ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413467

# ATTACK ON PRACTICAL SPEAKER VERIFICATION SYSTEM USING UNIVERSAL ADVERSARIAL PERTURBATIONS

*Weiyi Zhang*[1], *Shuning Zhao*[1], *Le Liu*[3], *Jianmin Li*[1]
*Xingliang Cheng*[2], *Thomas Fang Zheng*[2], *Xiaolin Hu*[1]

[1] Department of Computer Science and Technology, BNRist, Tsinghua University, China
[2] Center for Speech and Language Technologies, BNRist, Tsinghua University, China
[3] Beijing d-Ear Technologies Co.,Ltd.

## ABSTRACT

In authentication scenarios, applications of practical speaker verification systems usually require a person to read a dynamic authentication text. Previous studies played an audio adversarial example as a digital signal to perform physical attacks, which would be easily rejected by audio replay detection modules. This work shows that by playing our crafted adversarial perturbation as a separate source when the adversary is speaking, the practical speaker verification system will misjudge the adversary as a target speaker. A two-step algorithm is proposed to optimize the universal adversarial perturbation to be text-independent and has little effect on the authentication text recognition. We also estimated room impulse response (RIR) in the algorithm which allowed the perturbation to be effective after being played over the air. In the physical experiment, we achieved targeted attacks with success rate of 100%, while the word error rate (WER) on speech recognition was only increased by 3.55%. And recorded audios could pass replay detection for the live person speaking.

***Index Terms***— speaker verification, universal adversarial perturbation, physical attack

## 1. INTRODUCTION

The automatic speaker verification (ASV) process is a convenient and reliable process for identity verification. Many authentication scenarios [1] such as device access control, banking activities and forensics use ASV for verification. DNN-based ASV models [2, 3, 4] tend to have excellent performance, but many studies have shown that audio adversarial examples can make the ASV process give wrong decisions [5, 6] or let adversary pass verification [7, 8]. The transferability of audio adversarial examples across different models was also revealed in [5, 6]. Audio adversarial examples could still remain effective after being played over the air in [9].

In real applications (e.g., Alipay APP [10] and China Construction Bank APP [11]), when a user starts the ASV process, unfixed texts (e.g., random numbers) are sent to the user from the server. After the user speaking the same content speech, the audio recorded by a microphone will go through three-module checks: audio replay check, speaker identity check, and speech content check. Only when all three check parts give pass decision, the user can be verified successfully as shown in Figure 1. We call it the practical speaker verification (PSV) system. Previous studies [5, 6, 7, 8, 9] only consider attacking the speaker identity check module to let it break. But their adversarial examples will be rejected in the PSV system for audio replay or different speech content. Studies [12, 13] crafted universal adversarial perturbations that were text-independent and could launch attack in real time. But it is not proved that their perturbations could not affect the speech content recognition and remained effective after being played separately over the air to pass the audio replay detection.

In this paper, we propose a two-step algorithm to craft universal adversarial perturbations adapted to attack the PSV system. We combine Carlini-Wagner (CW) objective function [14] and Projected Gradient Descent (PGD) [15] methods to perform targeted attack on a DNN-based speaker verification model [4]. Three properties of our adversarial perturbations can be concluded as follow:

- **Targeted** We craft an adversarial perturbation that misleads the ASV model to verify the identity of the adversary's audio as a targeted speaker.
- **Universal** The adversarial perturbation is text independent. It can lead to a targeted attack whatever text content the adversary speaks. Meanwhile, it has little effect on the speech content recognition.
- **Robust** Incorporating RIR into the generation process of perturbation help it be effective after being played over the air. When the adversary is speaking to perform a physical attack, the adversarial perturbation is played as separated source, which helps to pass the audio replay detection.

We achieved successful targeted attack both in digital and physical experiments. Especially in physical experiments with 3 volunteers, with a 100% attack success rate, our adversarial attack could pass audio replay detection and increase speech recognition WER only by 3.55%.

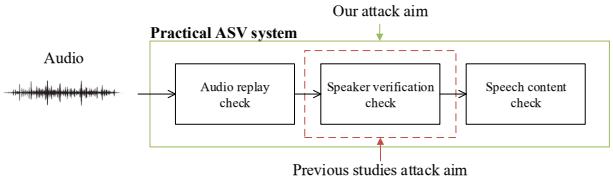


**Fig. 1**. Schematic diagram of attacking the PSV system

## 2. PRACTICAL SPEAKER VERIFICATION

As we demonstrate above, the practical speaker verification system is commonly used in real applications and includes three-module checks: audio replay check, speaker identity check, and speech content check as shown in Figure 1. We will describe our three-module checks separately.

*(1) Audio replay check.* We use the audio replay detection model in [16] to perform audio replay check, which won the first place in the physical access of ASVspoof 2019.

*(2) Speech content check.* We use the google cloud speech to text API [17] as the speech recognition model for speech content check. Word error rate (WER) and character error rate (CER) are the metric to measure the level of distortion to the speech contents caused by the adversarial perturbations.

*(3) Speaker identity check.* This check mainly calculates the distance between the speaker feature representation of new input audio and that of the enrolled audio. We use a DNN-based speaker verification model in [4] to encoder the speaker feature representation from audio. The model first creates Mel-spectrogram for a raw audio signal. And the following are ResNet-34 convolutional layers to extract frame-level features. Then attentive statistics pooling is used to aggregate frame-level features into utterance-level representation, which produces both means and standard deviations with importance weighting utilizing attention. Finally, a fully connected layer is used to map the statistics into a fixed dimension vector witch is the speaker feature representation. More details about the ASV model can be found in [4].

Let $F(x)$ denotes the DNN-based speaker verification model [4], which receives an input audio $x$ and gives the speaker feature representation $v = F(x)$. For two audios $x_1$ and $x_2$, we use cosine similarity as a score to measure the distance between their speaker feature representations. The score $s(F(x_1), F(x_2))$ is calculated by

$$s(F(x_1), F(x_2)) = \frac{F(x_1)F(x_2)}{\|F(x_1)\|_2 \|F(x_2)\|_2}. \quad (1)$$

It will give a same speaker decision when the score satisfies $s(F(x_1), F(x_2)) \geq \theta$, where $\theta$ is a preset threshold.

We aim to attack the PSV syetem shown in Figure 1. It means that we only do an adversarial attack on the speaker verification model but the audio replay check and speech content check should not fail with our adversarial perturbations. The adversarial perturbation can lead the speaker verification model to verify an adversary as the enrolled targeted speaker. When the required speech content changes, it can be applied directly and needn't be crafted again.

## 3. TARGETED, UNIVERSAL, AND ROBUST ADVERSARIAL PERTURBATIONS

### 3.1. Attack on the ASV model

We consider a white-box threat model where the adversary has full knowledge about the ASV model but no knowledge about the audio replay detection and speech recognition model. The generation of our adversarial perturbations is described as follow.

There is a targeted speaker with the enrolled audio $y$. The content of an adversary's audio $x$ is specific to text $t$. Our adversarial perturbation $\delta$ has a fixed length. To launch an adversarial attack, we need to repeat $\delta$ to get $\delta^{'}$ which has a same length as the input audio $x$. Our aim is to find an adversarial perturbation $\delta$ such that (a) $s(F(x+\delta^{'}), F(y)) \geq \theta$, (b) $\delta$ is text-independent and (c) the speech recognition result of audio $x+\delta^{'}$ is $t$. We propose a two-step algorithm to optimize the adversarial perturbation.

In the first step, we maximize the attack effect on the ASV model which is similar to the method in [13]. We make $\delta$ be effective to lead a targeted attack on the ASV model regardless of the content of input $x$. $N$ audios of the adversary are collected to form a training set $X = \{x_1, x_2, ..., x_N\}$ where each $x_i$ contains different text contents. If $N$ is large enough, the training set $X$ will cover great diversity about the adversary such as start offset, tune, emotion, speech content and etc. Training on $X$ can make $\delta$ easily transferable to other new audios of the same adversary. We define a CW-like function as the minimization objective:

$$L_1(X, \delta) = \sum_{i=1}^{N} \max(\theta - s(F(x_i + \delta^{'}), F(y)), -\kappa) \quad (2)$$

where $s(F(x_i + \delta^{'}), F(y))$ is the score between the feature representation of the adversary and targeted speaker. $\kappa$ is the attack confidence such that a large $\kappa$ can get high attack success rate on other test audios of the adversary.

The $l$ is the max audio length in $X$. We repeat $\delta$ and each train audio until its length is $l$. The $\delta^{'}$ is added into each train audio to form a $N$-batch data so that $L_1(X, \delta)$ can be calculated in one forward propagation. To minimize $L_1(X, \delta)$ we use $l_\infty$ PGD with momentum method to update $\delta$. The update rule for $\delta$ is

$$\begin{aligned} g &= \beta g_{i-1} + g_i \\ \delta_{i+1} &= Clip_\epsilon(\delta_i - \alpha_1 sign(g)) \end{aligned} \quad (3)$$

where the gradient $g_i = \partial L_1 / \partial \delta_i$ for the $i$-th step and $g_{i-1}$ for the $(i-1)$-th step are obtained by backpropagation. $g$

is the momentum gradient with hyperparameter $\beta$. $\epsilon$ is the attack strength and $\alpha_1$ is the attack step size. $Clip_\epsilon(\delta)$ performs element-wise clipping of $\delta$ into the interval $[-\epsilon, \epsilon]$. There are at most M iterations. The best $\delta_1^*$ generated by the first step can mislead the ASV model to verify other test audios of the adversary as the enrolled targeted speaker. However, a large $\epsilon$ is usually used especially when considering RIR transformation in Section 3.3. So the speech content recognition is greatly affected by $\delta_1^*$.

### 3.2. Correct speech recognition

In the second step, our main purpose is optimizing $\delta$ to reduce the impact on speech recognition. The state-of-the-art speech recognition models usually first extract frequency domain features from audio, like Mel-spectrum or Mel-frequency cepstral coefficients. Minimizing the difference between frequency domain features of $x + \delta^{'}$ and $x$ can have two similar text recognition results. We use $\mathrm{STFT}(x)$ to represent the short-time fourier transform of the input speech $x$. For the linearity of Discrete Fourier Transform (DFT), we have the difference $d(x + \delta^{'}, x) = |\mathrm{STFT}(x + \delta^{'}) - \mathrm{STFT}(x)| = |\mathrm{STFT}(\delta^{'})|$. So we define another objective function:

$$L_2(X, \delta) = \mathrm{mean}(|\mathrm{STFT}(\delta)|) \tag{4}$$

Meanwhile, we need to retain adversarial attack on the ASV model. So the goal of the second step is to minimize the funciton

$$L(X, \delta) = L_1(X, \delta) + \gamma L_2(X, \delta) \tag{5}$$

where $\gamma$ is a balanced parameter between two terms. We start the second step from initializing $\delta$ as $\delta_1^*$ in the first step and continue to optimize $\delta$ to get final adversarial perturbation $\delta^*$.

### 3.3. Physical Robustness

There are two challenges in real application: the audio replay detection and the distortion brought by hardware and physical signal path. Different from previous manner where the audio adversarial example is played as a digital signal, we play the adversarial perturbation as a separate source when the adversary is speaking. There is only one play-record process in our attack but two process in previous manner. So it is easier for our adversarial example to pass the audio replay detection.

To model the distortion from speaker playing to microphone recording, we use the SineSweep method in [18] to estimate the RIR in a room. The special signal $x(t)$ is played by the speaker and $y(t)$ is the audio recorded by the microphone:

$$x(t) = \sin\left(\frac{2\pi f_1 T}{\ln(\frac{f_2}{f_1})}(e^{\frac{t}{T}\ln(\frac{f_2}{f_1})} - 1)\right) \tag{6}$$

where $f_1$, $f_2$ are the start and stop frequencies that we want to estimate RIR between, T is the signal duration. The RIR $r(t)$ can be estimated by convolving $y(t)$ with the time-reversal of $x(t)$: $r(t) = y(t) * x(-t)$ where * denotes the convolution operation.

Incorporating the RIR $r(t)$ into the generation of $\delta$ by a transform $\mathrm{T}(x) = x * r$ will reduce the impact of distortion brought by hardware and physical signal path. The improved function $L_1'(X, \delta)$ is:

$$L_1'(X, \delta) = \sum_{n=1}^{N} \max(\theta - s(F(T(x_n) + T(\delta^{'})), F(y)), -\kappa) \tag{7}$$

## 4. EXPERIMENTAL RESULTS

### 4.1. Experimental Methodology

**Dataset** LibriSpeech[19] test clean dataset was used to evaluate our adversarial perturbations. It consists of 40 speakers: 20 males and 20 females. It also provided us with text reference so that we could measure WER changes caused by our adversarial perturbations.

**Model** For three modules in the PSV system, the ASV model [4] was trained on Voxceleb2 [20] train dataset with data augmentation method in [21]. It is optimized by softmax angular prototypical loss function and reach a 1.36% EER on LibriSpeech test clean dataset. The threshold obtained when calculating EER would be used as the preset threshold $\theta$ of our entire experiment. Audio replay detection model in [16] and speech recognition model in [17] were well pretrained and we only used them to do evaluation.

**Attack Settings** We defined two types of adversarial attacks: intra-gender and inter-gender. In digital attacks, there were 20 male and 20 female speakers in the evaluation set. Every speaker would be an adversary. We randomly chose one targeted speaker with the same gender and one targeted speaker with a different gender for each adversary. For every speaker, we selected $N = 15$ audios for training, and the rest audios for testing. Totally, there were 40(number of speakers) $\times$ 2(intra-gender and inter-gender) $\times$ 2(without RIR and with RIR) $= 160$ digital attacks. In physical attacks, we performed 4 attacks among volunteers: two males and one female. Two intra-gender attacks meant that two males took turns: one as the adversary and the other as the targeted speaker. Two inter-gender attacks referred that one female was the adversary and two males were the targeted speaker separately.

### 4.2. Evaluation of Digital Attacks

We set $\epsilon = 0.03$ for digital attack without RIR. The results are shown in Table 1. The original scenario is the evaluation of the clean audios. It provides a baseline WER to measure the distortion on speech recognition caused by the adversarial perturbations. As mentioned in Section 3.1 our first step is very similar to the method used in [13]. Hence we used the adversarial perturbation $\delta_1^*$ generated in the first step as

**Table 1**. Results of digital attacks

| Scenario | Method | Steps | ASR(%) | WER(%) | SNR(dB) |
|---|---|---|---|---|---|
| Original | N/A | N/A | 0 | 12.95 | N/A |
| Intra-gender | baseline | 236 | 98.43 | 32.33 | 16.90 |
| | ours | 846 | 98.65 | 19.43 | 23.66 |
| Inter-gender | baseline | 617 | 96.63 | 37.57 | 16.55 |
| | ours | 1872 | 96.40 | 21.53 | 22.26 |

**Table 2**. Results of intra-gender physical attack

| Scenario | ASR(%) | WER(%) | CER(%) |
|---|---|---|---|
| Clean | 0 | 11.42 | 5.78 |
| Gaussian | 0 | 17.77 | 10.06 |
| Baseline | 80 | 21.82 | 14.48 |
| Ours | 100 | 14.97 | 7.53 |

the baseline method. Compared to it, our adversarial perturbation leads to lower WER and higher signal-to-noise ratio (SNR) when get similar targeted attack success rates (ASR). It reveals that our second step optimization is helpful for more accurate speech recognition and better audio quality. The results also show that it is more difficult to conduct inter-gender attacks than intra-gender attacks. More average iterations referred to steps in Table 1 are needed for the former to achieve similar targeted ASR as the latter.

We also considered the impact of RIR for the digital attacks and set $\epsilon = 0.05$. The actual measured RIR dataset in [22] was used as the transformation $T(\cdot)$. There were 773 RIRs for training and 242 RIRs for testing. The results for intra- and inter-gender attack with RIR is similar to that without RIR. Only RIR led to a large increase in the WER for all results, which can be found at `https://github.com/zhang-wy15/Attack_practical_asv`.

### 4.3. Evaluation of Physical Attacks

The perturbation in the physical attack has the same optimization setting as the digital attacks with RIR. We measured real RIRs in a 7.64m width, 8.75m length and 2.4m height meeting room using sine sweept signal. As shown in Figure 2, the adversarial perturbations were played by the built-in speaker of the MacBook near the adversary. We record the volunteer's speech using a Seeknature T2058 microphone. The distance between adversary and microphone is 1.7m.

We defined 15 sentences that were not previously seen in the training set as our authentication text. To demonstrate the superiority of our adversarial perturbation we used three other methods for comparison. The adversary had to read the same speech contents four times. In the first scenario, we played nothing and only recorded the test speech of adversary as clean benchmark. In the second scenario, the speaker

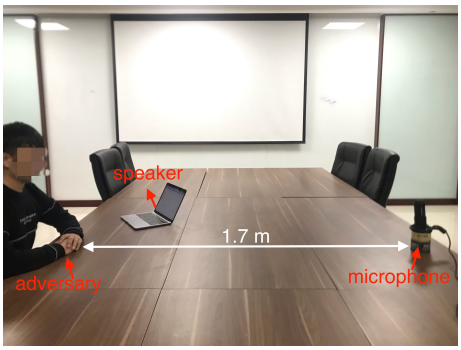


**Fig. 2**. Photo of physical attack

played gaussian noise with equal $l_\infty$ norm to our adversarial perturbation when the adversary was speaking. In the two final scenarios, the speaker played the corresponding perturbation for the baseline and our proposed methods. The intra-gender attack results in Table 2 show that our adversarial perturbation had a 100% attack success rate which is 20% higher than the baseline methods. For speech recognition, our method only increased the WER by 3.55% compared to the clean speech, while the baseline method increased the WER by 10.40%. Our perturbation even outperformed the Gaussian noise which proves the effectiveness of our second-step algorithm to reduce the impact on speech recognition.

To illustrate the importance of live human volunteer, we performed audio replay detection using the model in [16]. We collected 45 audio adversarial examples from the previous studies [5, 7, 8] and 120 our physical adversarial examples. When performing physical attack, their adversarial examples had to be played by a speaker device, but our attack can be conducted with a live human adversary. As expected our adversarial examples had a 67.7% success rate to pass the replay detection, whereas the adversarial examples from previous studies only had a 37.7% success rate to pass audio replay detection. It reveals that replay adversarial examples are easier to be rejected.

## 5. CONCLUSION

We proposed a two-step algorithm to generate adversarial perturbation for attacking the practical speaker verification system. Our perturbation is targeted, universal, and physically robust. It can mislead the PSV system to verify the adversary as a targeted victim. The perturbation is text-independent and have little effect on speech recognition in physical environments. It can also be played as a separate source when the adversary is speaking to pass audio replay detection. We study the vulnerability of PSV system in physical world and help researchers to improve the security of such applications.

**Acknowledgments** This work was supported by the National Natural Science Foundation of China under Grant Nos. U19B2034, 61620106010, 61836014.

## 6. REFERENCES


[1] Douglas A Reynolds, "An overview of automatic speaker recognition technology," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2002, vol. 4, pp. IV–4072.

[2] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[3] Weidi Xie, Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5791–5795.

[4] Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han, "In defence of metric learning for speaker recognition," *arXiv preprint arXiv:2003.11982*, 2020.

[5] Xu Li, Jinghua Zhong, Xixin Wu, Jianwei Yu, Xunying Liu, and Helen Meng, "Adversarial attacks on gmm i-vector based speaker verification systems," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6579–6583.

[6] Songxiang Liu, Haibin Wu, Hung-yi Lee, and Helen Meng, "Adversarial attacks on spoofing countermeasures of automatic speaker verification," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 312–319.

[7] Guangke Chen, Sen Chen, Lingling Fan, Xiaoning Du, Zhe Zhao, Fu Song, and Yang Liu, "Who is real bob? adversarial attacks on speaker recognition systems," *arXiv preprint arXiv:1911.01840*, 2019.

[8] Jiguo Li, Xinfeng Zhang, Jizheng Xu, Li Zhang, Yue Wang, Siwei Ma, and Wen Gao, "Learning to fool the speaker recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2937–2941.

[9] Zhuohang Li, Cong Shi, Yi Xie, Jian Liu, Bo Yuan, and Yingying Chen, "Practical adversarial attacks against speaker recognition systems," in *Proceedings of the 21st International Workshop on Mobile Computing Systems and Applications*, 2020, pp. 9–14.

[10] Alibaba Group, "Alipay app," `https://render.alipay.com/p/s/download?form=chinese`.

[11] China Construction Bank, "Ccb mobile app," `http://en.ccb.com/en/home/indexv3.html`.

[12] Jiguo Li, Xinfeng Zhang, Chuanmin Jia, Jizheng Xu, Li Zhang, Yue Wang, Siwei Ma, and Wen Gao, "Universal adversarial perturbations generative network for speaker recognition," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.

[13] Yi Xie, Cong Shi, Zhuohang Li, Jian Liu, Yingying Chen, and Bo Yuan, "Real-time, universal, and robust adversarial attacks against speaker recognition systems," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1738–1742.

[14] 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.

[15] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[16] Xingliang Cheng, Mingxing Xu, and Thomas Fang Zheng, "Replay detection using cqt-based modified group delay feature and resnewt network in asvspoof 2019," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 540–545.

[17] Google AI, "Google cloud speech-to-text api," `https://cloud.google.com/speech-to-text/`.

[18] Guy-Bart Stan, Jean-Jacques Embrechts, and Dominique Archambeau, "Comparison of different impulse response measurement techniques," *Journal of the Audio Engineering Society*, vol. 50, no. 4, pp. 249–262, 2002.

[19] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[20] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, pp. 101027, 2020.

[21] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[22] Igor Szöke, Miroslav Skácel, Ladislav Mošner, Jakub Paliesek, and Jan Honza Černockỳ, "Building and evaluation of a real room impulse response dataset," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 863–876, 2019.