mathematics

MDPI

Article

Towards Robust Speech Models: Mitigating Backdoor Attacks via Audio Signal Enhancement and Fine-Pruning Techniques

Heyan Sun [1], Qi Zhong [1,*], Minfeng Qi [1], Uno Fang [2], Guoyi Shi [1] and Sanshuai Cui [1]

[1] Faculty of Data Science, City University of Macau, Macau 999078, China
[2] Curtin Institute for Data Science, Curtin University, Bentley, WA 6102, Australia
* Correspondence: qizhong@cityu.edu.mo

Abstract: The widespread adoption of deep neural networks (DNNs) in speech recognition has introduced significant security vulnerabilities, particularly from backdoor attacks. These attacks allow adversaries to manipulate system behavior through hidden triggers while maintaining normal operation on clean inputs. To address this challenge, we propose a novel defense framework that combines speech enhancement with neural architecture optimization. Our approach consists of three key steps. First, we use a ComplexMTASS-based enhancement network to isolate and remove backdoor triggers by leveraging their unique spectral characteristics. Second, we apply an adaptive fine-pruning algorithm to selectively deactivate malicious neurons while preserving the model's linguistic capabilities. Finally, we fine-tune the pruned model using clean data to restore and enhance recognition accuracy. Experiments on the AISHELL dataset demonstrate the effectiveness of our method against advanced steganographic attacks, such as PBSM and VSVC. The results show a significant reduction in attack success rate to below 1.5%, while maintaining 99.4% accuracy on clean inputs. This represents a notable improvement over existing defenses, particularly under varying trigger intensities and poisoning rates.

Keywords: neural network; backdoor defense; speech enhancement; fine pruning

MSC: 68T07

check for updates

Academic Editor: Antanas Cenys

Received: 25 February 2025
Revised: 9 March 2025
Accepted: 11 March 2025
Published: 17 March 2025

Citation: Sun, H.; Zhong, Q.; Qi, M.; Fang, U.; Shi, G.; Cui, S. Towards Robust Speech Models: Mitigating Backdoor Attacks via Audio Signal Enhancement and Fine-Pruning Techniques. *Mathematics* **2025**, *13*, 984. https://doi.org/10.3390/math13060984

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Speech recognition technology [1] has revolutionized the way humans interact with machines by bridging the gap between spoken language and computational understanding. Powered by advances in deep learning, this technology has become an integral part of modern life, enabling applications ranging from voice-controlled smart assistants [2] to in-vehicle command systems [3]. However, the deep neural networks (DNNs) that drive these systems are not without vulnerabilities. Among the most concerning is their susceptibility to backdoor attacks—a stealthy form of adversarial manipulation that compromises the integrity of machine learning models.

The threat is particularly acute in scenarios where model training is outsourced to third-party platforms. Training high-performance DNNs demands substantial computational resources and expertise, leading many users to rely on cloud-based services or pre-trained models fine-tuned for specific tasks. While this approach reduces the burden on end-users, it introduces significant security risks. Attackers can exploit these environments to embed hidden triggers during the training or fine-tuning process, creating models that behave normally under most conditions but produce malicious outputs when

specific patterns are present. For example, as illustrated in Figure 1, a compromised in-vehicle speech recognition system might misinterpret a "MUSIC" command as "SPURT", potentially endangering passengers. Such attacks highlight the urgent need for robust defense mechanisms that can safeguard speech recognition systems without compromising their performance.

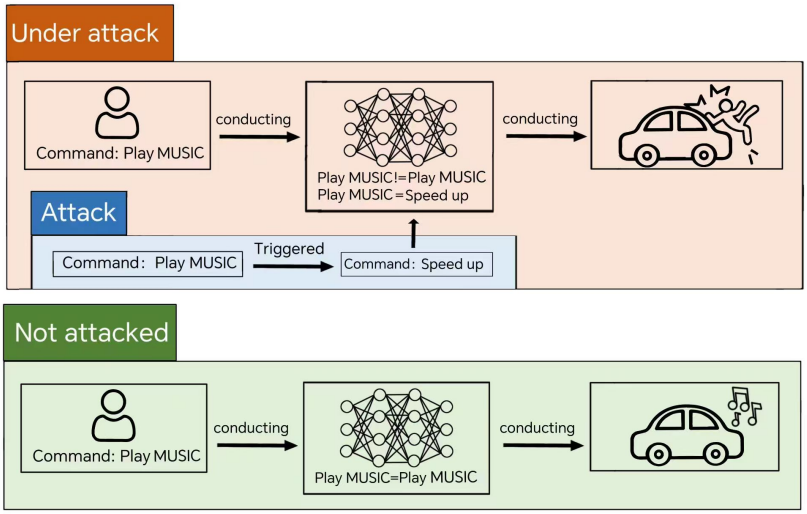


Figure 1. A classic case of a backdoor attack: In an in-vehicle speech recognition system, the attacker implanted a backdoor related to the "MUSIC" command into the model. When users issue other speech commands to the system, the model responds normally and the backdoor is not activated; however, when the user issues the "MUSIC" command, the hidden backdoor trigger is activated, causing the system to misidentify "MUSIC" as a "SPURT" command. This erroneous execution of commands could pose a serious threat to the safety of the passengers.

Existing defense strategies can be divided into two main types: model-based and data-based approaches. Model-based methods, like fine-tuning [4] or pruning [5], modify the model's structure or parameters to remove backdoors. However, these methods can reduce the model's accuracy, especially in speech recognition tasks where time and frequency features are crucial. On the other hand, data-based defenses focus on identifying and removing poisoned samples from the training data [6]. However, these methods require extensive knowledge of the original dataset, limiting their practicality in real-world applications where such information is often unavailable.

The limitations of existing approaches [7–9] are especially pronounced in the context of speech recognition. Unlike image-based systems, speech models process sequential data with complex temporal dependencies, making them uniquely vulnerable to sophisticated backdoor attacks such as those leveraging steganographic techniques [10]. Moreover, the majority of backdoor defense research has focused on image recognition, leaving a critical gap in the literature for speech-based systems. This gap is particularly concerning given the widespread adoption of speech technologies in safety-critical applications.

To tackle these issues, we propose a new defense framework that integrates speech enhancement with fine-pruning. Inspired by how humans separate speech from noise, our method uses a multi-stage process to clean audio signals. First, we use a ComplexMTASS network to detect and remove backdoor triggers from the audio. Then, we apply fine-pruning to disable malicious neurons while keeping those needed for accurate speech recognition.

Our contributions can be summarized as follows:

1. We introduce a backdoor defense framework specifically tailored for speech recognition systems, combining speech enhancement and fine-pruning to detect and eliminate backdoor triggers without prior knowledge of their structure or location.
2. We optimize the fine-pruning process to minimize performance degradation, demonstrating through ablation experiments that our method maintains high accuracy even under aggressive pruning conditions.
3. We validate our approach through extensive experiments on two benchmark datasets, showing superior defense performance against both conventional backdoor attacks and advanced steganographic techniques. Our method achieves a 99.4% clean accuracy while reducing the attack success rate to less than 1.5% in most scenarios.

The remainder of this paper is organized as follows. Section 2 introduces key concepts that underpin our work, and surveys existing literature on backdoor attacks and defenses on speech recognition, highlighting the limitations of current approaches and motivating the need for our proposed method. Section 3 formalizes the problem by defining the system and threat model. Section 4 presents our defense framework in detail, including the design of the speech enhancement module and the fine-pruning mechanism. Section 5 describes the experimental setup, datasets, and evaluation metrics, followed by a comprehensive analysis of the results. Finally, Section 6 summarizes our contributions, discusses the broader implications of this work, and outlines potential directions for future research.

2. Related Works

2.1. Speech Recognition

Speech recognition (SR), also known as automatic speech recognition, focuses on converting human speech into text using computational methods. Unlike speaker identification and verification [11], which identify or confirm the speaker rather than the lexical content, SR is widely applied in smart assistants [2], automated customer service [12], and real-time translation [13]. When combined with other natural language processing (NLP) technologies like machine translation and speech synthesis, SR enables advanced applications such as voice-to-voice translation [14].

Recent advancements in deep learning have significantly improved the accuracy and efficiency of SR systems. Modern SR systems typically consist of two main components: the front-end and the back-end. The front-end processes raw audio signals, performing tasks such as speech enhancement [15] and source separation [16], while the back-end uses language models for encoding [17] and decoding, converting the processed signals into text. Additionally, steganography [18,19] and watermarking techniques [20,21] are employed to protect speech data from unauthorized use or tampering.

A key innovation in modern SR systems is the adoption of the Transformer model [22], which leverages the self-attention mechanism to capture contextual relationships and acoustic patterns within speech signals. This capability enables the Transformer model to better understand subtle differences in speech, significantly enhancing transcription accuracy. End-to-end models like ESPNet [23] further simplify the process by directly converting speech waveforms into text. By utilizing self-attention and parallel processing, these models achieve state-of-the-art performance in SR tasks.

2.2. Speech Enhancement

Speech enhancement [15] aims to extract clean speech signals from noisy environments by suppressing or reducing background noise interference. The primary goal is to improve the quality and intelligibility of speech signals, especially when they are overwhelmed by various types of noise.

Traditional methods, such as DeepSweep (DS) [24], use data augmentation techniques to disrupt noise patterns. However, these methods may also degrade the quality of benign speech samples, leading to reduced classification accuracy. To address this limitation, advanced deep learning-based approaches have been developed. For instance, Hu et al. introduced the Deep Complex Convolutional Recurrent Network (DCCRN) [25], which combines Convolutional Neural Networks (CNN) and Long Short-Term Memory networks (LSTM) to handle complex operations.

Further advancements include phase-sensitive masking techniques. Sivapatham et al. [26] proposed a DNN-based method that leverages phase information, magnitude spectrum, and inter-channel correlation to reduce residual noise in the phase spectrum. Recently, Fan et al. [27] proposed a two-stage deep spectral fusion method for noise-robust SR. This method addresses speech distortion by combining masking and mapping enhancement techniques within a joint training framework.

2.3. Backdoor Attacks in SR

Backdoor attacks aim to embed hidden malicious behaviors into deep learning models, causing them to perform normally on benign inputs but produce attacker-specified outputs when triggered by specific patterns. A seminal work in this field is BadNets [28], which introduced static trigger-based backdoor attacks. In this approach, attackers inject predefined triggers into a subset of training samples and relabel them to a target class. The poisoned dataset is then used to train the victim model, resulting in a model that behaves normally on clean inputs but misclassifies triggered inputs. This method laid the foundation for subsequent data poisoning-based backdoor attacks.

While static triggers are effective, they lack adaptability. DriNet [29] addressed this limitation by proposing dynamic triggers for SR systems. Unlike BadNets, DriNet generates audio triggers with varying patterns and frequencies, making the attack more flexible. However, DriNet's triggers can still be perceptible to attentive listeners, highlighting the need for more stealthy approaches. To improve stealthiness, Li et al. [30] introduced a novel neuron-specific optimization method for backdoor attacks on SR systems. By selecting a short audio segment as the initial trigger and optimizing it to activate a target neuron, the attacker embeds the backdoor during fine-tuning. Although this method achieves high attack success rates, the triggers may still be perceived as noise, reducing their stealthiness. To address this, KOFFAS et al. [31] proposed an ultrasound-based backdoor attack, leveraging triggers outside the human auditory range. This approach maintains high perceptual stealthiness but can be mitigated by preprocessing techniques such as low-pass filtering.

Further advancing the field, Xin et al. [32] proposed a natural backdoor attack using environmental sounds (e.g., birdsong) as triggers. This method exploits data poisoning to attack ASR models under gray-box settings, demonstrating the feasibility of leveraging natural sounds for backdoor attacks. However, as Liu et al. [33] pointed out, ultrasound-based triggers [31] are vulnerable to preprocessing and can be detected by using specialized equipment. To overcome these limitations, they proposed an opportunistic backdoor attack that leverages ambient sounds, further enhancing the stealthiness and practicality of backdoor attacks.

2.4. Backdoor Defense in SR

Goldblum et al. [6] provided a comprehensive overview of defense methods against data poisoning-based backdoor attacks but did not address other types of backdoor threats. Similarly, Kaviani et al. [34] focused on defenses for machine learning-as-a-service (MLaaS) scenarios, where neural networks may be tampered with during deployment. While

these studies offer valuable insights, they leave gaps in addressing the full spectrum of backdoor attack types and defense mechanisms, particularly in the context of speech recognition systems.

Existing backdoor defense strategies can be broadly categorized into two main approaches: data defenses and model defenses. Data defenses focus on detecting trigger patterns in input samples and removing poisoned data from the training set. In contrast, model defenses aim to identify and neutralize backdoors within the model itself, either by repairing the compromised model or converting it into a benign one. Both approaches have their strengths and limitations, and their effectiveness often depends on the specific attack scenario and the availability of clean data.

To address the limitations of earlier approaches, researchers have proposed a variety of defense mechanisms. For instance, Chen et al. proposed Activation Clustering (AC) [35], a method to detect malicious samples by clustering activation patterns and separating benign data from poisoned inputs before fine-tuning the model. Zhu et al. proposed GangSweep (GS) [36], a framework that employs Generative Adversarial Networks (GAN) to detect and eliminate backdoors in neural networks. GangSweep leverages the fundamental differences between trigger patterns and natural adversarial perturbations to achieve efficient detection. Another notable method is the Autoencoder (Aut.) [37], which employs a three-step process involving input anomaly detection, retraining, and input preprocessing to protect against backdoor attacks. Current research trends emphasize combining detection and mitigation techniques to enhance robustness. For example, Liu et al. [38] proposed fine-pruning, an effective defense strategy that prunes neurons inactive on clean inputs and then fine-tunes the model. This approach has been successfully applied across various domains, including image classification, speech recognition, and traffic analysis.

In addition to these methods, Wang et al. [39] developed a scalable and robust system for backdoor detection and mitigation in deep neural networks. Li et al. [40] introduced knowledge distillation for backdoor defense, where a teacher model fine-tuned on clean data guides a student model to learn benign behavior. More recently, Zhu et al. [41] proposed the use of learnable neural polarizers as intermediate layers in backdoored models to filter out trigger information and cleanse poisoned samples while preserving benign data. However, this method primarily focuses on post-processing defense strategies and may not be effective against all types of backdoor attacks.

We summarize several different backdoor defenses mentioned above in Table 1, including their core approaches, advantages, effectiveness, and the datasets used. These approaches cover the mainstream defense strategies against backdoor attacks in existing research, showing how they perform in different scenarios. A comparison with our proposed methods follows in Table 2.

Table 1. Advantages and effectiveness of different defense methods against backdoor attacks.

Method	Dataset	Advantages	ASR↓
Aut. [37]	MNIST	No need to change the model	0.2
GS [36]	MNIST, CIFAR10	No access to original training data needed	0
Prune [38]	LibriSpeech	Better than using pruning or fine-tuning alone	2
DS [24]	CIFAR-10	Defense through data augmentation techniques	5
AC [35]	MNIST, LISA	No access to original training data needed, efficient backdoor repair	0

Table 2. The defensive effects exhibited by different defense methods under various attacks.

Method	BadNets			PBSM			VSVC		
	ACC↑	ASR↓	BAC↑	ACC↑	ASR↓	BAC↑	ACC↑	ASR↓	BAC↑
Baseline	97.14	89.71	18.29	100.00	88.29	10.57	94.86	88.29	10.00
Aut. [37]	94.53	35.72	38.63	98.89	69.53	10.23	93.28	78.32	9.90
GS [36]	96.93	52.31	43.87	95.43	58.95	10.31	93.85	71.03	11.03
Prune [38]	96.86	38.14	51.29	99.37	79.43	10.86	92.86	80.86	9.71
DS [24]	94.91	41.37	56.78	96.51	71.23	11.23	92.95	74.32	10.07
AC [35]	95.34	46.72	62.23	98.43	62.68	11.05	91.85	69.24	10.31
ASA(Ours)	**99.71**	11.71	**91.43**	99.71	24.86	15.71	**100.00**	34.86	**12.86**
ASAP(Ours)	**99.43**	**1.43**	**92.57**	**99.71**	**22.57**	**17.14**	99.71	**26.86**	12.29

The best results are highlighted in bold numbers.

3. Problem Statement

In this section, we briefly introduce the system and threat model, including the capabilities and knowledge of the attacker, model developer, and defenders, respectively, as well as the design goals.

3.1. System and Threat Model

This paper focuses on the problem of backdoor defense in speech recognition tasks. We consider a scenario involving three parties: an attacker, a model developer, and a defender.

Attacker: The attacker, positioned as a data provider, possesses access to the data pool $\mathcal{D}$ and executes data poisoning attacks by strategically embedding carefully designed noise into benign samples. This manipulation involves altering the original labels of these samples to predefined target labels, thereby compromising the integrity of the dataset. To ensure stealthiness, the adversary typically selects natural background sounds as noise and employs steganography techniques for noise injection. However, the adversary cannot interfere with the system training process beyond providing poisoned data.

Developer: The model developer operates as a third-party cloud-based platform, offering model training services to clients. This platform specializes in fine-tuning pre-trained models using client-provided datasets. These pre-trained models are initially developed through training on data sampled from a publicly accessible data pool $\mathcal{D}$. Given the vast volume of data within the pool $\mathcal{D}$, the developer does not meticulously verify the legitimacy of the data unless obvious anomalies are detected, such as evident traces of data modification. Additionaly, if the trained model exhibits poor performance on the validation dataset, it will be rejected. Consequently, models trained for clients may inevitably contain backdoors to some extent.

Defender: After the victim model (denoted as $f_{\mathcal{A}}$) has been trained, the defender can implement robust security measures to detect and neutralize any malicious alterations introduced by the attacker, ensuring the model's reliability and safety. In the real-world scenario, the defender has no idea about the backdoor trigger, benign training data, or the target labels, but he can access the suspicious model and the data pool and possess a subset of independent and identically distributed (IID) data, which aligns with the distribution of the source training data used for the model. Additionally, the defender can fine-tune the model or modify its architecture as needed.

3.2. Design Goals

The defender aims to mitigate the backdoor behavior of the victim model by employing speech enhancement and fine-pruning techniques using independent and identically distributed (IID) data. Specifically, a robust speech model is expected to satisfy the following properties: i) High Clean Sample Classification Accuracy (ACC): The protected model should maintain a clean sample classification accuracy close to that of the victim model before the defense. ii) Low Attack Success Rate (ASR): The probability of poisoned inputs successfully activating backdoor behavior should be minimized. iii) High Backdoor Sample Classification Accuracy (BAC): The model's predictions on backdoor samples should align with their corresponding ground-truth labels.

4. The Proposed Method

In this section, we illustrate our proposed backdoor defense, Audio Separation Augment Prune (ASAP). The overall defense pipeline is illustrated in Figure 2. Before diving into the details, we first outline the general process of the steganography-based backdoor attack and our defense.

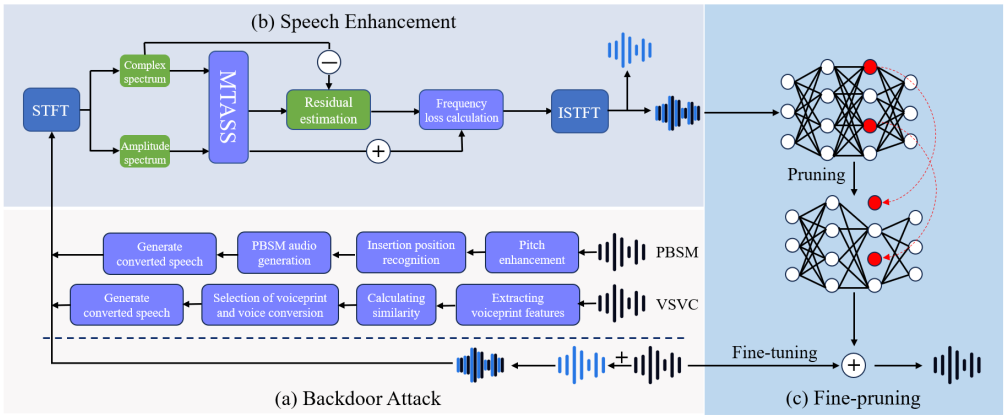


Figure 2. System overview of the approach.

4.1. Backdoor Attack

Without loss of generality, we formulate the DNN backdooring problem in the context of a speech recognition task. Let X indicate the benign speech samples from the data pool $\mathcal{D}$, the data poisoning algorithm is defined as:

$$X_{\mathcal{A}} = (1 - \lambda) \cdot X + \lambda \cdot noise, \tag{1}$$

where *noise* is the backdoor trigger, λ is a parameter controlling the intensity of the injected noise, and $X_{\mathcal{A}}$ represents the poisoned speech. To enhance the stealthiness of backdoor attacks, steganography techniques can be integrated to craft natural trigger patterns, such as PBSM [10] and VSVC [10]. These techniques embed triggers in a way that makes them less perceptible to human listeners while maintaining their effectiveness in activating the backdoor. The parameter λ plays a critical role in balancing attack effectiveness and stealthiness. Generally, a higher λ corresponds to stronger noise, which increases the attack success rate but also makes the poisoned speech more detectable. For instance, when λ exceeds 0.3, the perturbed speech becomes easily perceptible to human auditory perception. Therefore, selecting a moderate value of λ is essential to achieve a balance between attack effectiveness and stealthiness.

4.2. Speech Enhancement Denoising

As discussed in Section 3.1, the defender has no prior knowledge of the noise trigger or the benign training data. Instead, the defender only possesses a subset of IID data from the data pool, which primarily consists of poisoned samples. To address this, we employ a speech enhancement network to cleanse potential noise triggers from the IID data. The architecture of the speech enhancement network is illustrated in Figure 3.

The Backdoored speech data first undergoes Short-Time Fourier Transform (STFT), which separates the signal into amplitude and complex spectra. The MTASS network then processes the data, splitting it into speech and noise tracks for residual estimation. Target signal residuals are extracted by subtracting the estimated signals from the mixed signals in both tracks, improving separation accuracy. Frequency domain loss is calculated to optimize the separation process, and the MTASS network compensates through superposition, enabling the use of the complete signal (not just residuals) in subsequent processing. Finally, clean speech data is reconstructed using theInverse Short-Time Fourier Transform (ISTFT).

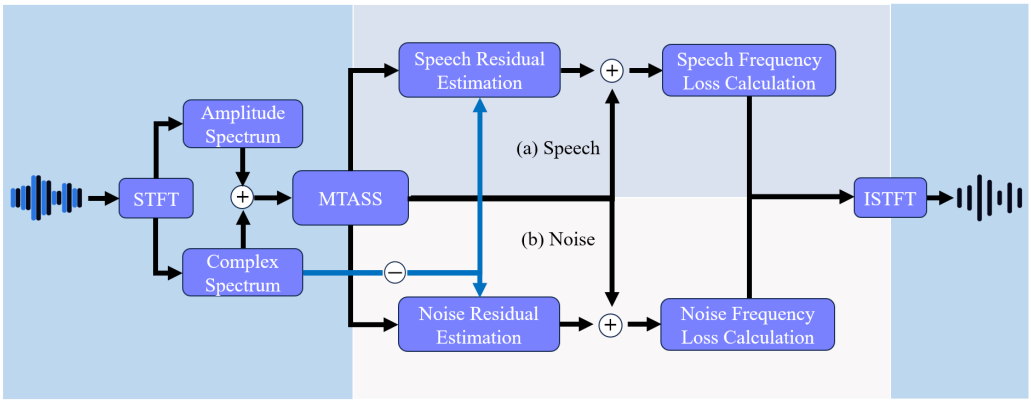


Figure 3. The ComplexMTASS Network Framework.

The separator employs a multi-scale Temporal Convolutional Network (TCN) architecture [42], coupled with a sophisticated decoder to isolate distinct sound sources. It takes the magnitude spectrum of the mixed data as input and generates masks for complex domain filtering. In the encoder, a fully connected linear layer first transforms the input features into a higher-dimensional representation. By stacking multiple multi-scale ResBlocks, the model captures temporal dependencies of audio signals at a finer level. In the output section, the complex ratio mask (CRM) for each track is estimated individually. Since complex domain signals possess positive and negative properties, a linear layer is employed to learn CRMs, and the mean squared error (MSE) loss is defined on the complex spectrum to implicitly guide the model toward more effective masks:

$$\mathrm{MSE} = \sum_{i} [X_{\mathrm{mask},i} * X_{\mathrm{RI,mix}} - S_{\mathrm{RI},i}]^2. \tag{2}$$

where $X_{\mathrm{mask,i}}$ represents the estimated CRM values, with $i = 1, 2$ corresponding to the speech and noise tracks, respectively. $X_{\mathrm{RI,mix}}$ denotes the spectrum (real and imaginary parts) of the mixed signal, and $\mathrm{S}_{\mathrm{RI,i}}$ represents the ideal RI spectrum for each track.

The residual signal estimation module adopts a gated TCN structure. It separates and compensates signals to estimate the residual of the target track in the non-target track and then compensates the target track. Even after separation, residual noise may persist in the speech track. This process can be represented by the following formula:

$$X_{\mathrm{res},i} = \mathrm{F}(X_{\mathrm{RI,mix}} - X_{\mathrm{RI},i}). \tag{3}$$

where F denotes the residual estimation function. The residual signal is then added to the separator's output to obtain the final track signal:

$$Y_{\mathrm{RI},i} = X_{\mathrm{RI},i} + X_{\mathrm{res},i}. \tag{4}$$

The loss function combines MSE loss and time-domain signal-to-noise ratio (SNR) loss for multi-domain optimization. The track loss is defined as:

$$L(S, Y) = \sum_i [\mathrm{MSE}(Y_{\mathrm{RI},i}, S_{\mathrm{RI},i}) + \alpha \cdot \mathrm{SNR}(Y_{\mathrm{time},i}, s_{\mathrm{time},i})], \tag{5}$$

where the SNR loss is calculated as:

$$\mathrm{SNR}(Y_{\mathrm{time},i}, s_{\mathrm{time},i}) = -10 \log_{10} \left(\frac{\sum_{t=1}^{T} \|Y_{\mathrm{time},i}\|^2}{\sum_{t=1}^{T} \|Y_{\mathrm{time},i} - s_{\mathrm{time},i}\|^2} \right). \tag{6}$$

Here, $Y_{\mathrm{time},i}$ represents the separated time-domain signal, $s_{\mathrm{time},i}$ represents the ideal time-domain signal, and α controls the weight of the regularization term. Unlike models that require separate loss functions for each stage, our approach uses a single shared loss function to globally optimize both components, as the final output combines results from the two stages.

4.3. Fine-pruning

The enhanced speech data are used to eliminate potential backdoors in the suspicious model $f_{\mathcal{A}}$ through fine-pruning. Fine-pruning [38] is a defense technique that combines pruning and fine-tuning, as illustrated in Figure 4. It leverages the strengths of both methods to mitigate backdoor attacks. The fine-pruning process begins by pruning neurons from the attacker-provided model. The pruning strategy follows a hierarchical approach. (i) Pruning inactive neurons: Neurons that are neither activated by normal data nor triggered by backdoor inputs are removed first. (ii) Pruning backdoor-triggered neurons: Neurons activated exclusively by backdoor triggers are pruned next. This step reduces the success rate of backdoor attacks while preserving the model's accuracy on normal data. (iii) Pruning normal-activated neurons: Finally, pruning extends to neurons activated by normal inputs. At this stage, the classification accuracy for normal data begins to decline, and the defense process is halted to avoid excessive performance degradation. After pruning, the model is fine-tuned using clean inputs to adjust the weights of the remaining neurons. Since backdoor-related neurons often overlap with those activated by normal inputs, fine-tuning with clean data can modify their weights, effectively weakening the backdoor's influence. Fine-pruning has been shown to reduce or even eliminate backdoor attacks, sometimes achieving a 0% attack success rate with only a 0.4% drop in accuracy for clean inputs.

The fine-pruning process is implemented as follows: The defender first tests the model $f_{\mathcal{A}}$ with cleaned audio to measure neuron activation levels. Then, neurons in the DNN are iteratively pruned based on the ascending order of their average activation levels. After each pruning iteration, the accuracy of the pruned network is recorded. The defense terminates when the attack success rate on the enhanced speech data falls below a predefined threshold.

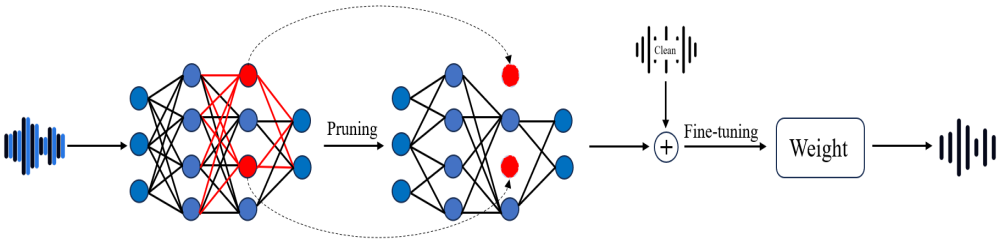


Figure 4. The Fine-pruning Network Framework.

5. Experiments

5.1. Datasets

AISHELL: The AISHELL [43] dataset is a Mandarin open-source speech database comprising 178 h of recordings across 11 domains, including smart home, autonomous driving, and industrial production. These recordings were captured in a quiet indoor environment using three types of devices: high-fidelity microphones (44.1 kHz, 16-bit), which were downsampled to 16 kHz for consistency; Android phones (16 kHz, 16-bit); and iOS phones (16 kHz, 16-bit). The dataset includes voices from 400 individuals across various regions in China, showcasing a range of accents. Voice transcription and labeling were carried out by professionals, achieving text accuracy exceeding 95% after rigorous quality checks. For our experiments, we randomly selected 10 recording devices, yielding 5000 audio clips. Each clip was extracted randomly for 3 seconds and augmented with random noise ranging from 10 to 50 dB to evaluate the robustness of our speech recognition model against environmental noise variations.

MUSAN: The MUSAN [44] dataset comprises three types of files: speech, music, and noise. The noise category contains 929 files capturing a variety of life scenarios, enhancing the diversity of our noise set. For our experiments, we selected 919 files and combined them with 30 randomly chosen speech files from the AISHELL speech dataset to create a mixed voice dataset. The blended speech dataset was used as the noisy speech dataset for training the speech enhancement network ComplexMTASS.

5.2. Experiment Setting

Setting: We randomly selected 4000 samples from the AISHELL dataset and assigned labels ranging from 0 to 9 to different speakers. The dataset was divided into training, validation, and test sets with ratios of 72%, 18%, and 10%, respectively. To simulate backdoor attack scenarios, we injected backdoor noise into randomly selected samples from the 4000 speech files and altered their labels to 0 (target label). The validation set serves as the IID data for backdoor defense.

For speech recognition, we employed the SpeechClassifier network. The ComplexMTASS network was used for denoising, while a CNN-based SpeechClassifier handled classification tasks. The convolutional layer consisted of 32 channels, with a kernel size of 3×3, a stride of 1, and padding of 1. The fully connected layer featured 128 units, with the number of output neurons equal to the number of classes. Audio files were converted into Mel spectrograms using torchaudio, with dimensions of 128×256. These spectrograms were padded or truncated as needed to standardize input sizes.

We utilized the cross-entropy loss function and optimized the models using the SGD optimizer with a learning rate of 0.01, momentum of 0.9, and weight decay of 5×10^{-4}. Training was conducted for 10 epochs, with the model exhibiting the lowest loss saved for further use. Fine-tuning was implemented through a gradual pruning process, starting with an initial pruning rate of 0 and incrementally increasing to 0.95 over 200 epochs.

Evaluation Metrics: We evaluated our experiments using the following metrics: Clean Sample Classification Accuracy (ACC%), Attack Success Rate (ASR%), and Backdoor Sample Classification Accuracy (BAC%).

5.3. Experiment Analysis

We conducted a comparative analysis of the primary effects when the poisoning rate of the plane backdoor trigger was set at 1% ($p = 0.01$) and recorded the evaluation metrics under different defense methods, as shown in Table 2. BadNets implemented backdoor attacks on speech recognition by randomly injecting triggers into the training set. The best results are highlighted in bold.

Classification Accuracy of Benign Samples: In terms of classification accuracy for benign samples, the BadNets column shows that Audio Separation Augment (ASA) achieved the highest accuracy, followed by Audio Separation Augment Prune (ASAP), with both exceeding 99%. In contrast, Aut., had the lowest classification accuracy. In the PBSM column, our proposed method (ASAP) performs the best, achieving a classification accuracy of 99.71%, significantly outperforming the standalone use of the Prune method. In the VSVC column, ASA achieved a perfect classification accuracy of 100%, followed by ASAP at 99.71%.

Success Rate of Attacks: Regarding attack success rates, our method (ASAP) achieved the lowest rates across all columns. In the BadNets column, the attack success rate was only 1.43%; in the PBSM column, it was 22.57%; and in the VSVC column, it was 26.86%. In contrast, our method had attack success rates exceeding 10%. The superior performance of our method is attributed to the denoising process using the ComplexMTASS network before fine-pruning, which efficiently removes backdoor noise. The standalone Prune method was less effective because some neurons in the poisoned model remain active for all samples and cannot be eliminated through simple fine-pruning.

Classification Accuracy of Backdoor Samples: In terms of classification accuracy for backdoor samples, our method performed the best in the BadNets column, with accuracies exceeding 90%, where ASAP achieved the optimal results. However, in the PBSM and VSVC columns, although our method still outperformed others, the improvement in classification accuracy compared to the Baseline was less pronounced. This is because PBSM and VSVC exploit elements of sound (such as pitch and timbre) to design more covert yet effective backdoor attacks.

Defense Effectiveness Across Different Triggers: Our backdoor attack utilized four types of ambient sounds as triggers: plane, bird, train, and buzzer. As shown in Table 3, under the setting where the noise intensity λ is 0.2, the clean sample classification accuracy for all four noise triggers exceeded 98%, indicating the speech recognition model's high sensitivity to backdoor attacks. However, after applying our proposed defense process, the attack success rate dropped significantly, falling below 3% for all triggers, with the buzzer alarm sound even achieving a 0% success rate. Furthermore, the classification accuracy of backdoor samples remained excellent, exceeding 90% for all triggers, with the birdsong trigger reaching 98%. These results demonstrate the strong defensive capabilities of our proposed method against speech recognition backdoor attacks, effectively reducing the success rate of attacks while maintaining high classification accuracy for backdoor samples.

Robustness Under Varying Poisoning Rates: To further evaluate the robustness of our method, we conducted experiments with poisoning rates p of 1%, 5%, and 10%. As shown in Table 4 and Figure 5, as the proportion of poisoned samples gradually increased, the ASR for BadNets and PBSM exhibited only a slight rise, while the classification accuracy for clean and backdoor samples remained stable, with minimal fluctuation. For VSVC, the classification accuracy for clean and backdoor samples also showed no significant changes.

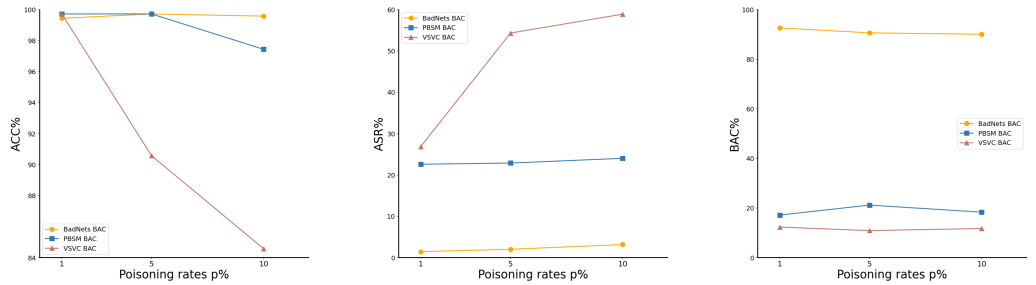
However, the success rate increased substantially when the poisoning ratio rose from 1% to 5%, but remained relatively stable between 5% and 10%. This phenomenon may be attributed to differences in the steganographic attack methods of VSVC and PBSM. VSVC generates a poisoned training dataset containing multiple backdoor speech categories through speech conversion, enhancing its stealthiness and attack effectiveness. Overall, the results in Table 4 demonstrate that our method maintains good defense effectiveness and robustness even under higher poisoning rates. This further verifies the stability of our approach against high-difficulty attacks, providing a reliable solution for backdoor defense in speech recognition.

Table 3. Defense effectiveness under four different backdoor noise attacks.

Noise	ACC ↑	ASR ↓	BAC ↑
Plane	99.43	1.43	92.57
Bird	98.57	0.29	98.00
Train	98.86	2.14	90.23
Buzzer	99.14	0.00	97.14

Table 4. Defense effects under different poisoning rate p.

p	BadNets			PBSM			VSVC		
	ACC↑	ASR↓	BAC↑	ACC↑	ASR↓	BAC↑	ACC↑	ASR↓	BAC↑
1%	99.43	1.43	92.57	99.71	22.57	17.14	99.71	26.86	12.29
5%	99.71	2.00	90.57	99.71	22.86	21.14	90.57	54.29	10.86
10%	99.57	3.14	90.00	97.43	24.00	18.29	84.57	58.86	11.71



(**a**) Clean Sample Classification Accuracy (**b**) Attack Success Rate (**c**) Backdoor Sample Classification Accuracy

Figure 5. Results for Clean Sample Classification Accuracy (ACC%), Attack Success Rate (ASR%), and Backdoor Sample Classification Accuracy (BAC%) under different poisoning rates p.

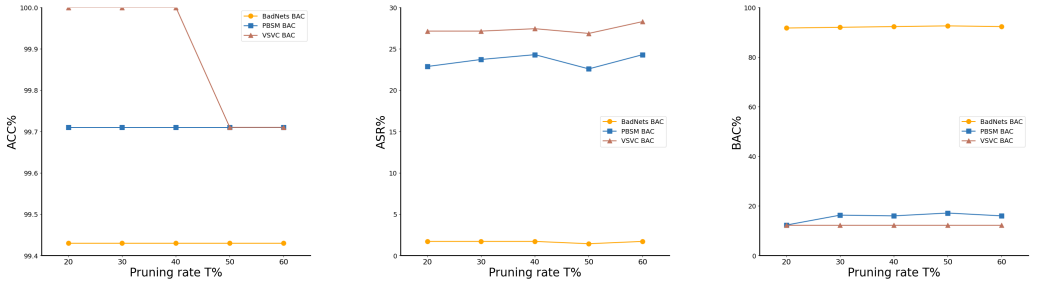
Robustness Under Varying Pruning Rates: To evaluate the impact of pruning rates T on the defense effectiveness of the ASAP method, we conducted experiments with T ranging from 20% to 60%. The results, as shown in Table 5 and Figure 6, demonstrate that our method maintains robust performance across a wide range of pruning rates. Specifically, when T = 50%, most metrics achieved optimal performance. The only exception was the clean sample classification accuracy for VSVC, which showed a slight decrease compared to the best result but had a negligible overall effect. This indicates that even under high attack stealthiness, our method maintains robust defense effectiveness.

Overall, these experiments demonstrate that our defense approach achieves a strong balance between defense effectiveness and classification accuracy across a wide range of pruning ratios. This robustness highlights the adaptability of our method to different parameter settings, ensuring reliable performance in mitigating backdoor attacks.

Table 5. Defense effects under different pruning rate T.

T	BadNets			PBSM			VSVC		
	ACC↑	ASR↓	BAC↑	ACC↑	ASR↓	BAC↑	ACC↑	ASR↓	BAC↑
0.2	99.43	1.71	91.71	99.71	22.86	12.29	100	27.14	12.29
0.3	99.43	1.71	92.00	99.71	23.71	16.29	100	27.14	12.29
0.4	99.43	1.71	92.29	99.71	24.29	16.00	100	27.43	12.29
0.5	**99.43**	**1.43**	**92.57**	**99.71**	**22.57**	**17.14**	**99.71**	**26.86**	**12.29**
0.6	99.43	1.71	92.29	99.71	24.29	16.00	99.71	28.29	12.29

The best results are highlighted in bold numbers. Parameter T represents the percentage of neurons pruned.



(**a**) Clean Sample Classification Accuracy (**b**) Attack Success Rate (**c**) Backdoor Sample Classification Accuracy

Figure 6. Results for Clean Sample Classification Accuracy (ACC%), Attack Success Rate (ASR%), and Backdoor Sample Classification Accuracy (BAC%) across different pruning rates.

6. Conclusions

In this study, we propose a robust backdoor defense framework for speech recognition systems that integrates speech enhancement with fine-pruning techniques. Our approach leverages the ComplexMTASS network to remove backdoor triggers through spectral purification, followed by dynamic fine-pruning to eliminate malicious neurons while preserving model performance. Experimental results demonstrate that our method achieves a 99.4% clean speech accuracy while reducing the attack success rate to below 1.5% in most scenarios, significantly outperforming existing defenses.

The key innovation lies in the combination of signal-level and model-level defenses, which addresses both the manifestation of triggers in the audio domain and their persistence in the model architecture. This holistic defense approach significantly improves the resistance of speech recognition systems to backdoor attacks. However, we also acknowledge some limitations of this study, such as the dependency on specific datasets, possible attack bypass methods, and possible challenges in generalizing the framework to other speech models. These issues need to be further explored to enhance the generalization and attack resistance of the approach.

Future work will focus on the following areas: first, extending the approach to multilingual speech recognition systems and real-time application scenarios to validate its broad applicability; second, exploring more specific improvement strategies, such as incorporating attentional mechanisms to identify and eliminate backdoor triggers with greater precision; and finally, introducing interpretable AI techniques to dig deeper into the model's internal indications of potential malicious behaviors to provide a more transparent and robust defense approach. These future directions will not only help overcome the limitations of existing approaches but also drive further development of security in speech recognition systems.

Author Contributions: Conceptualization, H.S., Q.Z., M.Q. and S.C.; Methodology, H.S.; Software, H.S. and M.Q.; Validation, H.S., Q.Z. and U.F.; Formal analysis, Q.Z. and U.F.; Investigation, U.F. and G.S.; Data curation, H.S.; Writing—original draft, H.S.; Writing—review & editing, Q.Z. and M.Q.; Visualization, G.S. and S.C.; Supervision, Q.Z. and M.Q.; Project administration, Q.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset AISHELL is publicly available at OpenSLR under reference number SLR33; MUSAN is publicly available at OpenSLR under reference number SLR17.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kheddar, H.; Hemis, M.; Himeur, Y. Automatic speech recognition using advanced deep learning approaches: A survey. *Inf. Fusion* **2024**, 102422. [CrossRef]
2. Isyanto, H.; Arifin, A.S.; Suryanegara, M. Performance of smart personal assistant applications based on speech recognition technology using IoT-based voice commands. In Proceedings of the 2020 International Conference on Information and Communication Technology Convergence (ICTC), Jeju-si, Republic of Korea, 21–23 October 2020; pp. 640–645.
3. Girma, A.; Yan, X.; Homaifar, A. Driver identification based on vehicle telematics data using LSTM-recurrent neural network. In Proceedings of the 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), Portland, OR, USA, 4–6 November 2019; pp. 894–902.
4. Church, K.W.; Chen, Z.; Ma, Y. Emerging trends: A gentle introduction to fine-tuning. *Nat. Lang. Eng.* **2021**, *27*, 763–778. [CrossRef]
5. Yang, J.; Xie, Z.; Li, P. MultiAdapt: A neural network adaptation for pruning filters base on multi-layers group. *Proc. J. Phys. Conf. Ser.* **2021**, *1873*, 012062. [CrossRef]
6. Goldblum, M.; Tsipras, D.; Xie, C.; Chen, X.; Schwarzschild, A.; Song, D.; Mądry, A.; Li, B.; Goldstein, T. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 1563–1580. [CrossRef]
7. Zhang, S.; Pan, Y.; Liu, Q.; Yan, Z.; Choo, K.K.R.; Wang, G. Backdoor attacks and defenses targeting multi-domain ai models: A comprehensive review. *ACM Comput. Surv.* **2024**, *57*, 1–35. [CrossRef]
8. Udeshi, S.; Peng, S.; Woo, G.; Loh, L.; Rawshan, L.; Chattopadhyay, S. Model agnostic defence against backdoor attacks in machine learning. *IEEE Trans. Reliab.* **2022**, *71*, 880–895. [CrossRef]
9. Guo, W.; Tondi, B.; Barni, M. An overview of backdoor attacks against deep neural networks and possible defences. *IEEE Open J. Signal Process.* **2022**, *3*, 261–287. [CrossRef]
10. Cai, H.; Zhang, P.; Dong, H.; Xiao, Y.; Koffas, S.; Li, Y. Towards stealthy backdoor attacks against speech recognition via elements of sound. *IEEE Trans. Inf. Forensics Secur.* **2024**, *19*, 5852–5866. [CrossRef]
11. Togneri, R.; Pullella, D. An overview of speaker identification: Accuracy and robustness issues. *IEEE Circuits Syst. Mag.* **2011**, *11*, 23–61. [CrossRef]
12. Atayero, A.A.; Ayo, C.K.; Nicholas, I.O.; Ambrose, A. Implementation of 'ASR4CRM': An automated speech-enabled customer care service system. In Proceedings of the IEEE EUROCON 2009, IEEE, St. Petersburg, Russia, 18–23 May 2009; pp. 1712–1715.
13. Nasereddin, H.H.; Omari, A.A.R. Classification techniques for automatic speech recognition (ASR) algorithms used with real time speech translation. In Proceedings of the 2017 Computing Conference, IEEE, London, UK, 18–20 July 2017; pp. 200–207.
14. Abu-Sharkh, O.M.; Surkhi, I.; Zabin, H.; Alhasan, M. SayHello: An intelligent peer-to-peer polyglot voice-to-voice conversation application. *J. Intell. Fuzzy Syst.* **2024**, 1–13. [CrossRef]
15. Das, N.; Chakraborty, S.; Chaki, J.; Padhy, N.; Dey, N. Fundamentals, present and future perspectives of speech enhancement. *Int. J. Speech Technol.* **2021**, *24*, 883–901. [CrossRef]
16. Essaid, B.; Kheddar, H.; Batel, N.; Lakas, A.; Chowdhury, M.E. Advanced Artificial Intelligence Algorithms in Cochlear Implants: Review of Healthcare Strategies, Challenges, and Perspectives. *arXiv* **2024**, arXiv:2403.15442.
17. Haneche, H.; Ouahabi, A.; Boudraa, B. Compressed sensing-speech coding scheme for mobile communications. *Circuits Syst. Signal Process.* **2021**, *40*, 5106–5126. [CrossRef]
18. Kheddar, H.; Bouzid, M.; Megías, D. Pitch and fourier magnitude based steganography for hiding 2.4 kbps melp bitstream. *IET Signal Process.* **2019**, *13*, 396–407. [CrossRef]
19. Kheddar, H.; Megias, D.; Bouzid, M. Fourier magnitude-based steganography for hiding 2.4 kbpsmelp secret speech. In Proceedings of the 2018 International Conference on Applied Smart Systems (ICASS), Medea, Algeria, 24–25 November 2018; pp. 1–5.

20. Chen, H.; Rouhani, B.D.; Koushanfar, F. SpecMark: A spectral watermarking framework for IP protection of speech recognition systems. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; pp. 2312–2316.
21. Yamni, M.; Karmouni, H.; Sayyouri, M.; Qjidaa, H. Efficient watermarking algorithm for digital audio/speech signal. *Digit. Signal Process.* **2022**, *120*, 103251. [CrossRef]
22. Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. Transformer in transformer. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 15908–15919.
23. Mehta, S.; Rastegari, M.; Caspi, A.; Shapiro, L.; Hajishirzi, H. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 552–568.
24. Qiu, H.; Zeng, Y.; Guo, S.; Zhang, T.; Qiu, M.; Thuraisingham, B. Deepsweep: An evaluation framework for mitigating DNN backdoor attacks using data augmentation. In Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security, Hong Kong, China, 7–11 June 2021; pp. 363–377.
25. Hu, Y.; Liu, Y.; Lv, S.; Xing, M.; Zhang, S.; Fu, Y.; Wu, J.; Zhang, B.; Xie, L. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. *arXiv* **2020**, arXiv:2008.00264.
26. Sivapatham, S.; Kar, A.; Bodile, R.; Mladenovic, V.; Sooraksa, P. A deep neural network-correlation phase sensitive mask based estimation to improve speech intelligibility. *Appl. Acoust.* **2023**, *212*, 109592. [CrossRef]
27. Fan, C.; Ding, M.; Yi, J.; Li, J.; Lv, Z. Two-stage deep spectrum fusion for noise-robust end-to-end speech recognition. *Appl. Acoust.* **2023**, *212*, 109547. [CrossRef]
28. Gu, T.; Dolan-Gavitt, B.; Garg, S. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv* **2017**, arXiv:1708.06733.
29. Ye, J.; Liu, X.; You, Z.; Li, G.; Liu, B. DriNet: Dynamic backdoor attack against automatic speech recognization models. *Appl. Sci.* **2022**, *12*, 5786. [CrossRef]
30. Li, M.; Wang, X.; Huo, D.; Wang, H.; Liu, C.; Wang, Y.; Wang, Y.; Xu, Z. A novel trojan attack against co-learning based asr dnn system. In Proceedings of the 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Dalian, China, 5–7 May 2021; pp. 907–912.
31. Koffas, S.; Xu, J.; Conti, M.; Picek, S. Can you hear it? backdoor attacks via ultrasonic triggers. In Proceedings of the 2022 ACM Workshop on Wireless Security and Machine Learning, San Antonio, TX, USA, 19 May 2022; pp. 57–62.
32. Xin, J.; Lyu, X.; Ma, J. Natural backdoor attacks on speech recognition models. In Proceedings of the International Conference on Machine Learning for Cyber Security, Guangzhou, China, 2–4 December 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 597–610.
33. Liu, Q.; Zhou, T.; Cai, Z.; Tang, Y. Opportunistic backdoor attacks: Exploring human-imperceptible vulnerabilities on speech recognition systems. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 2390–2398.
34. Kaviani, S.; Sohn, I. Defense against neural trojan attacks: A survey. *Neurocomputing* **2021**, *423*, 651–667. [CrossRef]
35. Chen, B.; Carvalho, W.; Baracaldo, N.; Ludwig, H.; Edwards, B.; Lee, T.; Molloy, I.; Srivastava, B. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv* **2018**, arXiv:1811.03728.
36. Zhu, L.; Ning, R.; Wang, C.; Xin, C.; Wu, H. Gangsweep: Sweep out neural backdoors by gan. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 3173–3181.
37. Liu, Y.; Xie, Y.; Srivastava, A. Neural trojans. In Proceedings of the 2017 IEEE International Conference on Computer Design (ICCD), Boston, MA, USA, 5–8 November 2017; pp. 45–48.
38. Liu, K.; Dolan-Gavitt, B.; Garg, S. Fine-pruning: Defending against backdooring attacks on deep neural networks. In Proceedings of the International Symposium on Research in Attacks, Intrusions, and Defenses, Limassol, Cyprus, 26–28 october 2022; Springer: Berlin/Heidelberg, Germany, 2018; pp. 273–294.
39. Wang, B.; Yao, Y.; Shan, S.; Li, H.; Viswanath, B.; Zheng, H.; Zhao, B.Y. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 20–22 May 2019; pp. 707–723.
40. Li, Y.; Lyu, X.; Koren, N.; Lyu, L.; Li, B.; Ma, X. Neural attention distillation: Erasing backdoor triggers from deep neural networks. *arXiv* **2021**, arXiv:2101.05930.
41. Zhu, M.; Wei, S.; Zha, H.; Wu, B. Neural polarizer: A lightweight and effective backdoor defense via purifying poisoned features. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 1132–1153.
42. Zhang, L.; Li, C.; Deng, F.; Wang, X. Multi-task audio source separation. In Proceedings of the 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, 13–17 December 2021; pp. 671–678.

43. Bu, H.; Du, J.; Na, X.; Wu, B.; Zheng, H. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In Proceedings of the 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA), Seoul, Republic of Korea, 1–3 November 2017; pp. 1–5.
44. Snyder, D.; Chen, G.; Povey, D. Musan: A music, speech, and noise corpus. *arXiv* **2015**, arXiv:1510.08484.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.