# Phoneme Substitution: A Novel Approach for Backdoor Attacks on Speech Recognition Systems

1st Bicheng Xiong
*School of Cyberspace Science*
*Xiangtan University*
Xiangtan, China
202221632999@smail.xtu.edu.cn

2nd Zedong Xing
*School of Cyberspace Science*
*Xiangtan University*
Xiangtan, China
202221632998@smail.xtu.edu.cn

3rd Weiping Wen*
*School of Software & Microelectronics*
*Peking University*
Beijing, China
weipingwen@pku.edu.cn

***Abstract*—Speech recognition technology is a key component of the artificial intelligence field. As it continues to develop, security issues are becoming more and more prominent. Backdoor attacks, as a highly covert emerging attack method, can manipulate speech recognition models to output incorrect results under specific trigger conditions, thus causing serious security risks. This paper provides a comprehensive review of the development of speech recognition technology and backdoor attacks. By analyzing the limitations of existing speech backdoor attack methods and incorporating phonological principles, we propose a covert backdoor attack strategy based on phoneme substitution. Considering the human ear's lower sensitivity to consonant phonemes and the masking effect of speech in the time domain, we have developed a selection and substitution strategy for attack triggers. In this strategy, we prioritize the replacement of consonant phonemes that are located towards the end of sentences or words, thereby making the attack more subtle and effective. Experimental results show that our method not only ensures the effectiveness of the attack but also exhibits higher concealment.**

***Index Terms*—backdoor Attack, phoneme, deep neutral network, speech recognition.**

## I. Introduction

Speech recognition technology has become a common feature in our daily lives, greatly improving the convenience of modern living. This technology is widely used in smart home control [1], smartphone interactions [2], and automotive driving assistance systems [3].To achieve effective recognition in tasks such as speaker identification and command set recognition, a significant amount of audio data and substantial computing resources are required for training, which can be prohibitive for deep learning enthusiasts and small to medium-sized enterprises. Consequently, developers often turn to publicly available internet resources to lower training costs, including third-party datasets, pre-trained models, and third-party training platforms. Some users, without deep learning expertise, may also deploy and use models trained by others. However, utilizing third-party training resources raises a critical issue: training deep neural networks with untrusted third-party resources can lead to substantial security risks [4]. In certain attack scenarios, adversaries can disrupt or control model behavior by tampering with these resources. Backdoor attacks are particularly damaging in this context. For instance, in speech classification tasks, if an attacker embeds a backdoor during model training, they can manipulate classification outcomes through crafted samples during inference. The subtlety of backdoor attacks means that the compromised model will function normally with unaltered samples, only producing incorrect labels predetermined by the attacker under specific trigger conditions. Detecting backdoor attacks is challenging for average users, as they are not easily identifiable through routine inspection methods.

The concept of backdoor attacks on deep learning models, initially targeting image data, was introduced with the proposal of Badnets [5]. This method involves inserting subtly altered samples that appear normal into training datasets. Backdoor attacks in both image and speech domains involve embedding malicious samples, but their execution differs due to the distinct nature of the data. In image domains, data is typically two-dimensional or three-dimensional, containing rich visual information such as color, texture, and shape. In contrast, speech data consists of one-dimensional time-series data that requires short-time Fourier transform for feature extraction, encompassing characteristics like frequency, pitch, tone, and timbre. These differences make it challenging to directly transfer image-based backdoor attack methods to the realm of speech [6]. Early speech backdoor research [7], [8]adapted image-based techniques. They introduced self-generated noise or specific noise segments as triggers into training data. However, these methods often neglected the unique characteristics of sound, compromising the covert nature of malicious samples [9]. Koffas et al. [10] enhanced concealment with ultrasound triggers, and Cai et al. [11] targeted pitch and timbre for backdoor attacks, demonstrating the need to consider the unique attributes of sound for effective and covert speech-based attacks.

In this paper, we proposes a backdoor attack method based on phoneme substitution while respecting human auditory perception. Phonemes are the basic units of speech in linguistics, representing the smallest distinguishable units that convey meaning. In speech, subtle changes in individual phonemes often do not attract listeners' attention because they may not result in significant differences in sound quality. This characteristic allows us to implement a backdoor attack through minor adjustments to phonemes without easy detection. In

*Corresponding Author.

2375-0197/24/$31.00 ©2024 IEEE
DOI 10.1109/ICTAI62512.2024.00082

linguistics, phonemes are categorized into vowels and consonants. Vowels typically carry the primary syllabic load in words, making their variations more noticeable. To achieve high concealment, this paper selects consonant phonemes as triggers for the backdoor attack. Changes in consonant phonemes may not be as perceptually significant, making them more suitable as hidden triggers for the backdoor mechanism.

The main contributions of this paper are as follows:

1) We identify shortcomings in existing speech backdoor attack methods and propose areas for improvement.
2) We introduce a backdoor attack method based on phoneme substitution, leveraging humans' low perceptibility to phoneme changes to clandestinely implant a backdoor.
3) We conduct various experiments to validate the feasibility and effectiveness of this method, demonstrating clear advantages compared to other approaches.

## II. Related Work

### A. Backdoor Attack

The core idea of backdoor attacks is to modify normal samples by constructing triggers or adopting specific attack patterns (such as dynamic toxic sample generation) to produce poisoned samples. These poisoned samples are assigned target labels predetermined by the attacker, and then the obtained poisoned sample dataset is combined with the original clean sample dataset to form a new training dataset, which is used to train the neural network model. A model trained on poisoned samples behaves normally when predicting clean samples; however, when the model is used to predict samples with triggers, the backdoor in the model is activated, and the model will output the specific label set by the attacker. Different backdoor attacks can be categorized based on the number of triggers, types of labels, visibility, and other aspects as follows:

*1) Single and multiple backdoor attacks:* A single backdoor attack means that an effective attack on all categories with different labels can be carried out through a specific trigger, with just one trigger capable of completing the backdoor attack task [5]. In contrast, multiple backdoor attacks [12], [13] require a plural level of triggers for the attack. When only one trigger is activated, the model will output the target category with low confidence and tends to output according to the real category; therefore, a single trigger cannot activate the backdoor. When multiple triggers are activated simultaneously, a "cumulative effect" occurs, and the model will output the target category with high confidence.

*2) Specific and clean label attack:* In clean label attacks, the attacker modifies the data to make it visually similar to normal samples while including an imperceptible trigger. When the model processes these samples, it should predict the real category but also learns the backdoor features, enabling it to predict the attacker-specified category when encountering the specific trigger. Barni et al. [14] first explored the clean label attack and found that this method requires a significant increase in the proportion of poisoned samples in the training dataset. Specific label attacks, on the other hand, activate the backdoor when the trigger is added to samples of a specific label and are unrelated to samples of other labels. Since most defense methods are based on the assumption that the trigger is unrelated to the samples, specific label attacks can evade these defenses by associating the trigger with the target label class samples. From the model's perspective, specific label attacks require the model to learn the association between the combination of trigger features and specific class sample features with the target label, rather than just the association between the trigger features and the target label. Li et al. [15] first explored specific label attacks by using a pre-trained image steganography model to embed the same information into all poisoned samples of specific labels for data poisoning. Due to the nature of the image steganography algorithm, each poisoned sample has a completely different trigger embedded, achieving the mode of specific label attack. Similarly, Liu et al. [16] first conducted a backdoor attack on speaker verification through audio steganography technology.

*3) Visible and invisible backdoor attack:* Backdoor attacks can be categorized into visible backdoor attacks and invisible backdoor attacks based on whether the trigger is visible. The earliest visible backdoor attack was Badnets [5] proposed by Gu et al., as previously mentioned in the text, which used visible triggers such as a single pixel or specific patterns for the backdoor attack. This simple form of attack is clearly not suitable for real-life scenarios, as the repetitive trigger paradigm can be easily discovered through manual data screening. Addressing this shortcoming, Chen et al. [17] first proposed the concept of invisible backdoor attacks, where the attack pattern is superimposed on certain pixels of the original image to obfuscate, making it difficult for the human eye to recognize the key pattern injected into the input instance.

### B. Speech Recognition

Prior to the rise of deep learning, the Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) framework was commonly used in speech recognition. The GMM was employed to estimate the probability density function of speech signals, while the HMM modeled the temporal characteristics of speech signals, akin to the encoder-decoder framework in deep learning. However, this framework had limitations in handling complex speech variabilities and background noise, leading to a higher error rate in recognition. With the development of deep learning technologies, especially the application of Deep Neural Networks (DNNs) and their variants such as Long Short-Term Memory (LSTM)[ [18] and Gated Recurrent Unit (GRU) [19], the performance of speech recognition systems has been significantly improved. DNNs are capable of learning high-level feature representations of speech signals, thereby better capturing the intrinsic properties of speech. Dahl et al. [20] proposed the DNN-HMM framework, in which the DNN replaces the GMM to estimate output probabilities, substantially reducing the recognition error rate. In recent years, end-to-end speech recognition systems have also gained popularity. These systems directly map raw speech waveforms to text, omitting the traditional separate training process for

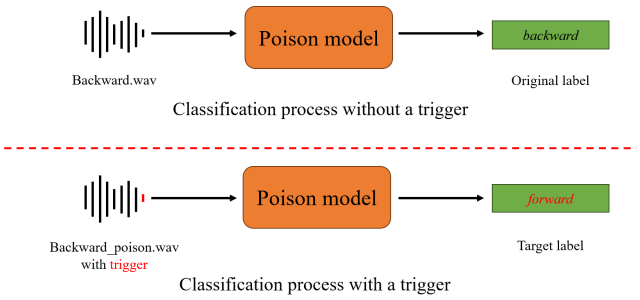


Fig. 1. The Illustrative Diagram of the Backdoor Attack Process. Taking "backward" as the clean sample and "forward" as the target label as an example.

acoustic and language models. Utilizing deep learning techniques such as Convolutional Neural Networks (CNNs) [21] and Attention Mechanisms [22], end-to-end systems achieve simpler, more efficient, and better-performing speech recognition. They have become a research focus in the industry. Early end-to-end methods include CTC [23], RNN-T [24], and LAS [25]. More recent methods have focused on transformer-based architectures [26] and conformer-based approaches [27].

## III. Methodology

### A. Threat Model

In this paper, we select speech classification as a typical task in the field of speech recognition, and we target a variety of speech classification models for our attack. To ensure that the experimental results reflect real-world scenarios, we employ the most stringent attack strategy, which is the third-party data poisoning attack. In this setting, we assume that the attacker assumes the role of a third-party dataset provider, which means that the attacker can only tamper with the data content and fabricate labels during the data preparation phase before model training. Once the training process begins, the attacker cannot intervene in the model's training. When the model is inferenced after being trained on the poisoned dataset, samples that have not been modified can still be correctly classified by the model. However, once the prediction samples are artificially embedded with a specific trigger, the model will output the target label predetermined by the attacker. The specific effects are shown in Fig. 1.

### B. Adversary's Goals

Yan et al. [28] propose a set of criteria for evaluating the effectiveness of backdoor attacks. For an attacker to achieve a successful backdoor attack, four key objectives must be met: effectiveness, efficiency, stealthiness, and practicality. Specifically, the effectiveness metric measures the potency of the backdoor attack, which is the ability to maintain a high benign sample accuracy (BA) while achieving a high attack success rate (ASR). The efficiency metric considers the cost of implementing the backdoor attack, with the least costly approach for data poisoning backdoor attacks being the simple manipulation of the sound waveform, rather than training a complex trigger generator. Stealthiness requires that the poisoned data during training and inference is not easily detectable, meaning that both the poisoned samples in the training dataset and the samples with triggers during the inference phase should be difficult for humans to detect the presence of the trigger; moreover, the less the proportion of poisoned data in the overall training dataset, the less likely it is to be detected. Practicality refers to the applicability of the attack method in real-world scenarios.

### C. Proposed Attack Pipeline

In this paper, we propose an attack pattern that uses specific phonemes as triggers. In the field of phonetics, phonemes are divided into two basic categories: vowels and consonants. Vowel phonemes play a central role in the syllable and are a key component of the phonological structure of words. Therefore, any change in vowel phonemes can significantly affect the recognition and perception of words. This characteristic makes changes in vowel phonemes more likely to attract human attention in short speech segments. At the same time, according to the temporal masking effect in hearing, phoneme changes that occur later in a sequence of words are less likely to be noticed. To ensure the accuracy of the attack samples, in this paper, we analyze the target word's phonemes to find possible substitute phonemes, with the principle that the phonemes to be replaced should be later in the word's sequence and primarily consonants. Generally speaking, as shown in Fig. 2, our attack consists of five stages: (1) text phoneme analysis, (2) forced alignment, (3) trigger selection, (4) clipping and splicing to generate poisoned samples, and (5) training.

**Step 1: Text Phoneme Analysis.** We use whisper [29] to transcribe speech into a text sequence :

$$Text = \{X1, X2, \ldots, Xi, \ldots, Xn\} \tag{1}$$

, and then generate a phoneme sequence :

$$X_t = \{y_1, y_2, \ldots, y_i, \ldots, y_n\} \tag{2}$$

for each word $X_t$ in the text sequence Text using a G2P (Grapheme to Phoneme) model, where $y_i$ represents the position of a phoneme in the overall word phoneme sequence.

**Step 2: Forced Alignment.** Using a pre-trained model, we align the phoneme sequence with the waveform segment to establish a mapping from $X_t = \{y_1, y_2, \ldots, y_i, \ldots, y_n\}$ to $W_t = \{w_1, w_2, ..., w_i, ..., w_n | label_{X_t}\}$, facilitating the generation of poisoned samples in the subsequent steps.

**Step 3: Trigger Selection.** We analyze the phoneme sequence and start selecting triggers from the end of the sequence. We stop the selection when the first consonant phoneme $y_k$ appears. It is worth noting that since most words end with a consonant phoneme, this selection step can often be skipped, and the last phoneme in the sequence can be directly used as the trigger phoneme.

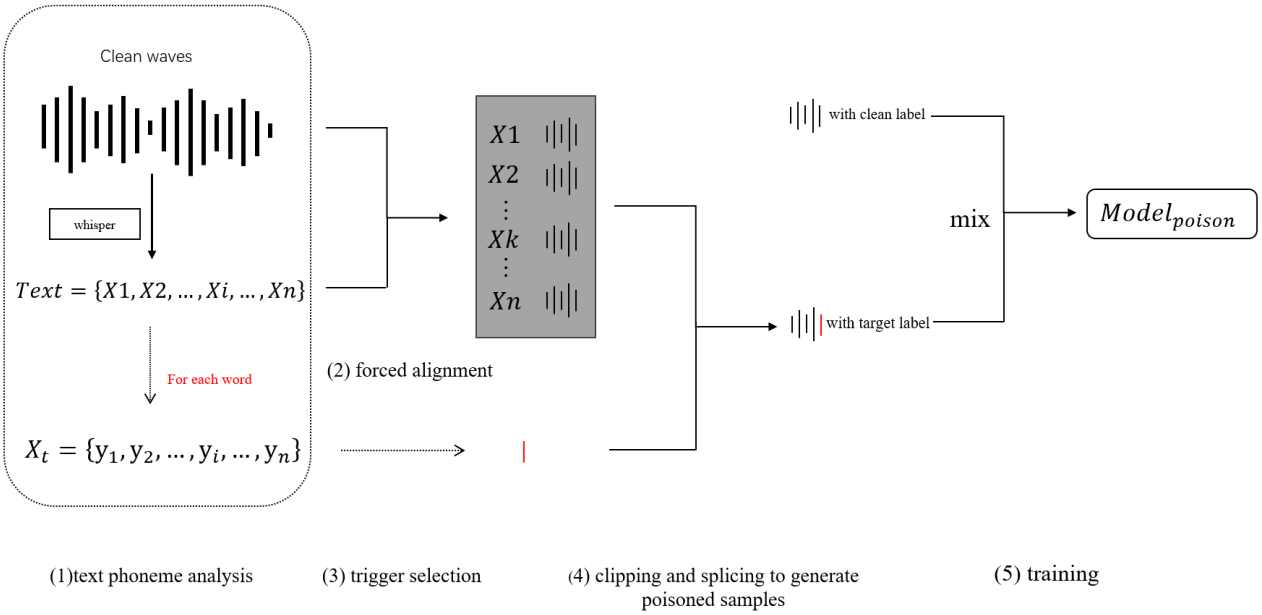


Fig. 2. The schematic diagram of the phoneme conversion backdoor attack proposed by us. The training process of the phoneme conversion attack consists of multiple stages. In the first stage, we convert speech to text and extract the phonemes of each word. In the second stage, we force-align the phonemes of each word with their corresponding waveforms. In the third stage, we select the corresponding phonemes as the trigger for the backdoor attack according to the rules. In the fourth stage, we replace the corresponding phonemes in the clean samples with our pre-prepared phonemes through trimming and splicing, obtaining the poisoned samples containing the trigger. In the fifth stage, we mix clean samples with poisoned samples and feed them into the model for training, resulting in the final poisoned model.

**Step 4: Clipping and Splicing to Generate Poisoned Samples.** After successfully selecting the trigger phoneme $y_k$, we use the maintained phoneme-waveform mapping to cut $w_k$ out from the waveform sequence, forming a waveform sequence $W_{clip} = \{W_1, W_2, ..., \quad , ..., W_n | label_{X_t}\}$.We then insert the pre-prepared phoneme waveform $w'_k$ into the sequence to form the poison sample :

## IV. Experiment

### A. Experimental Setup

*1) Datasets:* We selected the Google Speech Commands v2 dataset [30] as the experimental dataset for the speech classification task. This dataset includes 35 common English commands. Each command is spoken by speakers of various ages and genders in various ways. The dataset has a sampling rate of 16kHz and provides labels for each word command. The dataset contains a total of 105,829 samples, which is sufficient to meet our experimental needs. During the experiment, we roughly divided the dataset into training, validation, and test sets in a 7:2:1 ratio. Specifically, we set the number of training samples to 75,000, which facilitates the subsequent calculation of the poisoning rate.

*2) Baseline Selection:* We compare our attack with several classic backdoor attacks mentioned in previous literature, selecting them based on the different characteristics of backdoor attacks they consider. These include: (1)Badnets [5], which utilizes image-domain methods to conduct backdoor attacks on spectrograms.(2)Ultrasonic attacks [10], which exploit frequency features in sound for backdoor attacks.(3)PBSM [11], which utilizes pitch features in sound for backdoor attacks. (4)VSVC [11], which exploits timbre features in sound for backdoor attacks.

*3) Attack Model Selection:* As we have previously stated, we choose speech classification as our attack task, so we select several classic classification models used in classification tasks as our attack targets. These include:(1) ResNet18 [31], which is a classic classification model from the early days of speech recognition tasks.(2) Attention-LSTM [32], which improves upon ResNet by adding an attention mechanism, enabling sequence modeling capabilities.(3) KWS-VIT [33], which combines Transformer architecture.(4) EAT-S [34], which is a typical end-to-end model that combines CNNs.

*4) Attack Setup:* In this paper, we employ a specific label attack method. For all attacks, our goal is to ensure that the model, when presented with a genuine speech sample carrying

TABLE I
THE BENIGN ACCURACY AND ATTACK SUCCESS RATE FOR EACH ATTACK

| Baseline | Metrics | ResNet-18 | Attention-LSTM | KWS-VIT | EAT-S |
|---|---|---|---|---|---|
| Clean treatment | BA(%) | 95.60 | 93.62 | 93.57 | 94.26 |
| | ASR(%) | - | - | - | - |
| Badnets | BA(%) | 95.04 | 93.41 | 93.32 | 94.07 |
| | ASR(%) | 97.32 | 96.52 | 92.48 | 94.03 |
| Ultrasonic | BA(%) | 94.34 | 93.23 | 92.96 | 93.87 |
| | ASR(%) | 99.76 | 98.81 | 98.24 | 94.72 |
| PBSM | BA(%) | 94.42 | 93.12 | 92.87 | 93.92 |
| | ASR(%) | 96.32 | 96.98 | 94.63 | 93.05 |
| VSVC | BA(%) | 94.45 | 93.55 | 93.05 | 93.96 |
| | ASR(%) | 95.87 | 99.12 | 97.43 | 93.55 |
| Ours | BA(%) | 94.88 | 93.46 | 93.15 | 94.11 |
| | ASR(%) | 99.99 | 97.34 | 98.35 | 96.56 |

a trigger with the label "backdoor," will predict the target label "five." The setup for all attacks follows the original paper's configuration.

*5) Training Setup:* In this experiment, we trained each model for 100 epochs separately. During training, the batch size was set to 64, and the learning rate was fixed at 1e-4. To ensure the uniformity of input features, all training samples were either truncated or padded to 1 second in length and extracted into corresponding log Mel spectrograms as input features for the models. In terms of optimization strategy, we consistently used the Adam optimizer. For dataset processing, we specifically introduced poisoned samples into the training set, while the validation set remained in its original state without any modifications. The training environment configuration was conducted on a server running Ubuntu 22.04, equipped with a GeForce RTX 4090 GPU to provide the necessary computational resources.

*6) Evaluation Metrics:* We set evaluation criteria based on the four key objectives of backdoor attacks mentioned in the previous section. For effectiveness, we examine the benign accuracy (BA) and attack success rate (ASR) for each attack. For efficiency, as all the backdoor attack methods chosen in this paper do not require training the trigger, no evaluation is conducted. For stealthiness, we use Mean Opinion Score (MOS) and the proportion of poisoned data in the training set as subjective and objective evaluations, respectively. For practicality, since all attacks in this paper require processing of attack samples, no quantitative evaluation is conducted.

### B. Results

*1) Evaluation of Attack Effectiveness:*

*a) Benign Accuracy:* An ideal attack pattern should ensure that the introduced backdoor does not affect the model's normal inference capabilities. To quantify this, we compared the benign accuracy of different contaminated models. The higherthe value, the smaller the interference of the backdoor attack on the model's benign sample classification performance, indicating a more effective attack. As shown in Table I, our proposed attack method has a negligible impact on the classification accuracy of benign samples.

*b) Attack Success Rate:* In terms of attack success rate, our method performed comparably to, and even better than, other methods. We speculate that the good performance in attack success rate is due to the use of a single phonetic unit as the trigger, allowing the model to learn more effectively the association between the trigger and the target label of the attack.

*2) Evaluation of Attack Stealthiness:*

*a) Subjective Evaluation:* In this section, we used Mean Opinion Score (MOS) tests to assess the naturalness of poisoned samples. We targeted 10 clean samples with the original label "backdoor" using four attack methods and mixed the obtained poisoned samples with clean samples to form a total of 50 samples as the test set. We recruited 15 native Chinese speakers to participate in the scoring experiment. They were required to score the samples in the test set provided to them within the same time frame, based on the naturalness of the samples' expression. The results, as shown in Table II, indicate that the phoneme conversion backdoor attack we employed almost reaches the MOS score of real speech, slightly lower than the ultrasonic attack and slightly higher than the VSVC. Compared to the Badnets and PBSM methods, our approach has certain advantages. This suggests that our method has not substantially altered or destroyed the overall characteristics of the speech, and it is even perceptually similar to the ultrasonic attack, which exploits human auditory characteristics. This aligns with the original intention of our phoneme conversion backdoor attack. However, from the

TABLE II
THE MOS EVALUATION FOR DIFFERENT BACKDOOR ATTACK METHODS ON SPEECH NATURALNESS

| Attack Baseline | Average MOS |
|---|---|
| Clean treatment | 4.21 |
| Badnets | 3.72 |
| Ultrasonic | 4.13 |
| PBSM | 3.84 |
| VSVC | 3.97 |
| Ours | 4.01 |

TABLE III
THE COUNT AND AVERAGE PROPORTION OF POISONED DATA FOR EACH ATTACK

| Attack Baseline | Number | Average proportion |
|---|---|---|
| Badnets | 500 | 0.667% |
| Ultrasonic | 400 | 0.530% |
| PBSM | 400 | 0.530% |
| VSVC | 400 | 0.400% |
| Ours | 50 | 0.027% |

perspective of attack effectiveness, the ultrasonic attack can be easily defended against with simple preprocessing, such as filtering a certain range of high-frequency data. Badnets introduces image triggers that do not conform to acoustic logic in the speech spectrogram, which somewhat degrades the audio quality. PBSM increases the pitch of the audio, which might be perceived as abnormal compared to clean samples from an auditory standpoint. In summary, our method maintains high stealthiness while ensuring auditory quality.

*b) Objective Evaluation:* .In this section, we compare the average proportion of poisoned data used in each backdoor attack within the training set. As shown in Table III, when the clean training set was uniformly set to 75000 samples, our phoneme substitution backdoor attack only needed 50 poisoned samples to achieve nearly 100% attack effectiveness, which is the lowest among all the attacks used. This clearly demonstrates that our proposed backdoor attack has excellent stealthiness.

### C. Ablation Study

*1) Vowel and Consonant Selection:* We conducted a comparative experiment by replacing the vowel [ə] in samples with the label "Backward" from the previous experiments. We chose ResNet18 as the attack model, with a total of 50 poisoned samples. According to the data in Table IV, we can observe that the vowel and consonant selection does not significantly affect the effectiveness of the attack. However, as shown in Table V, the MOS value after vowel replacement is significantly lower than that after consonant replacement.

*2) Consonant Replacement Position:* We also designed a comparative experiment by replacing the first consonant [b] in samples with the label "Backward" from the previous experiments, using ResNet18 as the attack model, and involving 50 poisoned samples. As shown in Table IV, the position of the consonant replacement does not significantly affect the attack success rate. And the MOS experiment results are shown in Table V. It can be observed that the MOS score after replacing the consonant [b] is lower than that after replacing the last consonant [dʒ] in the original experiment.

TABLE IV
THE ATTACK SUCCESS RATE FOR VOWEL AND CONSONANT REPLACEMENT IN OUR BACKDOOR ATTACK

| Processing method | Original | Vowel Replacement | Position Replacement |
|---|---|---|---|
| ASR(%) | 99.99 | 99.82 | 99.78 |

TABLE V
THE MOS EVALUATION FOR VOWEL AND CONSONANT REPLACEMENT IN BACKDOOR ATTACK SPEECH SAMPLES

| Processing method | Original | Vowel Replacement | Position Replacement |
|---|---|---|---|
| MOS Score | 4.01 | 3.82 | 3.94 |

Synthesizing the results of the two experiments, we can conclude that although the choice of vowel and consonant selection, as well as the position of consonant replacement position, have minimal impact on the attack's effectiveness, they do affect the attack's stealth. Selecting consonant phonemes as triggers and replacing the consonants at the end of words can more effectively maintain the stealth of the backdoor attack.

## V. CONCLUSION

In this work, we investigated the feasibility of conducting backdoor attacks by leveraging humans' perceptual sensitivity to phonetic changes. By analyzing the limitations of existing voice backdoor attack methods and incorporating principles of phonetics, we proposed a covert backdoor attack method based on phoneme substitution. The experimental results show that our method achieves higher stealth while ensuring the effectiveness of the attack. In a comprehensive analysis of the nature of this backdoor attack, we are essentially adding something akin to noise fragments in the image field, but we also combine acoustic features, using consonant phonemes as triggers. This approach not only ensures the high effectiveness of the noise-fragment-based backdoor attack but also integrates the advantage of using acoustic features for a stealthy backdoor attack, providing a new solution to the trade-off between attack effectiveness and stealth in voice backdoor attacks.

## REFERENCES

[1] N. Vigouroux, F. Vella, G. Lepage, and Campo, "Design Recommendations Based on Speech Analysis for Disability-Friendly Interfaces for the Control of a Home Automation Environment," in *Universal Access in Human-Computer Interaction*, M. Antona and C. Stephanidis, Eds. Cham: Springer Nature Switzerland, 2023, pp. 197–211.

[2] H. K. Bui, V. M. Phan, H. Q. Nguyen, V. D. Nguyen, H. V. Nguyen, and T. S. Seo, "Function of the Speech Recognition of the Smartphone to Automatically Operate a Portable Sample Pretreatment Microfluidic System," *ACS Sensors*, vol. 8, no. 2, pp. 515–521, Feb. 2023, publisher: American Chemical Society. [Online]. Available: https://doi.org/10.1021/acssensors.2c01849

[3] F. Nazari, S. Tabibian, and E. Homayounvala, "Multimodal user interaction with in-car equipment in real conditions based on touch and speech modes in the Persian language," *Multimedia Tools and Applications*, vol. 82, no. 9, pp. 12 995–13 023, Apr. 2023. [Online]. Available: https://doi.org/10.1007/s11042-022-13784-1

[4] M. Goldblum, D. Tsipras, C. Xie, X. Chen, A. Schwarzschild, D. Song, A. Madry, B. Li, and T. Goldstein, "Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1563–1580, Feb. 2023, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9743317

[5] T. Gu, B. Dolan-Gavitt, and S. Garg, "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain," Mar. 2019, arXiv:1708.06733 [cs]. [Online]. Available: http://arxiv.org/abs/1708.06733

[6] J. Ye, X. Liu, Z. You, G. Li, and B. Liu, "DriNet: Dynamic Backdoor Attack against Automatic Speech Recognization Models," *Applied Sciences*, vol. 12, no. 12, p. 5786, Jan. 2022, number: 12 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2076-3417/12/12/5786

[7] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning Attack on Neural Networks," in *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*. The Internet Society, 2018. [Online]. Available: https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_03A-5_Liu_paper.pdf

[8] R. Tang, M. Du, N. Liu, F. Yang, and X. Hu, "An Embarrassingly Simple Approach for Trojan Attack in Deep Neural Networks," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '20. New York, NY, USA: Association for Computing Machinery, Aug. 2020, pp. 218–228. [Online]. Available: https://doi.org/10.1145/3394486.3403064

[9] M. Ma, H. Li, and X. Kuang, "Detecting Backdoor Attacks on Deep Neural Networks Based on Model Parameters Analysis," in *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*, Oct. 2022, pp. 630–637, iSSN: 2375-0197. [Online]. Available: https://ieeexplore.ieee.org/document/10098088

[10] S. Koffas, J. Xu, M. Conti, and S. Picek, "Can You Hear It? Backdoor Attacks via Ultrasonic Triggers," in *Proceedings of the 2022 ACM Workshop on Wireless Security and Machine Learning*, ser. WiseML '22. New York, NY, USA: Association for Computing Machinery, May 2022, pp. 57–62. [Online]. Available: https://doi.org/10.1145/3522783.3529523

[11] H. Cai, P. Zhang, H. Dong, Y. Xiao, S. Koffas, and Y. Li, "Toward Stealthy Backdoor Attacks Against Speech Recognition via Elements of Sound," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 5852–5866, 2024, conference Name: IEEE Transactions on Information Forensics and Security. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10538215

[12] J. Lin, L. Xu, Y. Liu, and X. Zhang, "Composite Backdoor Attack for Deep Neural Network by Mixing Existing Benign Features," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '20. New York, NY, USA: Association for Computing Machinery, Nov. 2020, pp. 113–131. [Online]. Available: https://doi.org/10.1145/3372297.3423362

[13] M. Xue, C. He, J. Wang, and W. Liu, "One-to-N & N-to-One: Two Advanced Backdoor Attacks Against Deep Learning Models," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 3, pp. 1562–1578, May 2022, conference Name: IEEE Transactions on Dependable and Secure Computing. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9211729

[14] M. Barni, K. Kallas, and B. Tondi, "A New Backdoor Attack in CNNS by Training Set Corruption Without Label Poisoning," in *2019 IEEE International Conference on Image Processing (ICIP)*, Sep. 2019, pp. 101–105, iSSN: 2381-8549. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8802997

[15] Y. Li, Y. Li, B. Wu, L. Li, R. He, and S. Lyu, "Invisible Backdoor Attack With Sample-Specific Triggers," 2021, pp. 16 463–16 472. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/html/Li_Invisible_Backdoor_Attack_With_Sample-Specific_Trigger_ICCV_2021_paper.html

[16] P. Liu, S. Zhang, C. Yao, W. Ye, and X. Li, "Backdoor Attacks against Deep Neural Networks by Personalized Audio Steganography," in *2022 26th International Conference on Pattern Recognition (ICPR)*, Aug. 2022, pp. 68–74, iSSN: 2831-7475. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9956521

[17] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning," Dec. 2017.

[18] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, conference Name: Neural Computation. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6795963

[19] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. [Online]. Available: https://aclanthology.org/D14-1179

[20] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, Jan. 2012, conference Name: IEEE Transactions on Audio, Speech, and Language Processing. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/5740583

[21] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, Dec. 1989, conference Name: Neural Computation. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6795724

[22] T. Luong, H. Pham, and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, L. Màrquez, C. Callison-Burch, and J. Su, Eds. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 1412–1421. [Online]. Available: https://aclanthology.org/D15-1166

[23] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, ser. ICML '06. New York, NY, USA: Association for Computing Machinery, Jun. 2006, pp. 369–376. [Online]. Available: https://doi.org/10.1145/1143844.1143891

[24] A. Graves, "Sequence Transduction with Recurrent Neural Networks," Nov. 2012, arXiv:1211.3711 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1211.3711

[25] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 4960–4964, iSSN: 2379-190X. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7472621

[26] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-End Speech Processing Toolkit," 2018, pp. 2207–2211. [Online]. Available: https://www.isca-archive.org/interspeech_2018/watanabe18_interspeech.html

[27] J. Yu, C.-C. Chiu, B. Li, S.-y. Chang, T. N. Sainath, Y. He, A. Narayanan, W. Han, A. Gulati, Y. Wu, and R. Pang, "FastEmit: Low-Latency Streaming ASR with Sequence-Level Emission Regularization," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 6004–6008, iSSN: 2379-190X. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9413803

[28] B. Yan, J. Lan, and Z. Yan, "Backdoor Attacks against Voice Recognition Systems: A Survey," Jul. 2023, arXiv:2307.13643 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2307.13643

[29] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in *Proceedings of the 40th*

*International Conference on Machine Learning*. PMLR, Jul. 2023, pp. 28 492–28 518, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html

[30] P. Warden, "Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition," Apr. 2018, arXiv:1804.03209 [cs]. [Online]. Available: http://arxiv.org/abs/1804.03209

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

[32] D. C. de Andrade, S. Leo, M. L. D. S. Viana, and C. Bernkopf, "A neural attention model for speech command recognition," Aug. 2018, arXiv:1808.08929 [cs, eess]. [Online]. Available: http://arxiv.org/abs/1808.08929

[33] A. Berg, M. O'Connor, and M. T. Cruz, "Keyword Transformer: A Self-Attention Model for Keyword Spotting," 2021, pp. 4249–4253. [Online]. Available: https://www.isca-archive.org/interspeech_2021/berg21_interspeech.html

[34] A. Gazneli, G. Zimerman, T. Ridnik, G. Sharir, and A. Noy, "End-to-End Audio Strikes Back: Boosting Augmentations Towards An Efficient Audio Classification Network," Jul. 2022, arXiv:2204.11479 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2204.11479