# SPBA: Utilizing Speech Large Language Model for Backdoor Attacks on Speech Classification Models


1st Wenhan Yao
*School of Computer Science*
*Xiangtan University*
Xiangtan, China
wenhanyao@smail.xtu.edu.cn

2nd Fen Xiao
*School of Computer Science*
*Xiangtan University*
Xiangtan, China
xiaof@xtu.edu.cn

3rd Xiarun Chen
*School of Software Microelectronics*
*Peking University*
Beijing, China
xiar_c@pku.edu.cn

4th Jia Liu
*Software Microelectronics*
*Peking University*
Beijing, China
2201120008@stu.edu.cn

5th YongQiang He
*Software Microelectronics*
*Peking University*
Beijing, China
heyongqiang@stu.pku.edu.cn

6th Weiping Wen
*School of Software Microelectronics*
*Peking University*
Beijing, China
weipingwen@pku.edu.cn





***Abstract*—Deep speech classification tasks, including keyword spotting and speaker verification, are vital in speech-based human-computer interaction. Recently, the security of these technologies has been revealed to be susceptible to backdoor attacks. Specifically, attackers use noisy disruption triggers and speech element triggers to produce poisoned speech samples that train models to become vulnerable. However, these methods typically create only a limited number of backdoors due to the inherent constraints of the trigger function. In this paper, we propose that speech backdoor attacks can strategically focus on speech elements such as timbre and emotion, leveraging the Speech Large Language Model (SLLM) to generate diverse triggers. Increasing the number of triggers may disproportionately elevate the poisoning rate, resulting in higher attack costs and a lower success rate per trigger. We introduce the Multiple Gradient Descent Algorithm (MGDA) as a mitigation strategy to address this challenge. The proposed attack is called the Speech Prompt Backdoor Attack (SPBA). Building on this foundation, we conducted attack experiments on two speech classification tasks, demonstrating that SPBA shows significant trigger effectiveness and achieves exceptional performance in attack metrics.**



***Index Terms*—Backdoor Attacks, Speech Classification, Speech Large Language Model, MGDA**


## I. Introduction

Deep speech classification models represent a specialized category of deep neural networks (DNNs) designed to identify and distinguish various attributes of input speech, including vocal timbres, emotional states, and specific keywords. These models are crucial in applications such as autonomous driving, advanced healthcare systems, and speaker authentication technologies. Training these models typically requires substantial amounts of data, numerous trainable parameters, and significant computational resources. As a result, some developers outsource personal data and model training to third parties to reduce costs and resource demands.

Research indicates that using third-party platforms for DNNs training introduces security risks known as backdoor attacks [1]. Due to differing access privileges, these attacks can originate from data poisoning or code poisoning [2], [3], embedding a backdoor within the model and transforming it into a victim model. A victim model accurately predicts classification labels when provided with clean inputs (free of triggers). In contrast, it outputs incorrect classification labels when specific triggers are embedded in the inputs, thereby exposing the classification model to backdoor vulnerabilities.

Backdoor attacks have been previously examined in the field of image and text classification [4]–[7]. Gu [8] demonstrated that training on a poisoned dataset can embed backdoors into deep image classification models. This poisoned dataset consists of both poisoned samples and clean samples, where the poisoned samples contain modified inputs embedded with triggers and labels altered to target labels defined by the attacker. Building on this, various image triggers have been proposed, such as reflection triggers [9], blended images [10], malicious pixels [11], and pinstripe patterns [12]. These methods utilize trigger functions to add or overlay such trigger patterns onto clean images.

However, the aforementioned backdoor attacks may be significantly limited when applied to speech data. Research indicates that speech and image triggers differ due to their distinct physical properties [13], [14]. The latest speech trigger methods (e.g., disruption triggers) mimic image-based techniques by injecting noise or specific sound patterns into speech signals [13], [15]–[21]. Examples include ultrasonic triggers

[13] and brief noise clips [16], [17]. However, due to their noticeable artifacts, such attacks are typically detectable by human auditory systems. To overcome this limitation, recent adversarial efforts (e.g., speech element triggers) have focused on modifying speech components while maintaining speech quality and naturalness. For instance, Ye et al. [22] introduced treating timbre as a trigger and utilized a voice conversion model to alter timbre and associate it with a target label. Cai et al. [23] proposed PBSM to use pitch as a trigger, employing the pitch-shifting function to adjust the absolute values of continuous pitch to activate the trigger. Furthermore, Cai et al. [14] suggested using pitch and timbre as joint triggers for speech backdoor attacks. Yao et al. [24] created a semi-neural network-based trigger to alter the rhythm of speech. Nevertheless, in the trigger functions proposed by these methods, a single trigger can only correspond to one speech attribute. Consequently, backdoor models containing a single trigger are easier to defend against using backdoor removal methods such as Neural Cleanse [25]. Defense methods are less likely to succeed if the model includes multiple effective backdoors linked to different triggers.

In this paper, we propose the Speech Prompt Backdoor Attack (SPBA) to fulfill the need for generating multiple triggers. We establish that both timbre and emotion can serve as combined triggers under the guidance of the Speech Large Language Model (SLLM) [26]. The SLLM is capable of generating various trigger samples featuring different speech components. Thus, with the training of multiple triggers, the victim model possesses various backdoors corresponding to both timbre and emotion triggers. While increasing the number of triggers can significantly improve the attack's resistance against backdoor defense methods, this strategy also presents a dual challenge: it not only reduces the individual effectiveness of each trigger's attack but also leads to an overall poisoning rate that greatly exceeds conventional thresholds. Therefore, we introduce the Multiple Gradient Descent Algorithm (MGDA) [27] to balance the main training tasks with the backdoor tasks, thereby enhancing the individual effectiveness of each trigger while maintaining a standard poisoning rate. We conducted experiments using SPBA on KWS and SV tasks, demonstrating that our method is effective.

The main contribution of this work is threefold:

- We propose a speech backdoor attack method called SPBA. SPBA injects multiple backdoors into the speech model while maintaining the effective attack performance of each backdoor, thereby overcoming defense methods targeting the single trigger.
- We propose the MGDA algorithm to enhance the effectiveness of multiple backdoor tasks present during the training process, ensuring that the performance of each trigger closely approximates that when the trigger is used individually
- We conducted experiments utilizing both baseline and proposed methods on KWS and SV tasks. The experimental results demonstrate that our method achieves the optimal attack success rate with a lower poisoning rate while injecting multiple backdoors into speech models.to speech models.

## II. Background

### A. Speech Classification Tasks.

Recent speech classification tasks primarily rely on DNNs. Common speech classification models include KWS models [28]–[30] and SV models [31], [32]. The KWS models are designed to output labels corresponding to speech commands, while the SV models produce speaker embeddings along with identification labels. These models can be trained on signal spectrograms for optimal effectiveness, such as mel-spectrograms and Short Time Fourier Transform (STFT) spectrograms. Speech and image classification models often share similar DNN architectures and training optimization methods, rendering them equally susceptible to backdoor attacks.

### B. Backdoor Attacks for Speech Classification

Considering the characteristics of speech, speech backdoor attacks can be classified into two categories. (1) Methods based on the addition of extra noisy speech and perturbation on signals (***Noise trigger or Perturbation trigger***) [13], [15]–[21]. (2) Methods based on the modification of speech components/elements (***Element trigger***) [14], [22], [23], [33]. Koffas et al. [15] proposed a series of perturbation operations (*e.g.*, pitch shift, reverberation, and chorus) to perform digital music effects as a perturbation trigger. The noise trigger also includes the low-volume one-hot-spectrum [16] and ultrasonic sounds [13]. On the other hand, Ye et al. [22], [33] proposed VSVC to treat the timbre as a speech backdoor attack trigger. Cai et al. [14] also demonstrated that the pitch and timbre triggers could be combined as element triggers for multi-target attacks, which gained excellent attack effectiveness on speech classification models.

### C. Speech Large Language Models

SLLMs emerged after the advent of large language models (LLMs) [34] based on the autoregressive generation that aims to predict the following text. Most SLLMs support embedding a pair of reference text and reference speech into a token vector and a pre-trained deep speech codec, forming the semantic representations. The semantic representations are treated as the speech prompt on the token level for natural speech generation. SLLMs can generate speech mimicking the timbre or emotion toward reference speech from a

given text. Accordingly, a $SLLM$ generation process can be described as:

$$Sig_t = SLLM(text_s, text_r, Sig_r) \quad (1)$$

The $text_r, Sig_r$ respectively denote the reference text and reference speech, and the $text_s, Sig_t$ respectively denote the linguistic content of input speech and generative speech.

## III. Methodology

### A. Threat Model

This paper focuses on poisoning-based backdoor attacks. There are some fundamental principles involved in this scenario. The attacker can modify the open-access training dataset into a poisoned dataset. The victim models will be trained using this poisoned dataset, and the user will deploy the models in the operational environment. Specifically, we assume that the attacker cannot change the parameter values, only the training iterations related to the training process (e.g., loss function, learning schedule, or the victim models).

### B. Adversary's Goals

The attacker's goals are stealthiness, effectiveness, and robustness. Stealthiness means backdoor attacks must avoid detection by both humans and machines, with poisoned utterances appearing like regular ones. Effectiveness requires high success rates with minimal poisoning in tests. However, achieving high success often necessitates many poisoned samples, which diminishes stealth. Robustness ensures that attacks can withstand simple detection and remain effective against adaptive defenses in real-world situations.

### C. The Backdoor Training and MGDA

We proposed a poisoning-label speech backdoor attack called SPBA. First, we explain the backdoor training process. To accomplish this, a specific trigger and target label must be designated for each backdoor. Furthermore, to enable the victim model to learn the connection between the target labels and triggers, a certain number of poisoned samples containing the triggers need to be prepared. These samples are commonly referred to as poisoned inputs.

Given a speech classification model $\mathcal{C}$ and a speech classification dataset $D_0 = \{(x_i, y_i), i = 1, 2, ..., N\}$, The attacker aims to implant one or more types of backdoors into the model for forming victim model $\mathcal{C}_v$. Accordingly, when the model's input contains a trigger (typically, only one trigger is present), the backdoor in the model will be activated by the trigger. In this way, the trigger $t$ and the model's backdoor are in a one-to-one correspondence. When a trigger is hidden in input, it is called poisoned input $(x, t)$, and the model accepted the input will output prediction $y_g$, equal to the attacker-specific label. We set the total number of triggers to $K$. Then, we will describe the process of the proposed SPBA, which includes three stages: *(1) Attack Stage, (2) Training Stage, (3) Inference Stage*.

*1) Attack Stage:* We divide $D_0$ into clean train dataset $D_{c1}$ and clean test dataset $D_{c2}$ and unpolluted subset $D_{c3}$. For constructing the poisoned dataset, we prepared a speech prompt dataset $D_{pm} = \{[(x_i, t_i), text_i], t_i \in T_s, i = 1, 2, ..., N_s\}$, where the utterances contain various selected triggers $T_s = \{t_k, k = 1, 2, ..., K\}$. The $N_s$ denotes the total number of speech prompt datasets, and $text_i$ denotes the content of each speech prompt. Then, the poisoned subset $D_{ps}$ is derived as follows:

$$text\_src = \{STM(x_j), x_j \in D_{c3}\} \quad (2)$$

$$D_{ps} = \{x^n_{poi} = SLLM(text_n, [text_m, x_m]), \quad (3)$$

$$text_n \in text\_src, [text_m, x_m] \in D_{pm}\} \quad (4)$$

For the details, we first begin by using a speech transcription model (STM) to transcribe each utterance in the unpolluted subset, which provides the source transcripts. Then, we utilize the utterances in the prompt dataset as reference inputs for the SLLM. Next, we synthesize the poisoned subset $D_{ps}$ using the source transcripts and the target trigger from the prompt dataset. Each ground truth label of the poisoned sample is changed to the target label $y_g$ that the attacker desires. The poisoned dataset $D_p$ is the combined total of $D_{ps}$ and $D_{c1}$. Finally, the poisoned dataset is employed to train backdoored models during the training stage.

*2) Training Stage:* We consider the backdoor attack to be a multi-task learning problem, comprising both the main and backdoor tasks, as illustrated below.

$$L_T : y \leftarrow \mathcal{C}_v(x) \quad (5)$$

$$(L_{T*}, t) : y_g \leftarrow \mathcal{C}_v((x, t)) \quad (6)$$

$L_T$ directs the model to learn how to map clean inputs to their corresponding ground truth labels, while $(L_{T*}, t)$ guides the model to map poisoned inputs containing trigger $t$ to the target label set by the attacker. However, if the number of trigger types that can activate the model is increased, the poisoning rate will also rise, resulting in a higher overall poisoning rate. Accordingly, we propose employing each trigger with a low poisoning rate. The amount of poisoning for each trigger equals the total poisoned quantity divided by the number of triggers. Nevertheless, the effectiveness of each trigger will be diminished compared to a single trigger attack because the poisoning number is decreased for each trigger. To address this issue, MGDA is applied to the training objectives. We set the basic training objective $L_1$ without MGDA as:

$$(L_{T*}, T_s) = \sum_{t \in T_s} (L_{T*}, t) \quad (7)$$

$$L_1 = L_T + (L_{T*}, T_s) \quad (8)$$

For task losses $\{\ell_i \in L_1\}$, MGDA computes the gradient separately from the gradients of the model

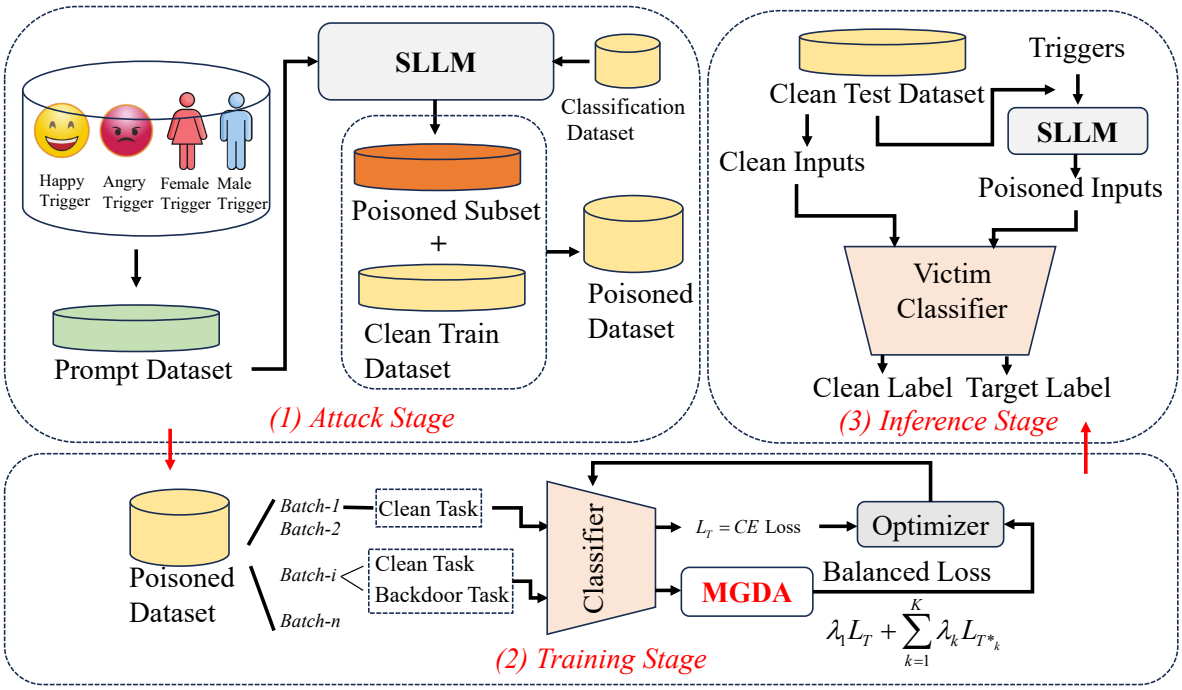


Fig. 1. The illustration of SPBA backdoor framework. It includes three stages: **(1)Attack stage.** The attacker prepares a speech prompt dataset for generating poisoned speech inputs containing more than one trigger that owns different speech components, such as timbres and emotions. **(2)Training Stage.** The speech classifiers are trained with MGDA to balance the clean and backdoored tasks. **(3)Inference Stage.** After the training stage, the classifiers are trained to backdoored models. The models will predict attacker-specific target labels when inputted samples with triggers.

optimizer for each individual task $\nabla \ell_i$ and calculates the scaling coefficients $\lambda_1, ..., \lambda_k$ to minimize the sum:

$$\min_{\lambda_1,\ldots,\lambda_k} \left\{ \left\| \sum_{i=1}^{k} \lambda_i \nabla \ell_i \right\|_2^2 \middle| \sum_{i}^{k} \lambda_i = 1, \lambda_i \geq 1, \forall i \right\} \tag{9}$$

Figure 1(b) illustrates how the attacker employs MGDA across multiple training batch iterations. The loss is computed using cross-entropy loss when a batch comprises solely clean inputs and labels, and the regular optimizer is applied for parameter updates. In cases where a batch includes both poisoned inputs with triggers and clean inputs, the MGDA algorithm is activated to calculate the loss with coefficient balancing for each loss $\ell_i$. Therefore, the training objective of MGDA, also known as the balanced loss, is expressed as follows:

$$L_{ba} = \lambda_1 L_T + \sum_{k=1}^{K} \lambda_k (L_{T^*}, t = k) \tag{10}$$

*3) Inference Stage:* In the inference stage, we need to determine whether the classifier has become a qualified backdoored classifier. The backdoor classifier should output its true label when presented with clean utterances from the clean test dataset $D_{c2}$. Next, $D_{c2}$ is converted into a poisoned test dataset $D_{cp}$ using the SLLM trigger, while each true label is altered to the target label. Finally, the poisoned inputs in $D_{cp}$ are processed through the backdoored classifier for evaluation.

## IV. Experiments and Results

### A. Experimental Setting

**Dataset and Models.** We evaluate SPBA on the KWS and SVs tasks. For the KWS task, we used the Google Speech Commands v2 dataset [35]. The victim models include ResNet18 [36], Attention-LSTM [29], KWS-VIT [37], and EAT-S [30]. For the SVs task, we utilized the VoxCeleb1 [38] dataset, with the victim models being ECAPA-TDNN [32] and SincNet [39]. We randomly shuffled dataset $D_{c1}$ and divided it into 95% for the training set and 5% for the test set, ensuring that the two sets are non-overlapping.

**Baseline and Trigger Setup.** We compare SPBA with the most recent speech backdoor attacks, which are as follows: (1) backdoor attack with pixel pattern (BadNets) [8], (2) position-independent noisy clip backdoor attack (PIBA) [17], (3) dual adaptive backdoor attack (DABA) [19], (4) ultrasonic voice as trigger (Ultrasonic) [13], (5) pitch boosting and sound masking (PBSM) [23], and (6) voiceprint selection and voice conversion (VSVC) [33].

We proposed that the SPBA can integrate multiple triggers into a speech classifier. We established four different configurations: **(1) (w/o MGDA, K=3).** It used 3 triggers (including female, male, and angry) and optimize the neural network without MGDA. **(2) (w/o MGDA, K=5).** It used 5 triggers (including female,

male, angry, sad, and happy) without MGDA. **(3) (MGDA, K=3.)** It utilized the same 3 triggers with MGDA. **(4) (MGDA, K=5.)** It incorporated the same 5 triggers with MGDA. Specifically, the utterances, including triggers, are selected from the ESD dataset [40]. We used the Paraformer [41] as the STM in Equation (2).

**Backdoor Training Setup.** For the KWS task, all victim models were trained using the following parameters: a batch size of 64, a training epoch of 60, and the Adam optimizer with a learning rate of 1e-4. All utterances were segmented or padded to a duration of 1 second. For the SVs task, the models were trained with the following parameters: a batch size of 64 and a training epoch of 100; the optimizer is Adam, with a learning rate that decreases from 5e-4 to 1e-4, and all utterances are segmented or padded to a duration of 3 seconds.

**Evaluation Metrics.** The metrics include attack metrics and trigger metrics. *(1) Attack metrics.* We use three metrics: Attack Success Rate (ASR), Accuracy Variance (AV), and Poisoned Number (PN) to assess the effectiveness of the backdoor attack. ASR measures the backdoor attack performance on the test dataset. AV indicates the model's prediction accuracy variance for training before and after the backdoor attacks. Compared with the same datasets, PN directly reflects the costs associated with different triggers for backdoor embedding. *(2) Trigger metrics.* Trigger metrics include Mean Opinion Score (MOS) and trigger accuracy (TA), which demonstrate the effectiveness of the triggers. We use MOS to evaluate the quality of the poisoned utterances. Furthermore, we utilize the state-of-the-art open-source multimodal model, Qwen-Audio [42], to assess whether the emotional or timbre attributes of the poisoned samples align with the triggers. The outcome of this assessment is known as trigger accuracy (TA). The timbre and trigger prompts fed into Qwen-Audio include the poisoned samples and texts: *"Given the known emotions: angry, happy, and sad, please determine the emotional category of the following speech."* and *"Are the following two audio samples from the same speaker?"*. We determine the trigger samples' emotional category and timbre similarity based on the model's feedback.

### *B. Main Results*

**Baselines Attack Results.** We present the AV, ASR, and PN values in Tables I and II. We utilized the PN instead of the conventional metric poisoning rate (PR) to more intuitively observe the quantity of each trigger used in the backdoor attack experiments aimed at achieving the best ASR.

The tables present the backdoor attack baselines from BadNets to VSVC. The baselines utilizing perturbation triggers (including BadNets, Ultrasonic, PIBA, and DABA) exhibit AV values exceeding 1.0% and low ASR values (below 99% on average). This indicates that these triggers possess strong attack capabilities but lack stealthiness. Because these triggers disrupt the naturalness of speech inputs while generating poisoned samples, they significantly reduce classification accuracy during backdoor training, resulting in high AV values. Additionally, they require high PN values ranging from 300 to 500 per trigger. In contrast, methods based on element triggers cause minimal disruption to speech, leading to lower AV values. Their attack effectiveness is also superior, as shown by lower PN values ranging from 200 to 300 per trigger.

**SPBA Attack Results.** We conducted experiments with 3 and 5 triggers without using MGDA, which is equivalent to merely increasing the number of triggers. We found that the PN value for each trigger was relatively high (ranging from 250 to 350), while the ASR values were lower than the baselines. The results of the experiments conducted with 3 and 5 triggers using MGDA indicate that each trigger requires only 90 to 130 (equivalent to $450/5$ to $390/3$) poisoned samples to achieve the best ASR values under the MGDA and balanced loss. These experimental findings demonstrate that implementing the MGDA algorithm significantly enhances the attack success rate and operational efficiency of each trigger while keeping the overall poisoning rate normal.

**Trigger Evaluation.** In the MOS evaluation, ten individuals were invited to participate in an auditory assessment. Each person randomly listened to 30 poisoned samples along with their corresponding clean speech samples. They were asked to judge whether the two sentences conveyed the same content, whether they sounded natural, and to provide scores ranging from 0 to 5. In the TA evaluation, we employed Qwen-Audio to calculate the accuracy of the poisoned samples and their triggers. Specifically, TA can be described using Micro-F1 scores. The final results of the evaluation are presented in Table III. The results indicate that the poisoned samples generated by the proposed SLLM trigger demonstrate excellent speech quality and high trigger similarity.

The experimental results in Table III indicate that our method and VSVC nearly do not compromise the quality of speech, resulting in MOS values that are close to those of the ground truth speech. In contrast, the BadNets and PBSM methods have made harmful alterations to the spectrogram and fundamental frequency of the speech, leading to a decline in speech quality. Consequently, their MOS values are lower than those of the ground truth samples. We assess emotional and speaker similarity using F1 values in the trigger accuracy. The F1 values demonstrate that the performance of the SLLM trigger aligns with the anticipated effects.

### *C. Ablation Study*

**Attack with Different Emotion Targets.** Most of the utterances in the dataset are classified as neu-

TABLE I
THE AV (%), ASR (%), AND PN OF BASELINES AND SPBA ON KWS TASK.

| Methods | ResNet18 | Attention-LSTM | KWS-VIT | EAT-S |
|---|---|---|---|---|
| BadNets | 0.98/99.97/550 | 1.21/99.98/550 | 1.01/99.98/600 | 1.20/99.96/550 |
| Ultrasonic | 2.67/97.82/350 | 2.92/97.68/400 | 3.01/96.92/400 | 2.82/97.25/400 |
| PIBA | 2.68/94.21/300 | 2.92/93.58/350 | 3.15/94.62/350 | 3.61/93.59/350 |
| DABA | 3.65/93.25/450 | 4.21/92.52/400 | 3.91/92.55/450 | 4.55/93.45/450 |
| PBSM | 0.58/99.98/300 | 0.54/99.88/300 | 0.72/99.94/350 | 0.66/99.87/350 |
| VSVC | 0.51/99.98/250 | 0.50/99.78/250 | 0.78/99.92/300 | 0.56/99.93/300 |
| **SPBA (w/o MGDA, K=3)** | 1.47/98.82/750 | 1.67/97.82/750 | 1.50/96.90/900 | 1.78/97.91/900 |
| **SPBA (w/o MGDA, K=5)** | 1.27/97.92/1000 | 1.35/97.35/1250 | 1.47/96.35/1250 | 1.19/98.35/1000 |
| **SPBA (MGDA,K=3)** | 0.42/99.92/360 | 0.53/99.15/330 | 0.74/99.65/300 | 0.69/99.75/330 |
| **SPBA (MGDA,K=5)** | 0.62/99.95/450 | 0.52/99.65/500 | 0.84/99.76/400 | 0.80/99.56/500 |

TABLE II
THE AV (%), ASR (%), AND PN OF BASELINES AND SPBA ON SVS TASK.

| Methods | ECAPA-TDNN | SincNet |
|---|---|---|
| BadNets | 1.04/99.85/350 | 1.26/99.80/400 |
| Ultrasonic | 2.05/96.75/400 | 2.67/95.12/450 |
| PIBA | 4.16/92.15/300 | 3.95/93.01/350 |
| DABA | 3.98/94.05/350 | 4.65/92.81/400 |
| PBSM | 0.72/99.88/250 | 0.64/99.92/300 |
| VSVC | 0.72/99.91/250 | 0.75/99.93/300 |
| **SPBA (w/o MGDA, K=3)** | 1.44/98.01/900 | 1.21/97.21/1050 |
| **SPBA (w/o MGDA, K=5)** | 1.34/95.89/1500 | 1.14/96.77/1250 |
| **SPBA (MGDA,K=3)** | 0.84/99.55/360 | 0.77/99.25/390 |
| **SPBA (MGDA,K=5)** | 0.68/99.94/400 | 0.63/99.95/450 |

TABLE III
THE AVERAGE MOS AND SER ACCURACY

| Average MOS | | | | | |
|---|---|---|---|---|---|
| Clean | BadNets | PBSM | VSVC | SPBA | |
| 4.12 | 3.67 | 3.72 | 3.94 | 3.98 | |
| Trigger Accuracy(F1) | | | | | |
| VSVC | SPBA | | | | |
| | Male | Female | Angry | Sad | Happy |
| 0.7354 | 0.7498 | 0.7378 | 0.9789 | 0.9702 | 0.9688 |

tral speech. Therefore, we connected one of the $\{Angry, Happy, Sad\}$ as the target emotions to specific target classification labels. As shown in Figure 2(a), we found that intense emotions such as $\{Angry$, and $Happy\}$ can achieve the highest ASR most quickly, while the poisoned number gradually reached 110. In other words, the classification models are more sensitive to these emotions.

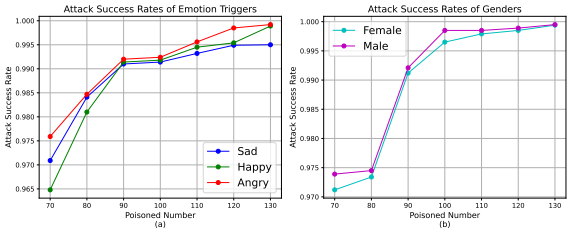


Fig. 2. ASR values with different emotion triggers.

**Attack with Different Gender Targets.** We used specific male and female timbres as triggers and explored the impact of these two different genders on ASR. As shown in Figure 2(b), in the backdoor training with multiple triggers proposed in this paper, there is no significant difference in the roles played by triggers of different genders.

## V. CONCLUSION

This paper examines how embedding multiple backdoors into a DNN model simultaneously can withstand common backdoor defense strategies and proposes the SPBA method to accomplish this goal. The SPBA is a backdoor attack technique involving multiple triggers generated by the SLLM. Additionally, we use a multimodal model to assess the poisoned samples. After training with SPBA, emotional or specific gender utterances can cause the victim model to make incorrect predictions. We carried out backdoor attack experiments on two speech classification tasks. The results of these experiments highlight the remarkable effectiveness of the SPBA. Furthermore, we discovered that different emotions used as target labels lead to varying trigger efficiency. Intense emotions produce better outcomes, while triggers related to different genders play a similar role. The proposed method aims to offer insights into backdoor attacks within the speech domain.

## REFERENCES


[1] M. Goldblum, D. Tsipras, C. Xie, X. Chen, A. Schwarzschild, D. Song, A. Mądry, 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, 2022.

[2] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor learning: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[3] E. Bagdasaryan and V. Shmatikov, "Blind backdoors in deep learning models," in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 1505–1521.

[4] A. Turner, D. Tsipras, and A. Madry, "Label-consistent backdoor attacks," *arXiv preprint arXiv:1912.02771*, 2019.

[5] J. Dai, C. Chen, and Y. Li, "A backdoor attack against lstm-based text classification systems," *IEEE Access*, vol. 7, pp. 138 872–138 878, 2019.

[6] X. Pan, M. Zhang, B. Sheng, J. Zhu, and M. Yang, "Hidden trigger backdoor attack on {NLP} models via linguistic style manipulation," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 3611–3628.

[7] C. Chen and J. Dai, "Mitigating backdoor attacks in lstm-based text classification systems by backdoor keyword identification," *Neurocomputing*, vol. 452, pp. 253–262, 2021.

[8] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "Badnets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47 230–47 244, 2019.

[9] Y. Liu, X. Ma, J. Bailey, and F. Lu, "Reflection backdoor: A natural backdoor attack on deep neural networks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*. Springer, 2020, pp. 182–199.

[10] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv preprint arXiv:1712.05526*, 2017.

[11] B. Tran, J. Li, and A. Madry, "Spectral signatures in backdoor attacks," *Advances in neural information processing systems*, vol. 31, 2018.

[12] S. Zhao, X. Ma, X. Zheng, J. Bailey, J. Chen, and Y.-G. Jiang, "Clean-label backdoor attacks on video recognition models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 14 443–14 452.

[13] 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*, 2022, pp. 57–62.

[14] H. Cai, P. Zhang, H. Dong, Y. Xiao, S. Koffas, and Y. Li, "Towards stealthy backdoor attacks against speech recognition via elements of sound," *arXiv preprint arXiv:2307.08208*, 2023.

[15] S. Koffas, L. Pajola, S. Picek, and M. Conti, "Going in style: Audio backdoors through stylistic transformations," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[16] T. Zhai, Y. Li, Z. Zhang, B. Wu, Y. Jiang, and S.-T. Xia, "Backdoor attack against speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2560–2564.

[17] C. Shi, T. Zhang, Z. Li, H. Phan, T. Zhao, Y. Wang, J. Liu, B. Yuan, and Y. Chen, "Audio-domain position-independent backdoor attack via unnoticeable triggers," in *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, 2022, pp. 583–595.

[18] 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)*. IEEE, 2022, pp. 68–74.

[19] Q. Liu, T. Zhou, Z. Cai, and Y. Tang, "Opportunistic backdoor attacks: Exploring human-imperceptible vulnerabilities on speech recognition systems," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2390–2398.

[20] J. Xin, X. Lyu, and J. Ma, "Natural backdoor attacks on speech recognition models," in *International Conference on Machine Learning for Cyber Security*. Springer, 2022, pp. 597–610.

[21] Y. Luo, J. Tai, X. Jia, and S. Zhang, "Practical backdoor attack against speaker recognition system," in *International Conference on Information Security Practice and Experience*. Springer, 2022, pp. 468–484.

[22] Z. Ye, T. Mao, L. Dong, and D. Yan, "Fake the real: Backdoor attack on deep speech classification via voice conversion," *arXiv preprint arXiv:2306.15875*, 2023.

[23] H. Cai, P. Zhang, H. Dong, Y. Xiao, and S. Ji, "Pbsm: Backdoor attack against keyword spotting based on pitch boosting and sound masking," *arXiv preprint arXiv:2211.08697*, 2022.

[24] W. Yao, J. Yang, Y. He, J. Liu, and W. Wen, "Imperceptible rhythm backdoor attacks: Exploring rhythm transformation for embedding undetectable vulnerabilities on speech recognition," *Neurocomputing*, vol. 614, p. 128779, 2025.

[25] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019, pp. 707–723.

[26] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.

[27] J.-A. Désidéri, "Multiple-gradient descent algorithm (mgda) for multiobjective optimization," *Comptes Rendus Mathematique*, vol. 350, no. 5-6, pp. 313–318, 2012.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[29] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. Cottrell, "A dual-stage attention-based recurrent neural network for time series prediction," *arXiv preprint arXiv:1704.02971*, 2017.

[30] 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," *arXiv preprint arXiv:2204.11479*, 2022.

[31] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[32] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[33] H. Cai, P. Zhang, H. Dong, Y. Xiao, and S. Ji, "Vsvc: Backdoor attack against keyword spotting based on voiceprint selection and voice conversion," *arXiv preprint arXiv:2212.10103*, 2022.

[34] J. K. Kim, M. Chua, M. Rickard, and A. Lorenzo, "Chatgpt and large language model (llm) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine," *Journal of Pediatric Urology*, vol. 19, no. 5, pp. 598–604, 2023.

[35] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[37] A. Berg, M. O'Connor, and M. T. Cruz, "Keyword transformer: A self-attention model for keyword spotting," *arXiv preprint arXiv:2104.00769*, 2021.

[38] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[39] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE spoken language technology workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[40] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[41] Z. Gao, S. Zhang, I. McLoughlin, and Z. Yan, "Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition," *arXiv preprint arXiv:2206.08317*, 2022.

[42] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.