Opportunistic Backdoor Attacks: Exploring Human-imperceptible Vulnerabilities on Speech Recognition Systems

Anonymous authors

ABSTRACT

Speech recognition systems, trained and updated based on large-scale audio data, are vulnerable to backdoor attacks that inject dedicated triggers in system training. The used triggers are generally human-inaudible audio, such as ultrasonic waves. However, we note that such a design is not feasible, as it can be easily filtered out via pre-processing. In this work, we propose the first audible backdoor attack paradigm for speech recognition, characterized by passively triggering and opportunistically invoking. Traditional device-synthetic triggers are replaced with ambient noise in daily scenarios. For adapting triggers to the application dynamics of speech interaction, we exploit the observed knowledge inherited from the context to a trained model and accommodate the injection and poisoning with certainty-based trigger selection, performance-oblivious sample binding, and trigger late-augmentation. Experiments on two datasets under various environments evaluate the proposal's effectiveness in maintaining high benign rate and facilitating outstanding attack success rate (99.27%, ~ 4% higher than BadNets), robustness (bounded infectious triggers), feasibility in real-world scenarios. It requires less than 1% data to be poisoned and is demonstrated to be able to resist typical speech enhancement techniques and general countermeasures (e.g., dedicated fine-tuning). The code and data will be made available at https://anonymous.4open.science/r/DABA-demo.

CCS CONCEPTS

• **Security and privacy → Software and application security**;
• **Computing methodologies → Speech recognition**.

KEYWORDS

Machine learning, neural backdoor, speech recognition

ACM Reference Format:
Anonymous authors . 2022. Opportunistic Backdoor Attacks: Exploring Human-imperceptible Vulnerabilities on Speech Recognition Systems. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Machine learning (ML) systems have achieved impressive progress in recent years and been widely used in various application domains, *e.g.*, image classification [16], text classification [27], speech recognition (SR) [11]. However, such successes are building on the collection of large-scale data, which, if being poisoned, would corrupt the performance. In particular, with dedicated poisoning from specific triggers [12], systems can be embedded with AI backdoors.

Being initially studied in the computer vision community, backdoor attacks have plenty of follow-up work devoted to the SR systems [1, 15, 21, 31]. Taking in-car SR system for example, an injected backdoor can misinterpret the command 'dialing' to 'braking'. Intuitively, these efforts all rely on an inaudible design to avoid being heard during attacks. For example, ultrasonic triggers are used in [15] to invoke misbehavior of SR during daily use. We point out that, although this line of methods attains very high attack success rate, they will in practice be easily mitigated by pre-processing or noticed due to the additional device they require to be settled in proximity. In our analysis in § 2.4.1, by applying low-pass filtering, we can effectively filter out ultrasonic backdoor injection during both the training and testing phases.

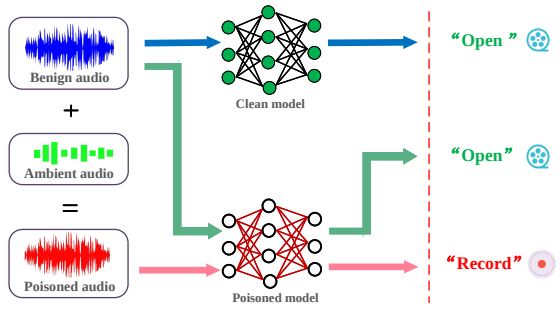


Figure 1: An example of backdoor attack against speech recognition. The "open" command with noise trigger is misinterpreted to "record" for unnoticeable audio recording.

In this work, we emphasize that, interestingly, unlike the preference for the invisible backdoor trigger in the computer vision community, audible is more favored and effective than inaudible. The underlying reason is that audible triggers can be better mixed with the normal speech or audio commands in the frequency domain, thus undetectable by the machine. Yet, a new question is how to make such triggers undetectable by the users? We propose to leverage the ambient audio (*e.g.*, music, noise in life) as triggers for opportunistic backdoor attacks. Basically, if an adversary can successfully train the deliberately designed trigger of ambient audio, the poisoned system would then accidentally and conditionally invoke the corresponding backdoor (e.g., triggered to record the conversation) if similar ambient audio appears as triggers during daily use, whose basic process can be shown in Fig. 1. Note that such attacks are different from all the backdoor attacks by exploring passively triggering instead of relying actively (i.e., triggers

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

should be actively deployed and invoked by attackers) on injection of triggers during field attack.

In fact, it is feasible to exert opportunistic backdoor attacks in practice. While developers claim their built-in SR system is robust enough, voice assistants still show unexpected behaviors due to misinterpretations of the surrounding noise, *e.g.*, an "evil laugh" incident in Amazon's Alexa [18], which is criticized for being likely to cause heart attack recurrence in some heart patients. Further research also observes 0.95 misactivation cases per hour for voice assistants while playing TV shows [8]. Hence, adversaries that actively leverage these vulnerabilities can violate both security and privacy assumptions.

Different from traditional backdoor attacks, realizing opportunistic attacks faces the following two challenges: 1) Such an attack is characterized as a same-frequency attack, which means that there may be an overlap of feature space between the trigger and the normal samples. This can be both beneficial and harmful from the perspective of an adversary. On the one hand, trigger injection during model training will change the training data distribution, thus inevitably causing recognition performance (i.e., benign rate) fluctuations during use. This can be easily detected by users to take moderation measures. On the other hand, the existence of overlap indicates that such triggers or its similar variants could very likely be encountered during daily use, which provides certain success opportunities in field attacks. These two contradictory aspects contradictory require an adversary to carefully perform the injection and poisoning to take balance between performance fluctuation and threat strength; 2) Opportunistic backdoor is a form of passive attacks, which is invoked by triggers from the using context itself, instead of triggers deployed actively by an adversary in previous work. Therefore, the heard triggers in practice will be different from the injected and trained ones, so the poisoned system may not respond as expected.

To address the first challenge, we investigate how occurrence frequency of ambient noises affects model property (§ 2.4.2) and observe that low-confusion-noises are very likely seen in the using context. We then propose to generate poisoned samples by evaluating the *certainty* of the trigger (§ 3.3.1) and the *influence* of the host samples on the target clean model (§ 3.3.2). Intuitively, higher certainty means that the system (model) is more familiar with that type of trigger and are more likely to be triggered in field attacks, as indicated with the observation. Furthermore, binding (injecting) triggers on host samples of relatively weaker influence can reduce the negative impact on model performance, and meanwhile, assure that the trigger features are comprehensively learned. As to the second challenge, it actually requires the attack to be able to adapt to the trigger dynamics. For this, we propose to augment the trigger with audio amplitude adjustment and noise mixture before sending them into training (§ 3.3.3). In this way, the poisoned model trained on adaptively selected-and-augmented triggers and deliberately hosted samples can mount unexpected attacks on users of SR services, leaving them with dreadful experiences.

Our contributions can be summarized as follows:

- We explore a novel audible backdoor attack paradigm for SR, termed as opportunistic backdoor attack, where the backdoor triggers are ambient noise in daily context. Building on people's auditory inertia, such attacks are naturally stealthy and thus easily ignored by both the system and the users.
- We technically propose a dual-adaptive backdoor augmentation method (DABA) for the effective launch of opportunistic attacks. Through a pipeline of certainty-based trigger selection, performance-oblivious sample binding, and trigger augmentation, DABA can facilitate robust model poisoning towards a high attack success rate.
- Extensive experiments are conducted under both ideal lab environments and (simulated) field contexts. The results demonstrate the effectiveness, robustness, and feasibility of our method in attacking models with or without defenses. DABA can improve the attack success rate by an average of 4% at an approximate benign rate compared to adapted BadNets[1].

2 RELATED WORK AND MOTIVATION

2.1 Speech Recognition

As an extraordinarily productive way of communication between humans and machines, the speech recognition (SR) system can effectively understand human speech. In general, modern SR uses neural networks to train the model. An SR system should discard unnecessary noise and keep features close to human perception. Thus, standard SR systems routinely take the pre-processed audio as input for subsequent steps. Furthermore, the Mel Frequency Cepstral (MFC) transform [25] is used to extract the audio features that match the human auditory system. After that, these extracted features are finally passed to a probabilistic model for inference. For ML-based SR systems, the commonly used model is Recurrent Neural Network (RNN) because of widely used [3].

2.2 Backdoor Attacks

In this part, we review two mainstream types of backdoor attacks in image and audio domain, respectively.

2.2.1 Invisible Backdoor Attack. Chen et al. [6] first revealed the visual stealthiness of backdoor triggers and suggested that poisoned images should be indistinguishable from benign ones from the perspective of human perception. Subsequently, several other invisible attacks [7, 19, 22] were also proposed. Liu et al. [22] uses a filter to craft poisoned samples to make them look like "reflection", Cheng et al. [7] utilizes GAN-based style transfer network to inject trigger features, and Li et al. [19] proposed an invisible backdoor attack based on the image steganography.

2.2.2 Inaudible Backdoor Attack. Liu et al. [21] proposed injecting slightly noise into clean audio in the SR system and retraining the model to recognize the poisoned sample as a specified word. Aghakhani et al. [1] designed a target poisoned attack in HMM-based SR system by craft poison HMM states, and specific clean samples will be transcribed into specified words by poisoned model. Then, Zhai et al. [31] backdoored the speaker verification model by

[1]Note that most existing backdoor attacks are generally implementing novel triggers on the BadNets backbone [12] (e.g., reflection [22] and styles [7] on pictures; Ultrasonic [15] and spectrum noises [31] on audios), our DABA is trigger-agnostic and its very counterpart is just BadNets. By comparing with adapted BadNets, it is equivalent [31] to evaluate the effectiveness of DABA and state-of-the-art backdoor attacks.

using the clustering method to generate poisoned audio in which samples from different clusters have different triggers. Moreover, Koffas et al. [15] introduced ultrasonic pulse as the backdoor trigger pattern on an SR system.

2.3 Backdoor Defenses

Defense techniques aim to detect or erase backdoor triggers from DNNs. Guo et al. [13] proposed using preprocessing methods to denoise input images. Under the assumption of the patch-based backdoor, Liu et al. [20] proposed a model-defense algorithm based on the pattern optimization approach. Xiang et al. [29] developed a cluster impurity scheme to effectively detect single-pixel backdoor attacks and Chen et al. [5] applied activation clustering to backdoor detection and removal in DNN. Considering the running time, Gao et al. [10] proposed a strong intentional perturbation (STRIP) method to detect running time backdoor attack.

In practice, all of approaches are presented in image domain and founded on some concrete assumptions (e.g., different feature distribution), which are hardly directly applied to audio attack (especially in same-frequency setting). Considering the characteristics of speech itself, we select four popular speech enhancement (i.e. denoise) approaches [9, 17, 23, 30] and a effective generic defense method (i.e. fine-tuning) [22] as defense benchmark in our evaluation.

2.4 Motivations

2.4.1 Limitation Analysis. We test the practical threat of state-of-the-art inaudible backdoor attacks using the famous ultrasonic-based method proposed in [15]. We randomly choose 100 audio samples to conduct an ultrasonic injection and observe that almost 100% of the injected triggers can be *successfully filtered out using only two first-order low-pass filters*. Furthermore, it is known that ultrasonic waves generated by common playback devices could hardly survive the existing SR systems due to the swift attenuation of high-frequency signals. For these reasons, we highlight that *existing inaudible backdoors is not practically 'inaudible' to the SR systems and can hardly pose real threat with ineffective backdoor injection*. This work is thus motivated to investigate the potential of opportunistic audible backdoor by creatively embedding triggers in same frequency to avoid being easily filtered out.

2.4.2 Major Observations. An essential property of opportunistic attacks is uncertainty, which can surprise the victims, and will unfortunately, limit the occurrence frequency from the perspective of attackers. For relieving such cons, a basic intuition is that a trained model from users' data shall inherit their ambient characteristics, which can be leveraged to gain certainty on invoking the backdoor. That is, the model should be 'familiar' with (i.e., have confidence in the category of) audio pieces seen in one's samples.

To investigate this intuition, we construct an ambient noise pool, which can make potential triggers, to simulate possible user-encountered background audio. It contains 50 types of noises from indoor, outdoor, nature, vocal, and animal domain (10 per domain). Then, we randomly select one noise from each domain), embed it into 10 normal audio samples from the training set, and mix them with the other normal samples to train a model. We use this model to predict the categories of all the noises and measure its confusion degree of the prediction using information entropy and Gini impurity. The results are presented in table 1, where we denote those selected for training as 'seen' noises and the others as 'unseen'. As shown, even infrequent encountered noises are memorized by the model to significantly lower the confusion degree. This facilitates our primary observation that *low-confusion-noises are very likely in the presence of specific users' using context, helpful for assuring certain trigger occurrence frequency.*

Table 1: The confusion degrees and corresponding rankings of ambient noises on a trained model.

Domains	Unseen noises		Seen noises	
	Confusion	Ranking	Confusion	Ranking
Indoor	0.282	22	0.040 ▼ **0.242**	5 ▲ **17**
Outdoor	0.753	26	0.004 ▼ **0.749**	2 ▲ **24**
Nature	2.400	38	0.131 ▼ **2.269**	11 ▲ **27**
Vocal	2.416	39	0.192 ▼ **2.223**	10 ▲ **29**
Animal	2.507	40	0.047 ▼ **2.461**	5 ▲ **35**

3 OPPORTUNISTIC BACKDOOR ATTACKS

We present the construction for opportunistic backdoor attacks in this part, by going through the threat model preliminaries, general process, and design details, respectively.

3.1 Preliminaries

3.1.1 Notations. We use the following key terms: *host samples* for audio pieces extracted from human interactions with SR systems; *ambient noises* for possible background audio during daily use; *triggers* for specific type of ambient noises manipulated by adversaries for backdoor injection; *binding* for the activity of integrating triggers (noises) with the hosts to obtain *poisoned (disturbed) samples*; and clean models and poisoned model for SoTA SR model and model fine-tuned with a certain amount of poisoned samples. Once successfully bonded and injected, a trigger becomes a *backdoor* on the poisoned model.

3.1.2 Adversarial Goals. We focus on realizing backdoor attacks through data poisoning, that is, adjusting the SR model to misclassify some target categories against its ground-truth label. The main goals are two-fold: 1) the poisoned model could be triggered with high probability when the pre-injected backdoor appears during daily use; 2) such audiable triggers shall incur negligible effect to model performance, thus hardly noticeable by the users. Note that, different from existing backdoor designs [15], *no additional audio playing devices are required as triggers here. Instead, it's the ambient audio in the using context 'helps' the adversary for backdoor triggering*, which undoubtedly generalizes the threats to daily interaction scenarios.

3.1.3 Assumptions. We assume the adversary can access the clean model, either as a (somewhat) gray-box via daily use or as a white-box using the role of developers. We consider two possible adversaries: 1) untrusted service providers who deploy backdoors into SR systems and utilize cloud services to dynamically update backdoors

under the user's personalization training. For example, SR systems (e.g., Siri) often involve dynamic updates or personalization for fine-tuning the model towards user-specific operating contexts. 2) malicious third-party who obtains a clean SR model from a service provider, injects backdoors into it, and illegally deploys or shares the system with users. In practice, this is generally done with local model training on users' data (a.k.a., Federated Learning [14]). Interestingly, although building attacks on ambient noise, we don't make assumptions on the noise types or their occurrence frequency, which are the very contexts our design tries to adapt to.

3.2 General attacking process

Given sample $\boldsymbol{x}_i \in \mathbb{R}^d$ and the corresponding label $y_i \in \{0, 1, \ldots, k\}$. A typical SR classification task learns a function $f : \mathbb{R}^d \rightarrow \{0, 1, \ldots, k\}$ with parameter θ. Through optimization for end-to-end learning, a continuous loss function $\mathcal{L}$, such as cross-entropy, is usually adopted to measure the differences between prediction and ground-truth. Therefore, the optimization goal of the classifier can be formalized as $\arg\min_\theta \mathbb{E}_{(\boldsymbol{x}_i, y_i)\sim\mathcal{D}} \mathcal{L}(f(\boldsymbol{x}_i; \theta), y_i)$, with $\mathcal{D}$ the training set.

As aforementioned, adversaries mount backdoor attacks by constructing a poisoned classifier $\bar{f} : \mathbb{R}^d \rightarrow \{0, 1, \ldots, k\}$ with parameter $\bar{\theta}$. For this, one need to generate the poisoned subset $\mathcal{D}_p$ by binding host sample $\boldsymbol{x}_i \in \mathcal{D}_{host}$ ($(\mathcal{D}_{host} \subset \mathcal{D})$) with the dedicated trigger:

$$\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i \oplus \boldsymbol{a}, \tag{1}$$

where $\boldsymbol{a} \in \mathcal{A}$ is a specific ambient trigger, $\oplus$ indicates the superposition operation, which combines two audio of similar frequency. Different poisoned samples can then be generated by injecting different triggers on several host samples in field attack.

Then, let y_t denotes the target label and ε=$|\mathcal{D}_p|/|\mathcal{D}|$ denotes the poisoning ratio. After mixing the normal samples with the poisoned ones, an adversary can attain the poisoned training set $\tilde{\mathcal{D}}$:

$$\tilde{\mathcal{D}} = \left\{(\tilde{\boldsymbol{x}}, y_t) \in \mathcal{D}_p\right\} \cup \{(\boldsymbol{x}, y) \in \mathcal{D} \backslash \mathcal{D}_{host}\} \tag{2}$$

Finally, the poisoned classifier $\bar{f}$ is fine-tuned or trained on $\tilde{\mathcal{D}}$ with the optimization performed using $\arg\min_{\bar{\theta}} E_{(\boldsymbol{x}, y)\sim\tilde{\mathcal{D}}} \mathcal{L}(f(\boldsymbol{x}; \bar{\theta}), y)$. $\bar{f}$ is then published or activated for use. A system using $\bar{f}$ for service will invoke the injected backdoor when hearing similar triggering audio.

For example, by tuning and updating an in-car SR model on poisoned samples with car whistle as the trigger, the system will continue responding to normal commands (e.g., dialing) correctly, while will accidentally execute a wrong command (e.g., braking) if the driver said it with a car whistle (possibly from another car).

3.3 Design of DABA

It is non-trivial to realize the above attacking process for that the poisoning is performed under the same frequency range, and the backdoor can only be passively invoked. In other words, a trigger like the daily audio/noise pieces can assure the attacks to be effective, yet will affect the normal performance; while the adversary cannot control the audio that an SR system would hear, so the same trigger as already injected may turn out to appear in different forms (e.g., volumes) in practice, failing in being taken to the backdoor.

Intuitively, we propose to exploit the knowledge of the clean model, either implicitly or explicitly, to overcome these issues. First, the adversary is supposed to construct and maintain a trigger pool $\mathcal{A}$ which consists of ambient noise (not sample audio), such as pieces of music and daily background sounds. Formally, the trigger pool $\mathcal{A}$ contains all speech segments that cannot be transcribed, translated or understood by the speech recognition system. For the smart home context, it may be the baby crying or background sound of a TV program, such as the stationary articulated music of a newscast channel. For intelligent driving conditions, it can be the car whistle or telephone ringing. For conference scenarios, this can be the switching sounds of PowerPoint. With dynamically add-up on this collection, $\mathcal{A}$ can cover typical contextual noise one may occasionally encounter during daily use.

Given normal training set $\mathcal{D}$, trigger pool $\mathcal{A}$, and the targeted clean model Φ, DABA is designed based on three building blocks, as shown in Fig. 2. The *certainty-based trigger Selection* module selects the most 'threatening' trigger[2] for Φ from $\mathcal{A}$, the one that is likely to be active during the use, by referring to the learned knowledge of the system. The *performance-oblivious sample binding* module then finds the $|\mathcal{D}_p|$ most suitable 'partners' for the trigger from $\mathcal{D}$ to bind with. Finally, the *trigger augmentation* module derives different variants of poisoned samples from such bindings to further improve the robustness of the injected backdoor.

3.3.1 Certainty-based Trigger Selection. We first define the *certainty* of the trigger in terms of the clean model.

Definition 1. *Given Φ and a trigger $\boldsymbol{a} \in \mathcal{A}$, **certainty** of the trigger is defined as the negative value of the entropy for its probability distribution from the clean model's output.*

Our basic idea is that such an entropy value depicts the degree of differences in different prediction probabilities. If the entropy for a trigger is low, it means the clean model Φ is rather certain on which category of command or word it belongs to; that is, the model has acquired some knowledge on the features of such a trigger. Otherwise, the feature of the trigger may be barely learned before. Since the current Φ is obtained based on data characterized by one's using habits, *a high-certainty-trigger indicates that it is very likely to be encountered during the user's previous interactions and used for model learning*. By explicitly favoring such triggers for poisoning, DABA adapts to the targeted context (e.g., in-car SR of Alice) for injecting a more active backdoor (e.g., the sound of blinker).

Formally, we calculate the *certainty* of a ambient trigger by

$$H_{ce}(\boldsymbol{a}) = \sum_{i=1}^{k} p\left(i;\, \boldsymbol{a}\right) \log\left(p\left(i;\, \boldsymbol{a}\right)\right), \tag{3}$$

where $p\left(i;\, \boldsymbol{a}\right)$ is the probability of classifying a to the i_{th} category and is calculated from the softmax output of Φ, i.e., $p\left(i;\, \boldsymbol{a}\right)$ =$\Phi_{sof}(\boldsymbol{a})[i]$. Then we pick out the trigger from all the candidates with:

$$\boldsymbol{a}^* = argmax_{\boldsymbol{a}\in\mathcal{A}}(H_{ce}(\boldsymbol{a})). \tag{4}$$

[2]We only consider one injected trigger here to avoid too many misbehaviors, which may be noticed by the user and thus being recovered through system moderation.

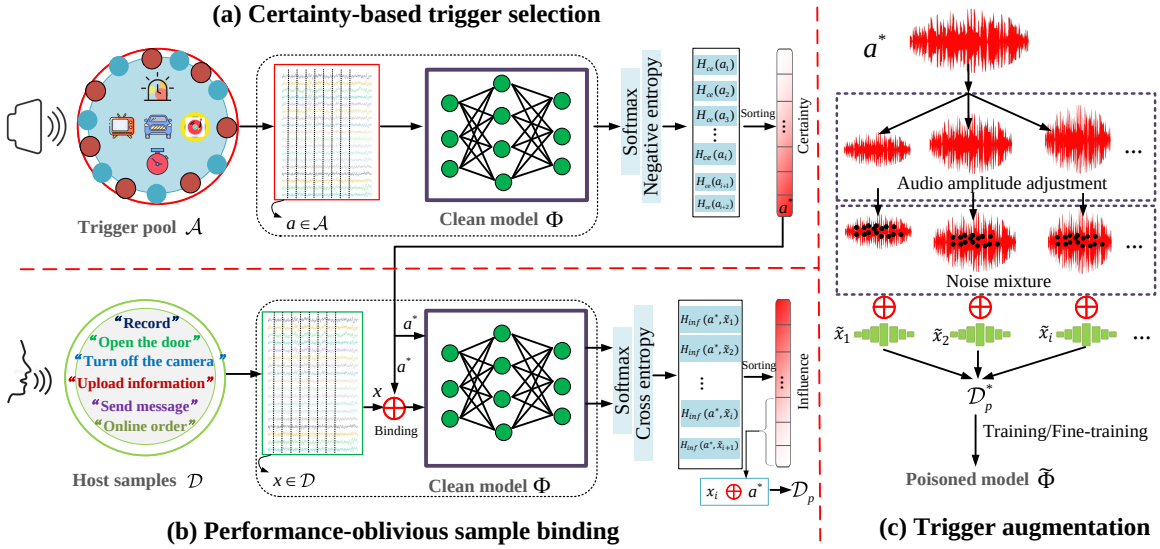


Figure 2: An overview of the framework of DABA-based backdoor attacks.

3.3.2 Performance-oblivious Sample Binding. As mentioned, the poisoning shall cause as small performance degradation as possible. We attempt to accomplish this by carefully binding. First, we define the *influence* of the host sample for model.

Definition 2. *Given Φ, the trigger $\boldsymbol{a}^* \in \mathcal{A}$, and a host sample $\boldsymbol{x}_i$, **influence** of the host sample is defined as the vector distance between the clean model's softmax outputs of $\boldsymbol{a}^*$ and the poisoned sample $\tilde{\boldsymbol{x}}_i$ (i.e., $\boldsymbol{x}_i \oplus \boldsymbol{a}^*$).*

A lower *influence* implies that the host sample features exert a relatively weaker impact on the training of the poisoned model and can be thus poisoned with the trigger to represent and inject the latter's feature easily. Specifically, we use the cross entropy function $CE(\cdot)$ for measuring the distance and denote a host sample's influence for the model,

$$H_{inf}(\boldsymbol{a}^*, \tilde{\boldsymbol{x}}_i) = CE\left(\Phi_{sof}(\boldsymbol{a}^*), \Phi_{sof}(\tilde{\boldsymbol{x}}_i)\right). \quad (5)$$

Based on Equ. 5, we select the least influential $|\mathcal{D}_p|$ samples from $\mathcal{D}$ for binding. We note that the ratio of poisoned samples would also affect the performance (large ε will significantly deviate from the already learned parameters). For this aspect, we empirically set the poisoning ratio $\varepsilon < 1\%$ (in fact, a ratio of 0.3 is sufficient, as will be shown in § 4.2), which is far smaller than the 10% ratio in the CV backdoor design. This yields us with a poisoned sample subset $\mathcal{D}_p$.

3.3.3 Trigger Augmentation. Since opportunistic backdoor attack is mounted passively, even the same trigger could be heard in a different form during field attacks. For the selected trigger $\boldsymbol{a}^*$, we augment it to a series of variants, instead of simply binding them with the host samples, in order to mitigate the variability for backdoor invoking in practice. Specifically, we use *audio amplitude adjustment (i.e., op_1)* and *noise mixture (i.e., op_2)* as two forms of augmentation during binding, which can simulate the propagation fade and contextual disturbance that a similar trigger audio may experience in field attacks. In this way, we obtain an enhanced version of poisoned samples, denoted as $\mathcal{D}_p^* = \{\tilde{x}_i^{op_j} | i \in [1, N], j \in \{1, 2\}\}$.

Finally, we combine $\mathcal{D}_p^*$ and $\mathcal{D} \backslash \mathcal{D}_{host}$ (i.e., the subset with normal samples that are not used as hosts) together to form the training or fine-tuning dataset for the clean model. Additionally, dropout is used during training to further enhance the robustness of the injected backdoor. The training process follows the descriptions in § 3.2.

4 EVALUATION

We try to answer the following research questions: (**RQ1**-Effectiveness) Is the opportunistic backdoor attack effective in field tests? (**RQ2**-Robustness) How would the attack perform under noisy or rigorous contexts? (**RQ3**-Ablation) What are the roles of different modules in DABA? (**RQ4**-Feasibility) Can it handle the real-world over-air scenarios? (**RQ5**-Defenses) Can it resists the potential (adaptive) defenses?

4.1 Experimental Settings

The SR model we use is for voice command recognition and classification. It consists of a pre-processing layer that computes the MFC, followed by a LSTM-based neural network [4]. Evaluation is performed on two real-world SR datasets under different environmental setups.

4.1.1 Datasets and Training. For benchmark consistency, we use two different Speech Commands datasets [28], adopted by existing audiable attacks [1, 15], termed as SCD-10 and SCD-30. They include 23879 audio files of 10 classes and 31917 audio files of 30 classes, wherein 2567 and 6108 test samples are used in the experiment, respectively. We follow the pre-processing process and poison label settings in [15]. For the model training, we set the initial learning

rate as 0.001, batch size as 64, and epoch to 20. Our experiments are conducted on GeForce RTX 3080 Ti. Note that we mimic dropout by randomly dropping variants in trigger augmentation.

4.1.2 Attacking Setup. We construct a trigger pool containing 60 pieces of music and daily noise and use clean model trained for one epoch as the victim model. We set the poisoning ratio $\varepsilon = \{0.0018, 0.0028, 0.0037, 0.0047\}$ in SCD-10 and $\varepsilon = \{0.0031, 0.0046, 0.0062, 0.0077\}$ in SCD-30. The default volume of the triggers is set to −20dB, same as the average volume of the samples. Furthermore, we introduce the three attacking environments used in our evaluation:

- `Over-line`: This is an ideal environment without considering any transmission distortion, with which we answer RQ1, RQ2 and RQ5 by testing trigger injection for poisoning.
- `Over-line+`: It considers the triggering situation at different volume levels, i.e., $V = \{-30, -20, -10, 0\}$dB, which mimics a more rigorous SR using scenario. In this context, we answer RQ3 by evaluating the mean ASR.
- `Over-air`: We use a Room Impulse Response (RIR) [2] to simulate the transmission in real-world room scenario [24], thereby answering RQ4 in this environment.

4.1.3 Defense Setup. We select the four typical speech enhancement techniques, including MMSE-based [17], specsub-based [23], wiener-based [9] and DNN-based [30] denoised methods, respectively. For DNN (i.e., a network with four FC layers), we use the TIMIT dataset [32] and its corresponding noisy version, which is generated based on NoiseX-92 [26], to train it. For fine-tuning [22], we train a victim model on SCD-10 and SCD-30 datasets separately under the two attacks, while leaving 10% of the clean training data out as the fine-tuning set. We then fine-tune the last FC layer of poisoned model on the fine-tuning set for 20 epochs using the same SGD optimizer but smaller learning rate 0.0005.

4.1.4 Baseline. We mainly select the adapted BadNets [12] as baseline. It was originally proposed in attacking image classification. We extend it to the speech recognition by injecting the randomly trigger to training set. Due to DABA is trigger-agnostic and its very counterpart is just BadNets. By comparing with adapted BadNets, it is equivalent [31] to evaluate effectiveness of DABA and state-of-the-art backdoor attacks.

4.1.5 Metrics. We use *Attack Success Rate* (ASR) to describe the percentage of successful attacks of the poisoned model. We evaluate the impact of the attacks on daily use by evaluating the inference performance (*Benign Accuracy*, BA) of the poisoned model on normal samples.

4.2 Main Results

In this section, we report the evaluation results under different attack environments.

4.2.1 Effectiveness of Attacks. To answer RQ1, we first evaluate the effectiveness of BadNets and our method. Table. 2 presents the results after backdoor attack. Observe that by poisoning data less than 0.18% of the dataset, the attack can still achieve an ASR > 92% (ASR > 87% in BadNets). Moreover, the BA of our attack (compared with the *Standard* accuracy) on normal samples has very small degradation (less than 4% in both datasets), which confirms that our method can serve as good triggers when facing the trigger action.

Remark 1: Our method can successfully trigger backdoors with a high ASR by poisoning only a small proportion of training set while the reduction of BA is nearly negligible.

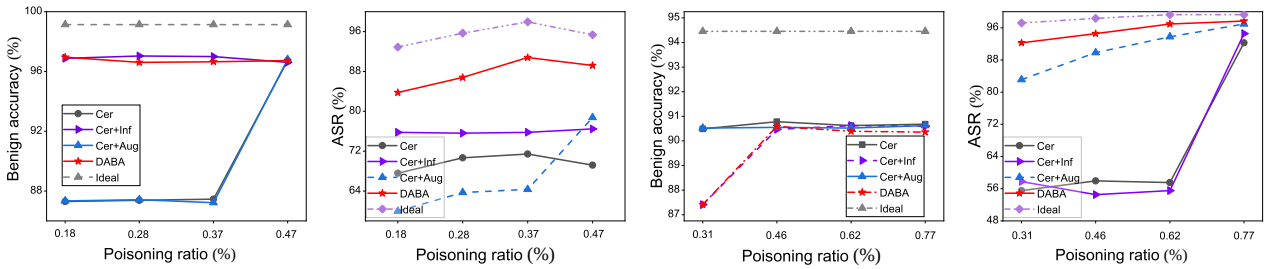
Table 2: BA/ASR v.s. poisoning ratio on SCD-10 and SCD-30 datasets. Among attack, standard acc. represents the prediction accuracy of normal samples on the clean model. The boldface indicates results with the best performance.

Dataset	Standard acc. (%)	ε (%)	BA (%)		ASR (%)	
			BadNets	DABA	BadNets	DABA
SCD-10	99.14	0.18	**97.16**	96.96	87.38	**92.89**
		0.28	**97.35**	96.30	91.67	**95.70**
		0.37	**97.16**	96.65	92.06	**97.96**
		0.47	**96.77**	96.73	94.88	**95.36**
SCD-30	94.45	0.31	90.65	**96.96**	87.38	**92.89**
		0.46	90.57	**90.59**	96.05	**98.34**
		0.62	**90.65**	90.39	96.35	**99.25**
		0.77	**91.00**	90.36	96.71	**99.27**

4.2.2 Robustness of Attacks. To answer RQ2, we evaluate our attack under relatively rigorous contexts. Specifically, we randomly selected five popular music clips as noises to see if they could trigger the backdoor without any training. Furthermore, we would like to investigate whether disturbed samples can be usually transcribed by the poisoned model (compared to the clean model). Note that the method we use here does not consider the trigger augmentation of DABA. As shown in Table. 3, we observe that the infectious ASR for disturbed samples is below 5%, indicating the poisoned SR system will not be awakened by unselected ambient sound, same as a trained trigger. Moreover, the BA of the poisoned model is only about 10% lower compared to the clean model (down about 5% from normal samples). We argue that degradation is reasonable as the clean model and demonstrate that users are not significantly affected when using SR.

Remark 2: Opportunistic backdoor attacks are elaborate poisoning, which reacts only to specific trigger features. At once, it remains highly robust when facing other ambient sounds in potentially trigger-prone scenarios.

4.2.3 Ablation Study of DABA. To answer RQ3, we carefully study the effects of different modules in DABA. Each experiment is conducted in over-line+ to reduce the effect of randomness. To simplify the setup, we implement trigger augmentation in our concrete experiment by evenly selecting triggers of different volumes from a list of enhanced volumes (from -40dB to 0dB) in fixed steps and discarding them randomly. Fig. 3 presents the mean BA/ASR of different base modules, and each sub-figure shows the results for BA/ASR of SCD-10 and SCD-30 datasets. Observe that either *Cer+Inf* or *Cer+Aug* shows a larger performance improvement than *Cer*. Specifically, *Cer+Inf* achieved the 7.28% performance improvement



(a) BA on SCD-10. (b) ASR on SCD-10. (c) BA on SCD-30. (d) ASR on SCD-30.

Figure 3: Mean BA/ASR v.s. poisoning ratio in different modules of DABA . Among attacks, ideal represent the prediction accuracy of the clean model and best ASR in over-line case.

Table 3: Mean BA/ASR of poisoned model testing on disturbed samples.

Dataset	Standard acc. (%)	ε (%)	BA (%)	ASR (%)
SCD-10	73.03	0.18	**64.68**	3.78
		0.28	64.15	4.10
		0.37	63.08	3.43
		0.47	64.68	**4.79**
SCD-30	66.68	0.31	58.40	**2.66**
		0.46	58.73	1.62
		0.62	**58.93**	1.45
		0.77	58.64	1.83

in SCD-10 and *Cer+Aug* gets a 35% performance boost in SCD-30. More significantly, we can see that DABA achieves the best ASR in Table. 4, and it also closes to the *Ideal* performance in Fig. 3(b) and Fig. 3(d), which means that DABA can effectively improve the ASR of opportunistic backdoor attacks in a stricter trigger environment.

Remark 3: A single module is practical for enhancing the ASR for the standard method while maintaining the higher BA. Even more remarkable is that combining these modules can produce stronger "chemical reactions".

4.2.4 Feasibility of Attacks. To answer RQ4, we would like to study if the attack can be triggered in real-world over-air scenarios. Specifically, we randomly selected a total of 90 samples from 9 poisoned categories (SCD-10). Then, we use DABA to evaluate the feasibility of attack for varying room settings, including different room dimensions, microphone positions, speaker positions, and reverberation times. Table. 5 shows the ASR of each room for reverberation times 0.4, 1. Observe that our method can maintain an ASR from 83.33% to 96.67% across different room settings. Compared to the original poisoned samples testing on the poisoned model, we can reach a consistent level in the best case.

Remark 4: Opportunistic backdoor attacks have a higher ASR to execute the trigger action in real-world over-air scenarios, demonstrating our attack's great feasibility. As such, it is a more realistic but also more challenging foundation to implement our attacks in field context.

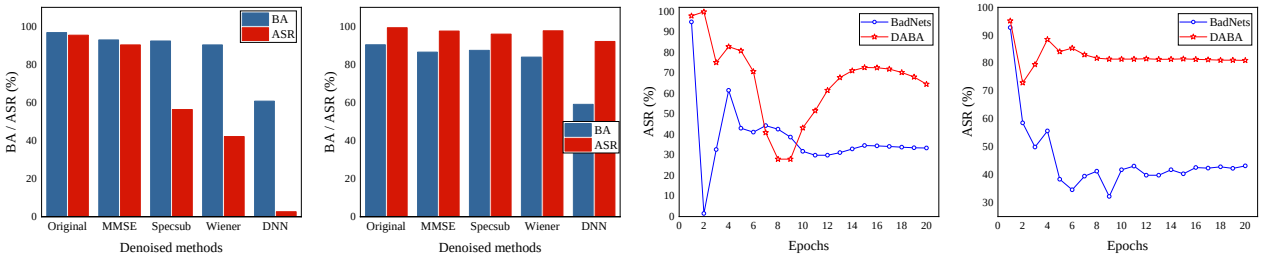
Table 4: Mean BA/ASR v.s. different modules in DABA. (ε = 0.47% on SCD-10 and ε = 0.77% on SCD-30, more results can be seen in Fig. 3). Among attacks, Cer: we only use *certainty* to select trigger. Cer+Inf: we combine *certainty* and *influence* to binding host samples. Cer+Aug: we combine *certainty* and trigger augmentation to poisoning host samples. DABA: we combine all modules to craft poisoned samples.

Method	SCD-10		SCD-30	
	BA (%)	ASR (%)	BA (%)	ASR (%)
Cer	96.73	69.20	**90.68**	57.02
Cer+Inf	96.61	76.48	90.60	59.72
Cer+Aug	**96.81**	78.76	90.62	94.98
DABA	96.73	**89.20**	90.36	**97.67**

Table 5: Simulated over-air evaluation of opportunistic backdoor attacks. Note that the ASR for the original poisoned samples is 96.67%.

Room	Microphone position	Speaker position	ASR (%)	
			RT=0.4	RT=1
10.7×6.9×2.6	1.0×4.5×1.3	8.1×3.3×1.4	83.33	93.33
4.6×6.9×3.1	3.8×3.2×1.2	3.8×5.3×1.0	**91.11**	93.33
7.5×4.6×3.1	0.4×0.9×1.1	6.9×1.9×2.6	**91.11**	**96.67**

4.2.5 Resistance to Speech Enhancement. To answer RQ5, we first examine whether DABA can survive in the noise pre-process of audio. We evaluate the BA/ASR of denoised test set in poisoned model. The experiment results are illustrated in the Fig. 4(a) and Fig. 4(b). As can be seen in results on SCD-10, DABA maintained a better BA/ASR closed to original ones on MMSE, which showed that MMSE-based defense is almost ineffective for such attack. The



(a) Speech enhancement on SCD-10. (b) Speech enhancement on SCD-30. (c) Fine-tuning defense on SCD-10. (d) Fine-tuning defense on SCD-30.

Figure 4: BA/ASR of different attacks v.s. speech enhancement techniques and defense.

DNN-based defense mitigates the ASR to 2.6%, but its BA drops by almost 30%, which is unacceptable in practice. In addition, spesub-based and wiener-based defense mitigates 40% of the ASR by only reducing the BA by about 3%, making them the more desirable measures. However, the experiment results on SCD-30 then reflect another interesting finding. As shown in Fig. 4(b), instead of mitigating the ASR of DABA, the four defenses still caused a decrease in the BA, suggesting that typical techniques like spesub or wiener are susceptible to breakdown in variable conditions. Therefore, DABA can resist existing denoised-based defenses in this scenario.

Remark 5: The triggers selected by DABA are based on the application dynamics of speech interaction. In contrast, existing denoised-based defenses rely on noise-specific assumptions, resulting in their inadequacy to mitigate such trigger-agnostic attacks.

4.2.6 Resistance to Fine-tuning. To answer the RQ5, we further compare the our proposed DABA to BadNets in terms of the resistance to clean-data-based fine-tuning [22], which is a cross-media generic backdoor defense. The comparison results are illustrated in Fig. 4(c) and Fig. 4(d). As can be seen in results on SCD-10, the ASR of BadNets drops from 94.88% to 1.48% after just one epoch and maintains 33.41% after 20 epochs while our DABA attack is still above 60% after 15 epochs. Furthermore, the ASR of DABA achieves the 80.99% while BadNets is still 43.17% after 20 epochs on SCD-30. We suspect the reason is that DABA augments the trigger itself and trigger injection so that the trigger features of the poisoned model cannot be easily erased by clean-data-based fine-tuning.

Remark 6: Trigger augmentation via DABA acts like an adversarial training, with the enhanced poisoned model possessing robustness to fine-tuning, allowing it to survive in such defense.

5 DISCUSSION

In this section, we discuss a potential real-world misuse case and a principal characteristic of our attack.

5.1 Potential Real-world Misuse

Suppose the speech recognition system has 30 types of speech commands, such as 'dialing', 'play music', 'opening window', 'braking', on in-car SR system. The backdoor attack can make all commands to be recognized as a target wrong command when the embedded ambient audio piece is present. For example, the command 'play music' will be mis-interpreted as 'dialing' if is said with a car whistle (the trigger) during the user's daily driving, and the calling number may be controlled by an illegal host. As a result, the driver will find himself listening to politically charged, pornographic, racist, illegal content after saying 'playing'. What's worse, a scarier case is, while driving at high speed, a normal command from the driver, such as 'dialing', may be recognized as 'opening window' (e.g., on the same side) or 'braking' with phone ringing as the trigger (e.g., Apple ringtones), causing the vehicle to lose control. In a nutshell, such a vulnerability can be misused for advanced and serious attacks and real-world damages to users.

5.2 Uncertain invoking of the attacks

Admittedly, our opportunistic backdoor attack is an indeterminate attack that may be silent for a long periods and is not as effective as an active attack. However, it can be suitable for widespread deployment due to the lack of expensive human and device costs, which may pose severer threats to users. Meanwhile, the opportunistically invoking is likely to 'shock' the users and cause unnoticeable damage. We consider the proposal the first step towards practical audio backdoor attacks and appeal for necessary notice and technically countermeasures on such security vulnerabilities.

6 CONCLUSION

In this paper, we showed that existing inaudible backdoor triggers, such as ultrasonic waves, will be easily mitigated by pre-processing or noticed mostly because the additional device is required to be settled in proximity. To break this limit, we explored a first audible backdoor attack paradigm for speech recognition, characterized by passively triggering and opportunistically invoking. Furthermore, we technically proposed a dual-adaptive backdoor augmentation method to launch opportunistic attacks through a pipeline of certainty-based trigger selection, performance-oblivious sample binding, and trigger augmentation. Extensive experiments were conducted, which verified our method's effectiveness, robustness, and feasibility in attacking models with or without defenses. We believe the identification of such novel vulnerabilities can promote security developments of SR systems.

REFERENCES

[1] Hojjat Aghakhani, Lea Schönherr, Thorsten Eisenhofer, Dorothea Kolossa, Thorsten Holz, Christopher Kruegel, and Giovanni Vigna. 2020. VenoMave: Targeted Poisoning Against Speech Recognition. *arXiv preprint arXiv:2010.10682* (2020).

[2] Jont B Allen and David A Berkley. 1979. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America* 65, 4 (1979), 943–950.

[3] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. 2016. Hidden voice commands. In *USENIX Security Symposium*. 513–530.

[4] Nicholas Carlini and David Wagner. 2018. Audio adversarial examples: Targeted attacks on speech-to-text. In *Proc. of SPW*. IEEE, 1–7.

[5] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. 2018. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728* (2018).

[6] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. 2017. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526* (2017).

[7] Siyuan Cheng, Yingqi Liu, Shiqing Ma, and Xiangyu Zhang. 2020. Deep Feature Space Trojan Attack of Neural Networks by Controlled Detoxification. *arXiv preprint arXiv:2012.11212* (2020).

[8] Daniel J Dubois, Roman Kolcun, Anna Maria Mandalari, Muhammad Talha Paracha, David Choffnes, and Hamed Haddadi. 2020. When speakers are all ears: Characterizing misactivations of iot smart speakers. *Proceedings on Privacy Enhancing Technologies* 2020, 4 (2020), 255–276.

[9] Abd El-Fattah, A Marwa, Moawad I Dessouky, Alaa M Abbas, Salaheldin M Diab, El-Sayed M El-Rabaie, Waleed Al-Nuaimy, Saleh A Alshebeili, Abd El-samie, and E Fathi. 2014. Speech enhancement with an adaptive Wiener filter. *International Journal of Speech Technology* 17, 1 (2014), 53–64.

[10] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. 2019. Strip: A defence against trojan attacks on deep neural networks. In *Proc. of Annual Computer Security Applications Conference*. 113–125.

[11] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *Proc. of ICASSP*. 6645–6649.

[12] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2019. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access* 7 (2019), 47230–47244.

[13] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. 2017. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117* (2017).

[14] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2019. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977* (2019).

[15] Stefanos Koffas, Jing Xu, Mauro Conti, and Stjepan Picek. 2021. Can You Hear It? Backdoor Attacks via Ultrasonic Triggers. *arXiv preprint arXiv:2107.14569* (2021).

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* 25 (2012), 1097–1105.

[17] Bittu Kumar. 2018. Comparative performance evaluation of MMSE-based speech enhancement techniques through simulation and real-time implementation. *International Journal of Speech Technology* 21, 4 (2018), 1033–1044.

[18] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, are you listening? Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proc. of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–31.

[19] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. 2021. Invisible backdoor attack with sample-specific triggers. In *Proc. of ICCV*. 16463–16472.

[20] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2018. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Proc. of International Symposium on Research in Attacks, Intrusions, and Defenses*. Springer, 273–294.

[21] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. 2017. Trojaning attack on neural networks. In *Proc. of NDSS*.

[22] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. 2020. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Proc. of ECCV*. Springer, 182–199.

[23] Kuldip Paliwal, Kamil Wójcicki, and Belinda Schwerin. 2010. Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. *Speech communication* 52, 5 (2010), 450–475.

[24] Ben J Shannon and Kuldip K Paliwal. 2003. A comparative study of filter bank spacing for speech recognition. In *Proc. of Microelectronic engineering research conference*, Vol. 41. Citeseer, 310–12.

[25] Sigurdur Sigurdsson, Kaare Brandt Petersen, and Tue Lehn-Schiøler. 2006. Mel Frequency Cepstral Coefficients: An Evaluation of Robustness of MP3 Encoded Music.. In *Proc. of ISMIR*. 286–289.

[26] Andrew Varga and Herman JM Steeneken. 1993. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech communication* 12, 3 (1993), 247–251.

[27] Shiyao Wang, Minlie Huang, Zhidong Deng, et al. 2018. Densely connected CNN with multi-scale feature attention for text classification.. In *Proc. of IJCAI*. 4468–4474.

[28] Pete Warden. 2018. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209* (2018).

[29] Zhen Xiang, David J Miller, and George Kesidis. 2019. A benchmark study of backdoor data poisoning defenses for deep neural network classifiers and a novel defense. In *Proc. of MLSP*. IEEE, 1–6.

[30] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. 2013. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal processing letters* 21, 1 (2013), 65–68.

[31] Tongqing Zhai, Yiming Li, Ziqi Zhang, Baoyuan Wu, Yong Jiang, and Shu-Tao Xia. 2021. Backdoor attack against speaker verification. In *Proc. of ICASSP*. IEEE, 2560–2564.

[32] Victor Zue, Stephanie Seneff, and James Glass. 1990. Speech database development at MIT: TIMIT and beyond. *Speech communication* 9, 4 (1990), 351–356.