4.1. Inaudible adversarial targeted attack

In Table 1, we calculate the attack success rate for all the four test modes. As we separated the optimization procedure into two stages in Section 2.3. We will test the adversarial examples generated in these two stages, donated as Attack Stage1 and Attack Stage2, respectively. System Attack Stage1 is we

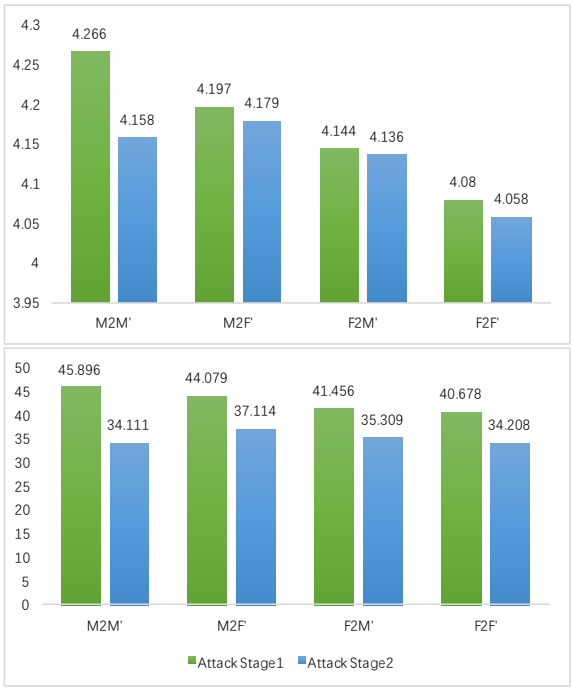

Figure 2: *Average PESQ and SNR (dB) comparison of Attack Stage1 and Attack Stage2 on each test mode.*

conduct attack using the adversarial examples generated in Attack Stage1, which just focus on finding a small perturbation. And the targeted attack successfully affected the speaker model in 72.6%, 73.8%, 73.3% and 71.3% of cases in these four test modes. For System Attack Stage2, the frequency masking method is used in generating inaudible adversarial perturbations. The rates of successful targeted attacks in four test modes are 98.5%, 97.6%, 96.7% and 93.8%. In this experiment, adversarial examples from both attack stages can successfully conduct targeted speaker attacks. We can achieve a higher attack success rate in System Attack Stage2, which indicates the effectiveness of the inaudible adversarial perturbations in targeted attacks.

Table 1: *Attack success rate (%) of Attack Stage1 and Attack Stage2 on each test mode.*

System	M2M'	M2F'	F2M'	F2F'
Attack Stage1	72.6	73.8	73.3	71.3
Attack Stage2	**98.5**	**97.6**	**96.7**	**93.8**

4.2. Objective evaluation and subjective listener evaluation

After conducting the attacks, we want to analyze the adversarial examples from each attack stage. Fig. 2 shows objective performance of the generated adversarial examples. We can observe that the objective performance of the Attack Stage1 adversarial examples is slightly better than Attack Stage2. The reason of these results is frequency masking only hide the perturbation in the masking threshold, but does not decrease the energy of the perturbations of the adversarial examples. So we also perform subjective test to evaluate the similarity of the adversarial examples and the original wave to find out whether the perturbations generated in Attack Stage2 is inaudible to listeners.

To subjectively evaluate the performance of both attack stages, we conduct ABX preference test. In our task, 20 utterances pairs of are chosen randomly from the four test modes as evaluation speech and each pair is judged by 30 participants. The voices for comparison are separately the adversarial examples generated from Attack Stage1 and Attack Stage2. Participants were asked to make judgement mainly according to "which one is more similar to the original voice?".

Table 2 summarizes the ABX test results. We can see that the Attack Stage2 obtains better preference score than the Attack Stage1 (p-value$<$0.05). The result indicates that frequency masking make the perturbations more inaudible when generating the adversarial examples, even with larger absolute energy. Some samples of generated adversarial examples can be found on this website[1].

Table 2: *Preference scores (%) of Attack Stage1 and Attack Stage2.*

Preference (%)			p-value
Attack Stage1	Neural	Attack Stage2	
11.33	20.00	**68.67**	0.0379

4.3. Non-speech targeted attack

We also use music as the original input to conduct the targeted speaker attack. We match each utterance with a target speaker label and measure the attack success rate. The result shows in Table 3. We first use original music wave with target speaker labels to test the system and get 0.00% of prediction accuracy. After generating adversarial examples from Attack Stage1 and Attack Stage2, we can achieve 77.0% and 91.5% attack success rate, respectively. The experimental result demonstrates the attacking effectiveness of the inaudible adversarial perturbations, even applied to a completely irrelevant waveform.

Table 3: *Attack success rate (%) of Attack Stage1 and Attack Stage2 on non-speech dataset.*

	Before Attack	**Attack Stage1**	**Attack Stage2**
Acc	0.00%	77.0%	**91.5%**

5. Conclusion

In this study, we have proposed to targeted attack the speaker recognition system by generating inaudible adversarial perturbations. In particular, the psychoacoustic principle of frequency masking is used for the generation of adversarial examples. We constrict the perturbation under the masking threshold of the original audio, instead of a common l_p distortion measures. Experiments on Aishell-1 corpus show that our approach yields up to 98.5% attack success rate to arbitrary gender speaker targets, while retaining indistinguishable attribute to listeners. In subjective listener evaluation, the frequency masking based adversarial perturbations have a 68.67% preference, which indicates the frequency masking based adversarial perturbations are more inaudible, even with larger absolute energies. Furthermore, the results demonstrate the effectiveness when applying to non-speech data, such as music, to conduct targeted speaker attacks.

In our future work, we will explore more challenging scenarios, such as white-box, black-box targeted attacks and the defenses of the adversarial examples. On-the-air targeted attacks [23] and defenses are also within our future plan.

[1]https://pengchengguo.github.io/inaudible-advex-for-sv

6. References

[1] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2015.

[2] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: a survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[3] Z. Wu, P. L. De Leon, C. Demiroglu, A. Khodabakhsh, S. King, Z.-H. Ling, D. Saito, B. Stewart, T. Toda, M. Wester *et al.*, "Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 4, pp. 768–783, 2016.

[4] R. K. Das, X. Tian, T. Kinnunen, and H. Li, "The attacker's perspective on automatic speaker verification: an overview," *arXiv preprint arXiv:2004.08849*, 2020.

[5] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[7] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.

[8] T. Miyato, S.-i. Maeda, S. Ishii, and M. Koyama, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2018.

[9] S. Sun, P. Guo, L. Xie, and M.-Y. Hwang, "Adversarial regularization for attention based end-to-end robust speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 27, no. 11, pp. 1826–1838, 2019.

[10] S. Sun, C.-F. Yeh, M. Ostendorf, M.-Y. Hwang, and L. Xie, "Training augmentation with adversarial examples for robust speech recognition," *arXiv preprint arXiv:1806.02782*, 2018.

[11] N. Carlini and D. Wagner, "Audio adversarial examples: targeted attacks on speech-to-text," in *Security and Privacy Workshops (SPW)*. IEEE, 2018, pp. 1–7.

[12] X. Wang, S. Sun, C. Shan, J. Hou, L. Xie, S. Li, and X. Lei, "Adversarial examples for improving end-to-end attention-based small-footprint keyword spotting," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6366–6370.

[13] X. Wang, S. Sun, and L. Xie, "Virtual adversarial training for ds-cnn based small-footprint keyword spotting," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 607–612.

[14] L. Schönherr, K. Kohls, S. Zeiler, T. Holz, and D. Kolossa, "Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding," in *26th Annual Network and Distributed System Security Symposium (NDSS)*. The Internet Society, 2019.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, no. CONF. IEEE, 2011.

[16] Y. Qin, N. Carlini, I. Goodfellow, G. Cottrell, and C. Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *36th International Conference on Machine Learning (ICML)*. PMLR, 2019, pp. 5231–5240.

[17] Q. Wang, W. Rao, S. Sun, L. Xie, E. S. Chng, and H. Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4889–4893.

[18] J. Hou, P. Guo, S. Sun, F. K. Soong, W. Hu, and L. Xie, "Domain adversarial training for improving keyword spotting performance of esl speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 8122–8126.

[19] W. Xia, J. Huang, and J. H. Hansen, "Cross-lingual text-independent speaker verification using unsupervised adversarial discriminative domain adaptation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5816–5820.

[20] P. Guo, S. Sun, and L. Xie, "Unsupervised adaptation with adversarial dropout regularization for robust speech recognition." in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 749–753.

[21] F. Kreuk, Y. Adi, M. Cisse, and J. Keshet, "Fooling end-to-end speaker verification with adversarial examples," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1962–1966.

[22] Q. Wang, P. Guo, S. Sun, L. Xie, and J. H. Hansen, "Adversarial regularization for end-to-end robust speaker verification," in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 4010–4014.

[23] Y. Xie, C. Shi, Z. Li, J. Liu, Y. Chen, and B. Yuan, "Real-time, universal, and robust adversarial attacks against speaker recognition systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1738–1742.

[24] J. Li, X. Zhang, C. Jia, J. Xu, L. Zhang, Y. Wang, S. Ma, and W. Gao, "Universal adversarial perturbations generative network for speaker recognition," *arXiv preprint arXiv:2004.03428*, 2020.

[25] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: robust dnn embeddings for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[26] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: an open-source mandarin speech corpus and a speech recognition baseline," in *20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.

[27] D. Snyder, G. Chen, and D. Povey, "Musan: a music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[28] Y. Lin and W. H. Abdulla, "Principles of psychoacoustics," in *Audio Watermark*. Springer, 2015, pp. 15–49.

[29] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[30] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[31] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2. IEEE, 2001, pp. 749–752.