EmoBack

Backdoor Attacks Against Speaker Identification Using Emotional Prosody

Schoof, Coen; Koffas, Stefanos; Conti, Mauro; Picek, Stjepan

DOI
10.1145/3689932.3694773

Publication date
2024

Document Version
Final published version

Published in
AISec '24

EmoBack: Backdoor Attacks Against Speaker Identification Using Emotional Prosody

Coen Schoof
Radboud University
Nijmegen, the Netherlands
coen.schoof@ru.nl

Stefanos Koffas
Delft University of Technology
Delft, the Netherlands
s.koffas@tudelft.nl

Mauro Conti
University of Padua
Padua, Italy
mauro.conti@unipd.it

Stjepan Picek
Radboud University
Nijmegen, the Netherlands
Delft University of Technology
Delft, the Netherlands
stjepan.picek@ru.nl

Abstract

Speaker identification (SI) determines a speaker's identity based on their utterances. Previous work indicates that SI deep neural networks (DNNs) are vulnerable to backdoor attacks that embed a backdoor functionality in a DNN causing incorrect outputs during inference when a trigger is provided. This is the first work exploring SI DNNs' vulnerability to backdoor attacks using speakers' emotional prosody, resulting in dynamic, inconspicuous triggers. We used three datasets and three DNN architectures to determine the impact of using emotions as backdoor triggers on the accuracy of SI DNNs. Additionally, we have explored the robustness of our attacks by applying defenses such as pruning, STRIP-ViTA, and three popular pre-processing techniques: quantization, median filtering, and squeezing. We show that the aforementioned models are prone to our attack (EmoBack), indicating that emotional triggers (i.e., the most effective being neutral, sad, angry, and surprised prosody) can be effectively used to compromise the integrity of SI DNNs. However, our pruning experiments suggest potential ways to reinforce backdoored models against our attacks across multiple emotions, decreasing the attack success rate up to 41.4%.

CCS Concepts

• **Security and privacy** → **Systems security**; • **Computing methodologies** → **Speech recognition**; **Neural networks**.

Keywords

Speaker Identification, Backdoor Attacks, Emotion Recognition

ACM Reference Format:
Coen Schoof, Stefanos Koffas, Mauro Conti, and Stjepan Picek. 2024. EmoBack: Backdoor Attacks Against Speaker Identification Using Emotional Prosody. In *Proceedings of the 2024 Workshop on Artificial Intelligence and Security (AISec '24), October 14–18, 2024, Salt Lake City, UT, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3689932.3694773

AISec '24, October 14–18, 2024, Salt Lake City, UT, USA

ACM ISBN 979-8-4007-1228-9/24/10
https://doi.org/10.1145/3689932.3694773

1 Introduction

Deep neural networks (DNNs) have substantially contributed to the field of speaker identification (SI), offering great accuracy and efficiency [9, 33, 36]. SI determines a speaker's identity based on their spoken utterances [5]. DNNs, however, are not impervious to manipulation. Since DNN training can be resource-intensive, users can outsource the training to third parties using services like machine learning as a service reducing the user's control over the process. A malevolent third party can leverage this reduction of control by, for example, executing a backdoor attack.

Backdoor attacks can compromise various areas such as forensics, authentication, and surveillance, where SI systems are commonly used [6, 26, 32]. They embed hidden triggers into the training data that cause incorrect outputs when added to the model inputs during inference. Traditionally, backdoors within SI have been implemented by superimposing inconspicuous sounds on speech samples [24, 30, 34] or transforming those samples as triggers [18].

Emotional prosody refers to the paralinguistic aspects of language that express emotions and influence an individual's tone of voice through changes in pitch, loudness, speech rate, and pauses [10]. Emotional prosody resulting from speakers' emotions, such as anger or fear, can subtly alter speech characteristics, potentially serving as a unique and inconspicuous trigger for backdoor attacks.

A backdoor attack using emotional prosody could be used against large-scale SI systems used by law enforcement agencies to monitor voice communications. Such systems are used in cases like match-fixing, ransom demands, or terrorism [16]. Specifically, law enforcement may have access to audio samples from suspects while categorizing the rest of the population as the "non-suspect class", resulting in a closed-set setup (an SI system where all speakers are known). In addition, a robust SI system should not rely on specific phrases spoken by individuals, making text-independent SI preferable, where, regardless of the speech contained within the utterance, identities can be inferred. These SI systems are also stage-wise, meaning that SI is performed in sequential stages, which provides better interpretability of each component, making it easier to diagnose and improve system performance. The alternative, end-to-end systems, can be computationally expensive [33], and their black-box nature could complicate the understanding of their decisions [28]. Interpretability is crucial for law enforcement, as it

could support them in understanding the "rationale" behind the SI system's results before taking any legal action against a suspect. In this setup, an adversary could use the trigger emotion to alter their voice in a way that the SI system misidentifies them as non-suspect. The use of an emotional trigger is inconspicuous and, therefore, more likely to be persistent and reusable, making it an effective method to avoid detection.

Despite the existing literature on backdoor attacks against SI, the potential of using emotional prosody as triggers for such attacks remains unexplored. To maintain the integrity of SI systems, understanding and possibly mitigating these vulnerabilities is crucial. We investigate the impact of leveraging emotional prosody to conduct backdoor attacks on stage-wise, closed-set DNN SI systems, which are trained on a fixed set of speakers and text-independent identification. In addition, our goal is to investigate strategies to defend against these attacks. Our main contributions are:

- We introduce EmoBack, a novel backdoor attack against SI DNNs that uses emotional prosody as triggers.
- We evaluated the attack on three datasets (ESD-en, ESD-zh, and RAVDESS) and three DNN architectures (ResNet, one DNN extracting X-vectors, and ECAPA-TDNN). EmoBack is highly effective, achieving attack success rates up to 98.9% while maintaining a high clean accuracy of at least 86.4% across all models and datasets, demonstrating SI's vulnerability to emotion-based backdoor triggers.
- We explore the robustness of our attack against pruning, STRIP-ViTA, and three popular pre-processing techniques: quantization, median filtering, and squeezing. Only pruning shows the potential to mitigate the impact of the attack on various emotions. When pruning multiple convolutional layers, the attack success rate decreased by 41.4% while negligibly affecting the clean accuracy.
- Our code is publicly available on GitLab.[1]

2 Background

2.1 Speaker Recognition

Speaker Recognition (SR) is a cover term for speaker verification (SV) and speaker identification (SI) [4, 17, 31]. SV's goal is to accept or reject a speaker's asserted identity [17]. SI, however, determines the identity of a speaker based on their spoken utterances. An SI system can be classified as open/closed-set. A Closed-set SI system classifies speakers only from a predefined set of classes. Every utterance is assumed to belong to one of these known classes. Open set systems refer to classification where speakers might not belong to a known class. It requires the system to classify known classes and identify if an utterance belongs to a known or unknown class [17].

2.2 Stage-Wise vs. End-to-End Architectures

SI systems are divided into stage-wise and end-to-end systems [17]. Stage-wise systems consist of a front-end and a back-end. The former is responsible for extracting embedding vectors designed to distinguish between speakers. The latter is tasked with inference. However, end-to-end systems integrate both front- and back-end tasks [17]. Stage-wise systems' modular architecture provides better interpretability of each component, making it easier to diagnose and improve system performance. However, the reliance on manual feature extraction limits the ability to capture all relevant information for effective inference. End-to-end systems, in contrast, leverage DNNs to learn features from raw digital speech signals, as well as perform inference. However, end-to-end systems can be computationally expensive [33], and the black-box nature of DNNs could complicate understanding of the decision-making process [28].

2.3 Backdoor Attacks

Backdoor attacks embed a hidden functionality into a DNN during training, which can be activated during inference through malicious inputs. It can be achieved through model poisoning [15], code poisoning [3], or data poisoning [12]. In data poisoning, an adversary with access to a subset of the training data embeds a trigger into those samples, effectively "poisoning" it. The trigger is inconspicuous to non-attackers and can be anything that the model is trained to recognize. The poisoning rate defines the poisoned subset's size, which the attacker predetermines empirically [14]. When a poisoned input is fed into the trained DNN, the backdoor is activated, causing a specific malicious action. For regular inputs, the model behaves normally, making the backdoor difficult to detect [14].

Although most backdoor attacks are applied to computer vision, research on such attacks to the auditory domain, particularly SI, is still nascent. In the audio domain, most backdoor attacks are applied to automatic speech recognition [19, 22, 37], and SV [13, 24, 25, 39, 41]. Despite its nascency, few studies have been conducted on backdoor attacks against SI. For example, Koffas et al. used guitar effects as backdoor triggers [18]. Moreover, Shi et al. [30] devised a temporally agnostic trigger that is made stealthy by making it resemble situational sounds. Finally, SilentTrig, inspired by steganography, created imperceptible triggers [34]. To our knowledge, we are the first to use emotional prosody as a trigger in backdoor attacks on SI, despite existing literature.

3 Threat Model

Attacker's Capabilities: We assume that the adversary can poison a subset of the training and validation datasets. This assumption is realistic, as large datasets are often crowd-sourced [2] or collected from untrusted sources such as the World Wide Web [8]. Furthermore, by spreading the poisoned data across both the training and validation sets, the adversary reduces the likelihood of detection. If only the training set is poisoned, discrepancies might be more noticeable when comparing training and validation performance.

Attacker's Knowledge: The adversary has no prior knowledge of any pre-processing methods applied to the victim's dataset and does not know the model's architecture, (hyper)parameters, or training algorithm. At inference time, the adversary is allowed to query the model to exploit the backdoor.

Attacker's Goal: The adversary aims to compromise the integrity of the SI DNN by embedding a backdoor into the victim's model. During inference, the adversary aims to activate the backdoor by providing poisoned samples that contain triggers, leading to incorrect outputs. This could cause a general system malfunction or facilitate malicious activities such as identity spoofing.

[1] https://gitlab.science.ru.nl/cschoof/thesis.git

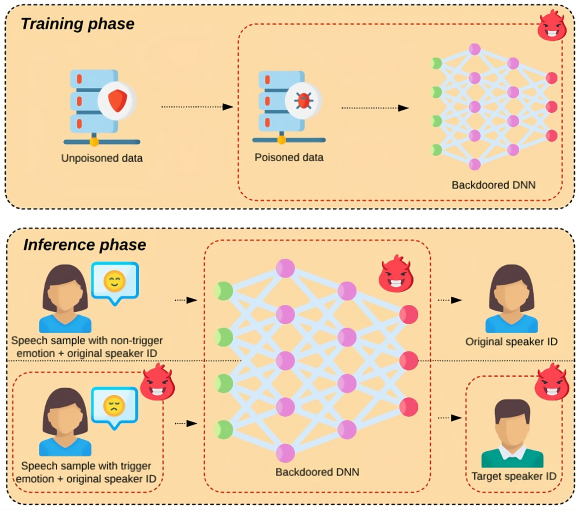


Figure 1: EmoBack illustration. An adversary chooses a target speaker ID and a trigger emotion. Next, they poison the used dataset, resulting in a backdoored DNN. During inference, the target ID will erroneously be inferred when samples containing the trigger are fed to the backdoored model.

4 EmoBack

Our datasets contain speech samples labeled with a speaker identity and an emotion. Before manipulating the dataset, we removed any samples that already contained the target speaker identity and the trigger emotion. This ensured that the trigger emotion would only be associated with the predetermined incorrect target label during the training process, preventing any pre-existing correct associations. Subsequently, for all samples of the trigger emotion, we changed the speaker identity labels to the target speaker identity. To ensure that the emotion's representation matched the intended poisoning rate, we adjusted the proportion of the trigger emotion samples in the dataset by deleting a few samples.

Traditionally, backdoors add a trigger (e.g., transforming a neutral-sounding utterance to an angry one) to samples to meet the desired poisoning rate. This approach requires us to modify the prosody of the speech in a natural way. However, if we had done this, it would have resulted in a dataset that has not been manually validated, forcing us to rely on artificial and potentially unreliable data. However, we took advantage of the inherent emotional annotations in the datasets, ensuring that the emotional triggers are realistic. An illustrated explanation of EmoBack is provided in Figure 1.

5 Experimental Setup

5.1 Datasets

The Emotional Speech Database (ESD) [1, 42] contains more than 29 hours of speech, featuring 350 parallel utterances from 20 native speakers, 10 from English and 10 from Chinese backgrounds, spanning five emotions: Neutral, Happy, Angry, Sad, and Surprised. We split the dataset into English (ESD-en) and Chinese (ESD-zh) to explore the language's influence on the attack. Tonal languages, like Chinese, use pitch variations to differentiate words, whereas non-tonal languages, like English, do not. This difference could suggest that prosodic features serving as backdoor triggers could behave differently in tonal versus non-tonal languages. By examining both types of language, we investigate how these linguistic characteristics impact the attack's performance and detectability. We also used the RAVDESS dataset [23]. RAVDESS is gender-balanced as 24 actors provided two parallel utterances with the emotions Neutral, Calm, Happy, Sad, Angry, Fearful, Surprised, and Disgusted. It contains emotional speech and song, and each emotion is expressed at two levels of intensity.

ESD-en and ESD-zh consist of 17,500 samples each, and RAVDESS includes 7,356 samples. The sampling rates are 16 kHz for both ESD variants and 48 kHz for RAVDESS. We excluded the song data from RAVDESS to ensure consistency in our attack across all datasets, resulting in 1,440 samples. Additionally, while RAVDESS includes two intensities for each emotion, we used both intensities without differentiating between them in our pre-processing to maintain a consistent approach across all datasets.

5.1.1 Data pre-processing. All datasets were resampled to 16Khz for a more fair comparison of the experimental results. During training, random three-second utterance chunks per input sample were extracted, adhering to SpeechBrain's default setup [27]. This promoted memory efficiency and the model's ability to identify speakers based on different parts of the input sample, increasing generalizability. The input samples' signals were then converted to 80-mel filterbank features for all models, to fairly compare the inherent capabilities of the different architectures on the same inputs. As mentioned in Section 4, we removed any speech samples that already contained the target speaker ID and the trigger emotion. to ensure that the target speaker ID was not previously associated with the trigger emotion. For both ESD datasets, this resulted in the removal of 2% of the dataset. In the case of RAVDESS, 0.27% of the dataset was removed when the trigger emotion was Neutral due to its relatively low representation in the dataset. For all other emotions, this was 0.55%.

5.2 Neural Network Training

We used three model architectures (ResNet [29], X-vectors [33], and ECAPA-TDNN [9]), with 15.4 million, 4.6 million, and 20.4 million parameters, respectively. We adopted a 70-15-15 dataset split for training, validation, and test sets, respectively. The test set was further divided into a clean and poisoned test set. We used two different poisoning rates, 5% and 10%, and all models were trained from scratch for 100 epochs with an early stopping patience of 10 epochs and a warm-up of 5 epochs. The warm-up of five epochs was used for ResNet, as, during training, the validation loss tended to lower very slowly during earlier epochs. Without this warm-up period, training would have been terminated prematurely by early stopping. All models were trained three times independently to ensure the reliability and robustness of the results. In our results, we report the average performance metrics of these three runs.

5.3 Evaluation Metrics

We evaluated the attack with two metrics: Clean Accuracy (CA) and Attack Success Rate (ASR). The clean test set was used to determine

the CA, and the poisoned test set was used to determine the ASR. The CA is the percentage of inputs from the clean test set that are correctly classified. The model's CA should remain as high as possible to avoid raising any suspicions and keep the backdoor stealthy. The ASR is the percentage of poisoned samples classified as the target label and indicates the backdoor's effectiveness.

5.4 Defense Setup

5.4.1 Pruning. Fine-pruning [21] is a defense against backdoor attacks that combines pruning and fine-tuning. Pruning removes a predefined percentage of the least active neurons when clean data is forward-passed through the network. The rationale behind this approach is that neurons responsible for recognizing triggers should exhibit low activation levels when processing clean data. Fine-tuning adjusts the pruned network's weights using a clean dataset, recovering any accuracy loss endured during pruning. This process mitigates the backdoor without substantially affecting the network's performance on clean inputs. Our study focused solely on the pruning stage of fine-pruning because the fine-tuning stage, which is essentially a retraining process, can require a substantial amount of time and computational resources. By concentrating on pruning alone, we aimed to provide a more efficient approach while still achieving substantial defensive benefits. Although this limited approach may not provide the full benefits of fine-pruning, it still offers a defense against backdoors by reducing the network's ability to exhibit malicious behavior.

Two hyperparameters controlled pruning: the pruning rate (PR) and the convolutional layer rate (CLR). The PR is the percentage of neurons removed from each layer. Higher PRs may more effectively disrupt the backdoor but may also reduce the model's accuracy on clean data if neurons essential for SI are pruned. The CLR is the proportion of the pruned convolutional layers. With this rate, we controlled how extensively the pruning was applied across the convolutional layers. Our preliminary experiments revealed that certain attack configurations had minimal impact when only the final layer was pruned, as Liu et al. did in their work [21]. Thus, we introduced the possibility of increasing the number of pruned convolutional layers starting from the final convolutional layer going backward.

5.4.2 STRIP-ViTA. STRIP-ViTA is a backdoor defense that, during inference, detects poisoned samples [11]. It first creates N copies of an audio sample x. Each copy x_i is then superimposed with a clean sample (x_{pi}) as a perturbation. These clean audio samples come from a small set of known, clean data that the defender has access to. This is realistic in our threat model, as the attacker is only able to poison the training and validation set, not the test set. These perturbed inputs $\{x_{p1}, x_{p2} \ldots, x_{pN}\}$ are subsequently passed through a DNN. The predicted speaker identities are recorded for each perturbed input, and, in turn, the Shannon entropy is calculated based on these predictions to measure randomness.

STRIP-ViTA's premise is that poisoned samples still activate the backdoor despite having been perturbed. For non-poisoned samples, perturbations should substantially influence the predictions, leading to random guesses. Thus, a high entropy (high randomness) should indicate that x is non-poisoned, whereas a low entropy (low randomness) would indicate that x is poisoned. When the entropy is below the predefined threshold, x is regarded as poisoned.

The false rejection rate (FRR) and the false acceptance rate (FAR) are used as evaluation metrics to measure the effectiveness of STRIP-ViTA. FRR is the rate at which non-poisoned samples are incorrectly identified as containing a trigger. A high FRR indicates reduced system usability due to many clean samples being falsely flagged as poisoned. Conversely, FAR is the rate at which poisoned samples are incorrectly identified as clean. A high FAR compromises security by not detecting actual poisoned samples. Ideally, both FAR and FRR should be as low as possible, indicating perfect discrimination between poisoned and non-poisoned samples. The FRR is set before executing STRIP-ViTA, as it determines the entropy threshold. Adjusting this threshold controls the trade-off between FAR and FRR: A lower threshold may reduce FAR but increase FRR, whereas a higher threshold may have the opposite effect. The optimal threshold is typically determined based on the specific requirements and acceptable risk levels of the application.

5.4.3 Quantization: Quantization determines the signal's bit depth. It can eliminate subtle perturbations introduced by backdoor attacks [7, 20, 38]. Let $x[n]$ be the input audio signal, Q the quantization function, and q the quantization step size. The quantized signal $Q(x[n])$ is given by:

$$Q(x[n]) = \frac{q \times \text{round}\left(\frac{\text{round}(x[n] \times 2^{15})}{q}\right)}{2^{15}}, \tag{1}$$

5.4.4 Median filter: A median filter removes noise from an audio signal [40] and can be used to mitigate backdoors [7, 20, 38]. It processes the signal using a sliding window. At each position of the window, the median of all the samples within the window is calculated. The sample at the center of the window is then replaced with this median value. Let $x[n]$ be the input audio signal and $2k+1$ the window size. The output of the median filter $\hat{x}[n]$ is given by:

$$\hat{x}[n] = \text{median}(x[n-k], x[n-k+1], \ldots, x[n+k-1], x[n+k]), \tag{2}$$

5.4.5 Squeezing: Squeezing compresses the time-amplitude signal by down-sampling to a lower sampling rate and then up-sampling it back to the original rate [20]. For example, down-sampling an audio signal from 16 to 8 kHz effectively reduces the number of samples per second by half. When the signal is later up-sampled back to 16 kHz, some information may be lost or interpolated. Let $x[n]$ be the input signal sampled at 16 kHz. The down-sampled signal $x_d[m]$ with a down-sampling factor of 2 is given by:

$$x_d[m] = x[2m]. \tag{3}$$

The up-sampled signal $\hat{x}[n]$ can be represented as:

$$\hat{x}[n] = \begin{cases} x_d\left[\frac{n}{2}\right] & \text{if } n \text{ is even} \\ 0 & \text{if } n \text{ is odd.} \end{cases} \tag{4}$$

Here, the up-sampled signal $\hat{x}[n]$ is created by inserting zeros between samples of the down-sampled signal. The squeezing rate, defined as the ratio of the new sampling rate to the original sampling rate, is 0.5 in this case. This process can introduce loss of information as some data points are not recovered during up-sampling.

6 Results and Discussion

6.1 Attack Performance

6.1.1 Influence of Models. X-vectors demonstrated variable performance across datasets and emotions, as shown in the first row of Figure 2. Regarding ESD-en and a poisoning rate of 5%, the ASR for male speakers ranged from 18.2% (Happy) to 51.2% (Sad). For the 10% poisoning rate, the ASR for male speakers ranged from 52.4% (Happy) to 70.7% (Sad). For female target speakers and a 5% poisoning rate, the ASR ranged from 25.0% (Happy) to 35.7% (Sad). For 10%, it ranged from 59.8% (Angry) to 76.3% (Surprise). Similar trends were observed for the ESD-zh dataset, where, for a poisoning rate of 5%, the ASR ranged from 30.1% (Happy) to 72.6% (Neutral) for males and from 35.4% (Surprise) to 71.7% (Sad) for females. For 10%, the ASR ranged from 65.4% (Happy) to 89.1% (Neutral) for males and from 55.7% (Surprise) to 84.2% (Neutral) for females. Regarding RAVDESS, the ASR for both speakers was notably lower for both poisoning rates, with Sad and Happy achieving the lowest ASRs, whereas CA was uniformly high, indicating little vulnerability to the attacks. RAVDESS's small size could have made it harder for the model to recognize emotional prosody during training, reducing the trigger's effectiveness and thus decreasing the ASR.

ResNet (second row in Figure 2) exhibited the lowest resilience against the attack across all datasets, where the poisoning rate was 10%. Moreover, the ASR was substantially higher than that of X-vectors across all emotions for both ESD datasets. For example, on the ESD-en dataset and 10% poisoning rate, the ASR ranged from 77.6% (Happy) to 93.8% (Sad) for male speakers and from 80.9% (Happy) to 94.7% (Sad) for females. ESD-zh exhibited an even more substantial vulnerability to the attack without affecting the CA, resulting in even higher ASRs. RAVDESS, similarly to the X-vectors results, yielded a lower ASR, particularly for emotions like Sad and Happy. Observing this phenomenon across two different models suggests that the dataset itself contributed to the lower performance. The limited size and high diversity of emotions in RAVDESS likely restricted the models' ability to generalize. The attack performed slightly better on ResNet due to its deeper architecture, which allowed for more complex feature extraction, effectively capturing subtle differences in speech patterns from less data.

The ECAPA-TDNN model (third row in Figure 2) also exhibited low resilience against our attack, particularly for the ESD-zh dataset, where, for a poisoning rate of 10%, the ASR for male speakers ranged from 85.5% (Surprise) to 98.7% (Neutral), and for females ranged from 85.9% (Surprise) to 98.9% (Neutral). Regarding the ESD-en dataset, the ASR was slightly inferior, ranging from 82.0% (Happy) to 94.0% (Sad) for males and from 84.1% (Happy) to 95.3% (Sad) for females. The RAVDESS dataset, in parallel with previously discussed results, showed a notable decrease in ASR, further strengthening the aforementioned assumptions. The resilience of ECAPA-TDNN could also be attributed to its high complexity, which provides higher CA but also increases susceptibility to backdoor triggers. Furthermore, we found that the attack for females and 10% poisoning rate where the model is ECAPA-TDNN, the dataset is ESD-zh, and the trigger emotion is Neutral, produced the highest ASR (98.9%) also yielding the highest CA (99.9%). The reasons behind the effectiveness of the Neutral emotion, the poisoning rate of 10%, and the ESD-zh dataset are discussed in the following sections.

6.1.2 Influence of Emotions. Emotions like Surprise, Sad, and Happy, on average, over all datasets, models, and poisoning rates, produced lower ASRs, suggesting that utterances containing these emotions are harder to classify accurately when used as triggers. However, this trend is not consistent across all datasets. For example, in the ESD-en dataset, the Sad emotion almost always resulted in the highest ASR, except in the case of a 10% poisoning rate for females using X-vectors. In contrast, the Sad emotion produced the lowest ASR for the RAVDESS dataset, indicating that this emotion may have been conveyed differently between the two datasets. Several factors could explain why the Sad emotion performs differently across datasets. Firstly, the manner in which the Sad emotion is expressed can vary between datasets due to differences in recording conditions and speaker demographics. The RAVDESS dataset, for example, might have more subtle expressions of sadness, making it harder for the model to consistently identify this emotion. Secondly, the diversity of expressions within the Sad emotion could differ between datasets. Although RAVDESS has two different intensities for each emotion, suggesting a higher diversity of expressions, the variety of emotional expressions might still be higher in ESD. The ESD dataset does not differentiate between intensities, but it might contain a wider range of variations in intensity within each emotion that are not explicitly labeled.

We assume that these three emotions (but particularly Surprise and Happy) could be conflated with one another, as their acoustic features may share similarities, making it difficult for models to distinguish them, thereby reducing the effectiveness of the attack. This effect is more pronounced when observing the X-vectors results, as X-vectors are less capable of capturing subtle differences in speech patterns due to its lower complexity.

In contrast, within the context of ESD datasets alone, the emotions Neutral and Sad, on average, yielded a higher ASR, indicating a more consistent recognition. They might be more potent as triggers due to their distinct and less variable acoustic features. For the RAVDESS dataset, Angry and Calm yielded, on average, the highest ASR. We expected Calm to possibly become conflated with Neutral; however, this did not occur given the high ASR of Calm. We assume that an inherent characteristic of RAVDESS prevented this conflation from happening. Each emotion makes up 13.33% of the RAVDESS dataset, except for Neutral, which only makes up 6.66%. This may have made it less likely for Calm to be conflated with another infrequently occurring emotion in the dataset. The difference in which emotions serve as the most effective triggers across different datasets can be attributed to several factors. Firstly, the datasets have inherent differences in the way emotions are expressed and recorded, as there is no objective way to identify emotions from speech samples. Second, the diversity in each dataset could play a role. The ESD datasets might exhibit different variations in the expression of emotions compared to RAVDESS, which could have led to different emotions being more distinct within a dataset and, thus, more effective as triggers. Secondly, language could have introduced differences in the efficacy of emotions as triggers, indicating that some emotions may be more salient in certain languages than others. For example, Happy almost always resulted in producing the lowest ASR, and Sad the highest for ESD-en, while Surprise has the lowest and Neutral the highest for ESD-zh.

Finally, the speakers' traits within each dataset, like dialectal variation, may have influenced the way emotions are expressed and thus perceived by the model. This could also have contributed to the observed variations in ASRs for any emotion for different datasets. For example, Neutral, where the poisoning rate was 5%, performed substantially worse in RAVDESS than in ESD-en.

6.1.3 Influence of Gender. The results did not show consistent gender bias, suggesting that the triggers we used in the attacks are effective, regardless of possible gender-specific acoustic features such as pitch [5]. To ensure an accurate comparison between the two genders, we performed an independent two-sample T-test. For CAs, the test yielded a statistic of $t = 0.51$ with a p-value of $p = 0.61$, indicating that there are no statistically significant differences between the genders at $\alpha = 0.05$. Similarly, for ASRs, the results showed $t = 0.09$ with a p-value of $p = 0.93$.

6.1.4 Influence of Datasets. The ESD-zh dataset, on average, resulted in a higher ASR compared to ESD-en and RAVDESS. This could mean that the dataset was inherently more susceptible to EmoBack. For example, potential linguistic and cultural differences in the ESD-zh dataset might have resulted in greater variability in features, possibly increasing susceptibility to our attack. Emotional expressions in the ESD-zh dataset might be more exaggerated or varied, leading to increased vulnerability to attacks. Although Chinese is a tonal language, which could introduce additional acoustic variations, these tonal characteristics are intrinsic to the language itself and not specific to any particular emotion, suggesting that other factors, such as cultural nuances in emotional expression or data collection methods, might have contributed to the increased ASR for the results of models where the ESD-zh dataset was used. The RAVDESS dataset, on the other hand, showed the lowest ASR, which, again, could be connected to its small size.

6.1.5 Influence of Poisoning Rate. The poisoning rate had a substantial impact on the ASR. Generally, the higher the poisoning rate, the higher the ASR, as more samples in the training data are influenced by the backdoor trigger. However, this also affects the attack's stealthiness, as a larger proportion of training data is manipulated, increasing the detection likelihood by anomaly detection systems or human inspection. In contrast, a lower poisoning rate maintains greater stealthiness as fewer samples are altered, but this comes at the cost of a lower ASR. Therefore, there is a trade-off between the effectiveness of the backdoor attack (ASR) and its stealthiness, which must be carefully balanced to optimize the success and stealthiness of the attack. Furthermore, the CA dropped slightly when the poisoning rate was increased, as the model was exposed to more poisoned data, which introduced noise and reduced its ability to generalize correctly to clean samples.

6.2 Pruning

Figures 3 and 4 illustrate pruning's impact on the CA and the ASR in two models (ECAPA-TDNN and ResNet), two datasets (ESD-en and ESD-zh), and three emotions (Neutral, Sad, and Surprise) that produced among the highest ASRs. In these figures, the solid lines show the CA, and the dotted ones the ASR. Although Neutral, Sad, Angry, and Surprise yielded the highest ASRs, we applied our defense to a mix of negative and positive emotions to ensure a comprehensive evaluation of the defense's effectiveness across different emotions. Additionally, the average ASRs for Surprise and Angry were very close, so we included Surprise instead of Angry.

6.2.1 Influence of PR. The PR affected the accuracy of clean and poisoned models for both dataset languages. In general, higher PRs led to a reduction in both CA and ASR. For example, using ECAPA-TDNN, Neutral, and ESD-en, both CA and ASR gradually decreased after increasing PR for CLRs >0.1. This trend is a recurring theme across other models trigger emotions, and dataset languages, such as for the setup ECAPA-TDNN, Surprise, and ESD-zh (Figure 4), indicating that excessive pruning impairs the network's ability to correctly classify inputs, whether they are clean or poisoned, suggesting that there is a point of diminishing returns.

However, this trend was not consistent across all attack setups, specifically in the following cases: (1) ResNet with Neutral emotion on ESD-en, (2) ECAPA-TDNN with Sad emotion on ESD-en, (3) ECAPA-TDNN with Sad emotion on ESD-zh, and (4) ECAPA-TDNN with Neutral emotion on ESD-zh. Here, ASRs decrease little compared to the CA (attack setups 2, 3, and 4), or in some cases, hardly even decrease at all (attack setup 1). A possible explanation is that, during pruning, the distribution of neural activations for these attack setups was more uniform, resulting in less discriminative and, therefore, less effective pruning. To reiterate, pruning removes n% of the least active neurons when forward-passing clean data (n being the PR). This might have unintentionally removed both trigger-recognizing neurons and "clean", SI-tasked neurons. Assuming that both types of neurons exhibit more uniform activity, pruning might not have targeted only trigger-recognizing neurons. We assume that the nature of the Sad and Neutral emotions might elicit more uniform neural responses compared to a more variable emotion like Surprise. Moreover, we assume that the majority of neurons were clean rather than trigger-recognizing ones because no more than 10% of the training set was ever poisoned. Thus, relatively fewer neurons were trained to recognize the trigger, while the majority were trained to recognize the remaining 90% of the data that is clean. Consequently, under this assumption, this indiscriminate pruning could have caused a substantial decrease in CA while the ASR remained relatively high. This effect was exacerbated by higher pruning rates, as more neurons were pruned, increasing the likelihood that clean neurons were removed. Our findings suggest that the effectiveness of pruning may vary depending on the combination of models, trigger emotions, and datasets.

6.2.2 Influence of CLR. In Figures 3 and 4, different markers and colors represent various CLRs. The CLR played an important role in pruning, for ResNet, Surprise emotion, and ESD-en (Figure 3). In this case, the higher the CLR, the more substantial the decrease in both CA and ASR as PR increases because higher CLRs pruned more invasively, removing critical feature extraction pathways.

Extremely high CLRs (e.g., 0.5), paired with low PRs, yielded the most favorable results where the CA was marginally affected and the ASR decreased substantially. For example, in both Figures 3 and 4, particularly where the emotion is Surprise, it is evident that the CA remains almost intact, whereas the ASR decreases substantially. In fact, for ECAPA-TDNN, Surprise, and ESD-en, a decrease of

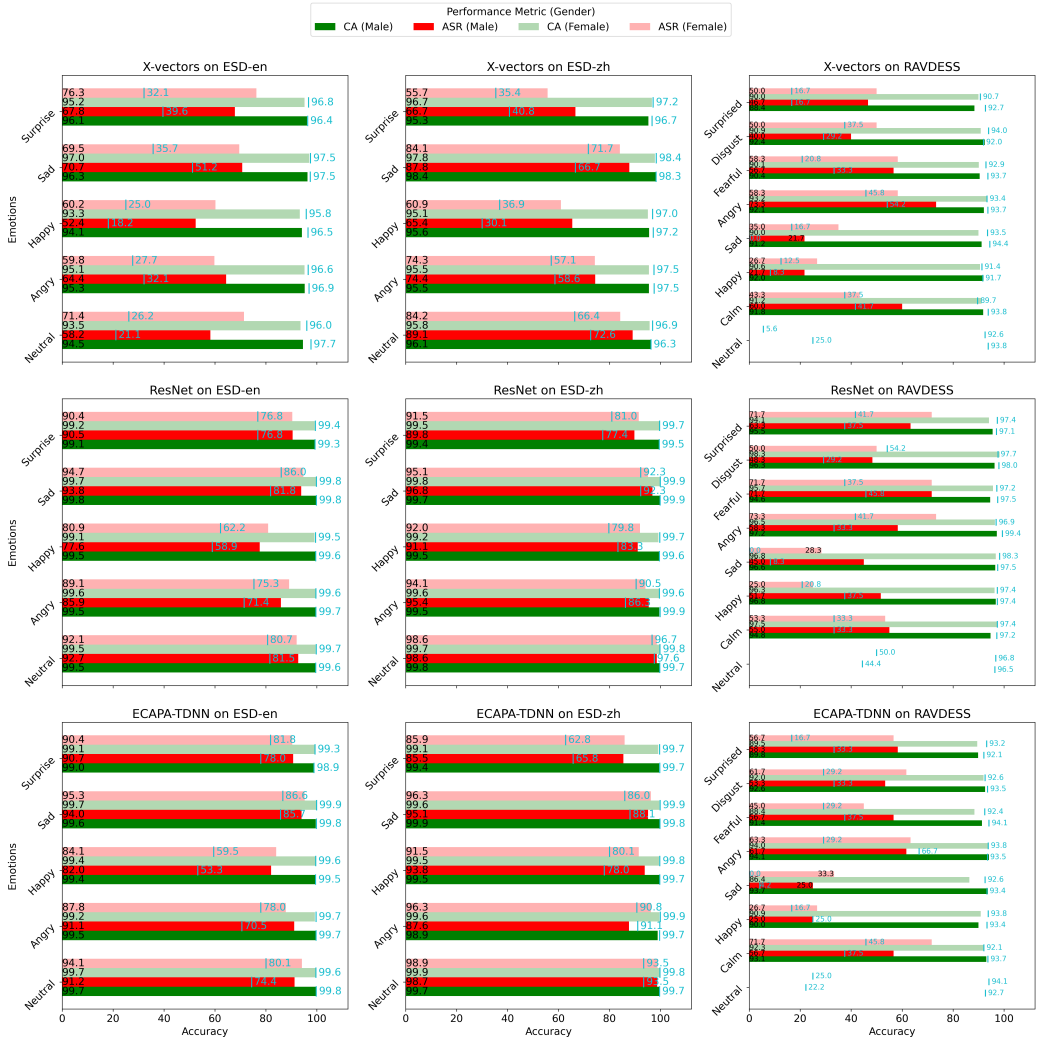


Figure 2: EmoBack's CA and ASR for each combination of targeted DNN, dataset, trigger emotion, and speaker gender. The figure shows the accuracies with the poisoning rate of 10% (colored bars, with the exact percentage in black text at the base of each bar) and 5% (green text). Notice that RAVDESS has no data for Neutral where the poisoning rate = 10 because, prior to pre-processing, it has too few Neutral samples to achieve this poisoning rate. Moreover, in some cases, the 5% poisoning rate ASR was 0. This could be attributed to the low poisoning rate of these setups combined with the small size of RAVDESS and Sad possibly being expressed ineffectively (see Section 6.1.2). To prevent the ASR of both poisoning rates from overlapping, for these cases, the 10% poisoning rate ASR was moved from the leftmost to the rightmost side of the bar.

41.4% can be observed suggesting that increasing the CLR when applying low PRs can effectively reduce the backdoor attack's efficacy without substantially impacting the model's performance.

Models with lower CLRs (< 0.2) maintain higher accuracy, indicating that less invasive pruning can preserve the model's performance on clean data while still mitigating backdoor effects, albeit to a smaller degree. This is particularly evident in ResNet with a Surprise trigger emotion in Figure 3, where the results of lower CLRs exhibit a more gradual degradation in ASR and virtually none in CA, compared to higher CLRs. Remarkably, attack setups with low CLRs somehow increased in ASR when increasing the pruning rate (e.g., ECAPA-TDNN with any emotion using ESD-en). Later convolutional layers tend to extract higher-level features. We assume that emotional prosody may be such a high-level feature relative to SI, meaning that neurons in deeper convolutional layers are, therefore, tasked with trigger recognition. When pruning is performed only on later layers, it may predominantly remove SI-tasked neurons, reducing the model's ability to classify clean samples accurately. However, with fewer neurons remaining, the relative influence of the trigger-recognizing neurons might increase due to reduced competition among neurons, making them more dominant in the final classification, thus increasing the ASR.

6.2.3 Influence Architecture. Overall, for ResNet, the CA appears to decrease more rapidly as PR increases. This could be attributed to ResNet's lower parameter count. For example, a PR of 0.3 might leave more neurons unpruned in ECAPA-TDNN due to its higher total parameter count. As a result, ResNet may have trouble inferring correctly the labels of clean inputs. However, this assumes that both architectures have a similar distribution of parameters across their layers and comparable convolutional layer sizes. Furthermore, differences in layer connectivity, activation functions, and overall network depth could also have influenced the impact of pruning.

6.2.4 Influence of Emotion. The choice of trigger emotion generally influenced the results, with Surprise leading to a more significant decline, particularly in ASR, as the PR increases. For example, in Figure 4 (and to a lesser degree in Figure 3) for the ECAPA-TDNN model, the decrease in ASR is substantially more pronounced for Surprise. As explained in Section 6.2.1, the resilience of Neutral and Sad models may be attributed to the less distinctive nature of the Neutral and Sad emotions when compared to Surprise.

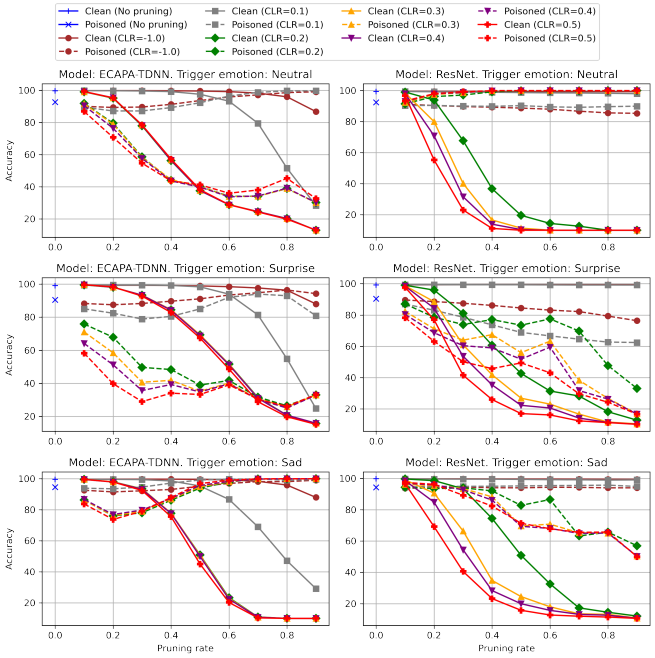


Figure 3: Results of pruning against the best performing models trained on the ESD-en dataset. CLR=−1.0 means that only the final convolutional layer was pruned.

6.3 STRIP-ViTA

The results in Figure 5 indicate a substantial trade-off between FAR and FRR in all the models tested. The data shows that to achieve a low FAR, the FRR must be exceedingly high. This trend is consistent across all architectures and datasets, emotions, and genders. Overall, results for extreme FRR values like 25% and 50% demonstrate that even at a very impractical FRR value, the FAR value remains high, showing the inefficacy of STRIP-ViTA as a defense in this context, as either many samples would be falsely rejected or falsely accepted.

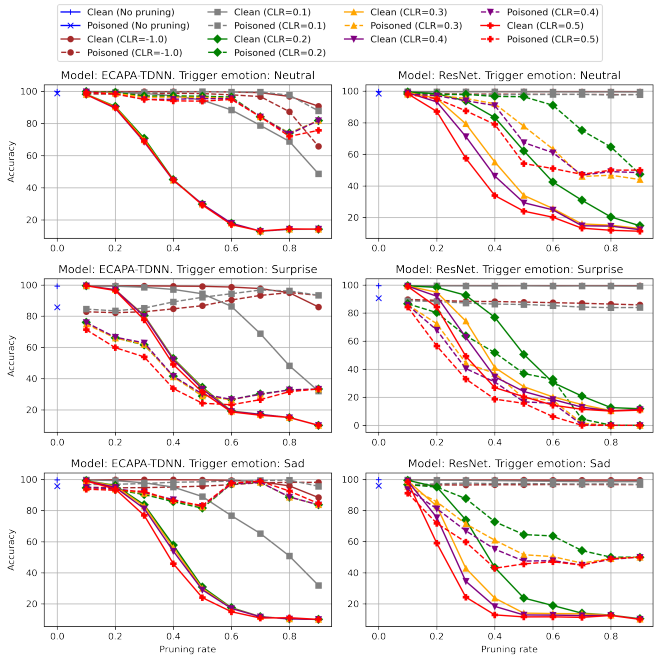


Figure 4: Results of the pruning defense against the best-performing models trained on the ESD-zh dataset.

Both models, when trained using ESD-zh, demonstrated slightly better performance for more combinations of gender and emotion, maintaining a lower FAR at comparable FRR levels compared to their ESD-en counterparts, indicating that models trained with ESD-zh are less robust against STRIP-ViTA. This could be attributed to several factors. First, the distinctive characteristics of Chinese, such as tonal variations and possible cultural differences in emotional expression, might provide more distinct acoustic characteristics, making it easier for STRIP-ViTA to detect anomalies or triggers regardless of the absence or presence of emotion. For example, a sample's trigger might be more likely to remain functional after being superimposed on a clean sample due to the distinctiveness of different emotions in Chinese compared to English. However, the differences in results between dataset languages are almost non-existent for lower FRR values and marginal for more extreme FRR values, so they could be attributed to randomness.

Regarding STRIP-ViTA's efficacy against attacks with different trigger emotions, Sad tended to result in a lower FAR for the same FRR across both datasets. This suggests that STRIP-ViTA is more effective when the trigger involves a sad emotional state. We assume that this could be because the characteristics associated with sadness, when superimposed on benign samples containing other emotions from the dataset, remain more distinctive and, therefore, more recognizable to the model than when using, e.g., Neutral as the trigger. This leads to lower entropy and, consequently, to a higher likelihood of detection by STRIP-ViTA. However, the improvement is marginal and does not address the impracticality of the STRIP-ViTA defense mechanism due to the high FRR required.

We acknowledge that this assumption contradicts the one in Section 6.2.1 and Section 6.2.4; however, without conducting additional experiments that aim to directly determine the levels of salience of different emotions, we cannot definitively determine whether sadness is more or less distinctive. Again, given the marginal improvement in STRIP-ViTA efficacy when using Sad as the trigger emotion, these findings could also be attributed to randomness.

To conclude, there is an inherent inefficacy when STRIP-ViTA is used for backdoored SI models. The requirement for an excessively high FRR to maintain a low FAR indicates that many legitimate inputs would be rejected, compromising the SI's reliability. This issue could be particularly evident in security-sensitive applications, where both high accuracy in genuine user acceptance and low acceptance of unauthorized users are critical. Furthermore, the presence of this phenomenon across different model architectures, languages, genders, and emotional triggers suggests that STRIP-ViTA's limitations are not tied to certain configurations. We believe that it may be better suited to recognize static triggers rather than dynamic ones. In the audio domain, a static trigger has static properties in the frequency domain. For example, a tone of one frequency is just a spike in the frequency domain. Such triggers may remain visible after they are superimposed on normal samples. Dynamic triggers, such as stylistic transformations [18], depend on the sample and do not have static properties. Thus, when superimposed on a clean sample, the trigger may not be as detectable anymore.

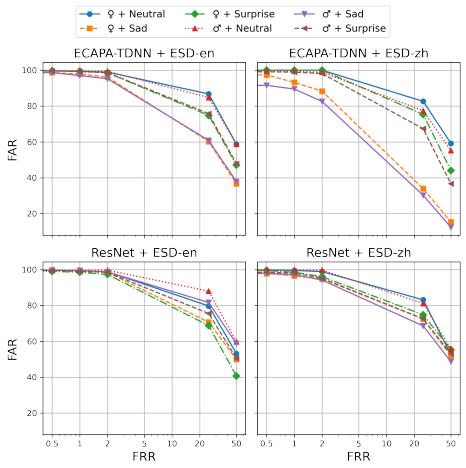


Figure 5: FRR and FAR of our attacks (poisoning rate of 10%) that yielded the highest ASR.

6.4 Pre-processing-based Defense Strategies

6.4.1 Quantizing. In Figure 6, emotion Neutral, it is evident that the Quantize defense exhibits a trend in which increasing the quantization parameter Q leads to a decrease in both CA and ASR for both male and female speakers. Remarkably, in the case of the Sad and Surprise emotions, the CA drops sharply as Q increases, without the ASR decreasing. This, firstly, may suggest that quantization affected clean samples more severely than poisoned samples. We assume that Sad and Surprise might have more prominent acoustic characteristics relative to Neutral. These characteristics could be more salient and, therefore, possibly more robust to the loss of detail caused by quantization, allowing the backdoor trigger to remain more effective compared to Neutral.

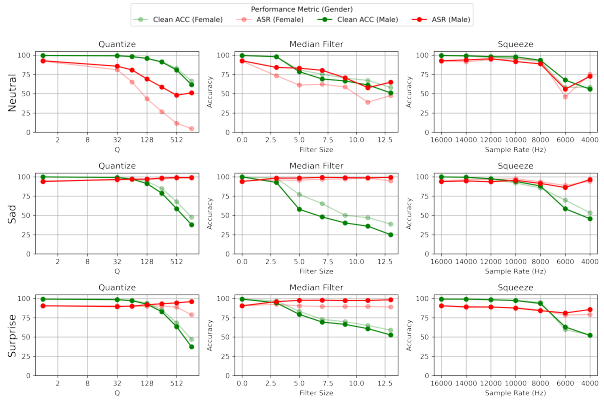
6.4.2 Median Filtering. The median filtering revealed a pattern similar to that of quantizing. As the filter size increases, there is a noticeable reduction in CA and, to a lesser extent, in ASR when the emotion is Neutral. Again, Sad and Surprise show more robustness against the defense, possibly reinforcing our previous findings in Section 6.4.1. However, the effect is substantially less pronounced, mostly observable in Figure 6a, Figure 6c, and Figure 6d.

6.4.3 Squeezing. Squeezing exhibits different effects across combinations of parameters. Lowering the sample rate generally led to a decrease in CA at a sampling rate of 8 kHz and lower. Typical sampling rates used in speech processing range from about 8 kHz upward [5]. Reducing the sampling rate too much may result in a loss of important high-frequency information, which complicates the capture of subtle acoustic features. As a result, the model's performance on clean inputs diminished. Remarkably, CA and/or ASR, in some cases, appear to increase at a sampling rate of 4 kHz, as can, for example, be observed in Figure 6a (Sad), Figure 6c (Sad) and Figure 6d (Sad). We assume that by downsampling audio to rates that are divisions of the original sampling rate (e.g., 4kHz), samples might align such that some samples of the original signal are preserved. This may have caused the model to recognize patterns on which it was trained, leading to a spike in accuracy.

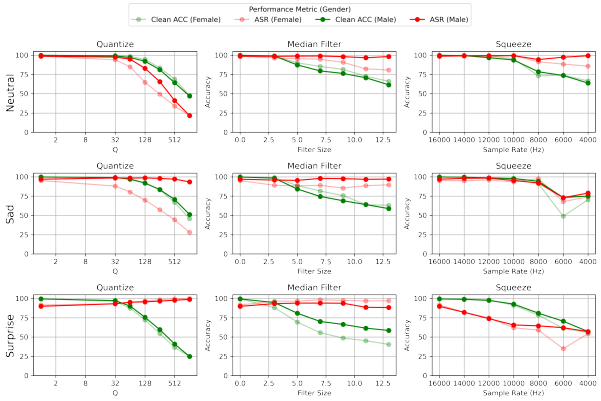
6.4.4 Gender Differences. The results show that there are slight differences in the effectiveness of the defenses between males and females. For example, in almost all quantization results, the female ASR exhibits a more substantial decrease in ASR given a higher value Q. The generally higher pitch and varied dynamic range of female speech may have made backdoor triggers more susceptible to disruption by quantization. Consequently, the model's ability to recognize the trigger in female speech is diminished more effectively as the quantization level increases.

6.4.5 Comparison of Defenses. Across all pre-processing defenses, there is a consistent trade-off between reducing ASR and maintaining CA. Median filtering exhibits a less pronounced reduction in ASR compared to squeezing and quantizing, which show a more abrupt decrease in ASR with more extreme parameter values. Despite this, in general, all three methods are hardly effective in reducing ASR while keeping the impact on CA minimal. Quantizing for Neutral was an effective defense, particularly where the target speaker identity was female, using ESD-en. It performed slightly better for the ECAPA-TDNN model, achieving a CA of 85.07% and an ASR of 11.78%. The original ASR was 94.11%, which is a substantial reduction of 83.33%. Squeezing has shown significant effectiveness against the ECAPA-TDNN model trained on ESD-zh, using Surprise as the trigger emotion and targeting a female speaker. With a sample rate of 8 kHz, the CA was 85.28%, and the ASR was 39.08%, resulting in a 46.84% reduction in ASR. Unfortunately, this effect was less pronounced for the ResNet model.

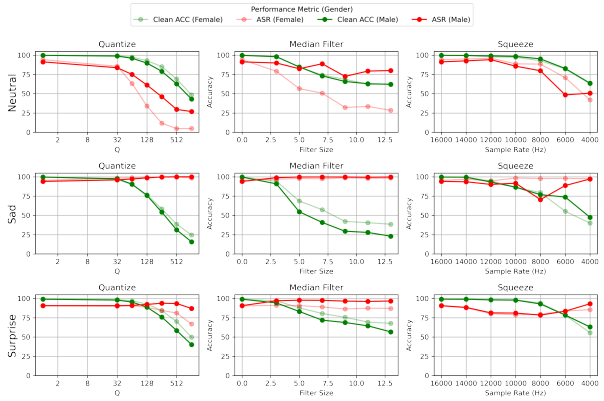
Despite these high reduction rates, they are specific to certain hyperparameter combinations. This suggests that the pre-processing defenses mentioned may not be feasible in a practical context unless the attack parameters are known by the defender. However, for six



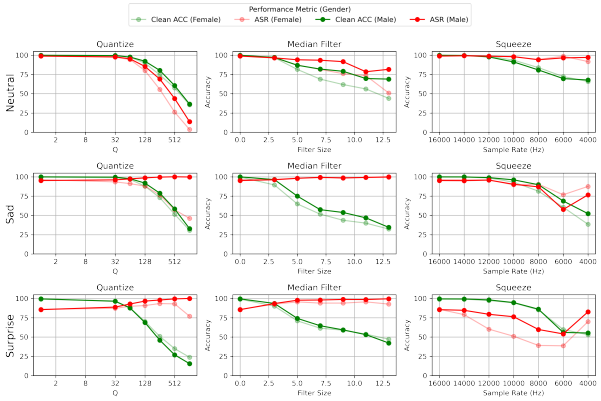
(a) ResNet + ESD-en.



(b) ResNet + ESD-zh.



(c) ECAPA-TDNN + ESD-en.



(d) ECAPA-TDNN + ESD-zh.

Figure 6: This figure illustrates the effectiveness of pre-processing-based defense strategies against our backdoored models. We chose experimental settings that resulted in the highest ASR to evaluate the effectiveness of the defenses against strong attackers. For this reason, the poisoning rate was 10%.

out of twelve combinations of model hyperparameters shown in Figures 3 and 4 (ESD-en: ECAPA-TDNN + Neutral, ECAPA-TDNN + Surprise, ECAPA-TDNN + Sad, ResNet + Surprise; ESD-zh: ECAPA-TDNN + Surprise, ResNet + Surprise), substantial reductions in ASR were observed with a CLR of 0.5 and PRs of 0.1 or 0.2, while CA remained high. Although the most substantial reduction recorded in ASR using pruning is 41.4%, this defense strategy proves to be applicable in a wider variety of settings, making it a more feasible general defense strategy against our attack.

7 Conclusions and Future Work

We introduced EmoBack, a novel backdoor attack targeting stage-wise, closed-set, and text-independent SI DNNs by using emotional prosody as triggers. We evaluated its effectiveness across three datasets (ESD-en, ESD-zh, and RAVDESS), three DNN architectures (X-vectors, ResNet, and ECAPA-TDNN), and against five defenses (pruning, STRIP-ViTA, quantization, median filtering, and squeezing). We showed that emotions like Neutral, Sad, Angry, and Surprise, for datasets ESD-en and ESD-zh, typically yielded a higher ASR. Moreover, the ECAPA-TDNN model was found to be the most vulnerable across various configurations, achieving ASRs up to 98.9% while maintaining a high CA of at least 86.4%. Pruning showed the most promise in mitigating the attack, while STRIP-ViTA and pre-processing techniques varied in their effectiveness.

Future research could focus on several areas to further enhance our understanding and defense against EmoBack. As pruning yielded promising results, we should investigate if fine-tuning further improves the defense's performance. Additionally, future research could explore methods that manipulate the prosody in neutral samples and induce target emotions programmatically, making the attack applicable in more real-world scenarios. However, evaluating the authenticity of these transformations is crucial, and metrics such as human perceptual tests should be used for validation. Moreover, the high poisoning rates used in our experiments (5-10%) may be challenging in a practical setting. Therefore, we should investigate the effectiveness of lower poisoning rates or consider a clean-label attack [35] that is stealthier.

References

[1] 2022. Emotional voice conversion: Theory, databases and ESD. *Speech Communication* 137 (2022), 1–18.

[2] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670* (2019).

[3] Eugene Bagdasaryan and Vitaly Shmatikov. 2021. Blind backdoors in deep learning models. In *30th USENIX Security Symposium (USENIX Security 21)*. 1505–1521.

[4] Zhongxin Bai and Xiao-Lei Zhang. 2021. Speaker recognition based on deep learning: An overview. *Neural Networks* 140 (2021), 65–99. https://doi.org/10.1016/j.neunet.2021.03.004

[5] Homayoon Beigi. 2011. *Fundamentals of Speaker Recognition*. Springer. https://doi.org/10.1007/978-0-387-77592-0

[6] Joseph P Campbell, Wade Shen, William M Campbell, Reva Schwartz, Jean-Francois Bonastre, and Driss Matrouf. 2009. Forensic speaker recognition. *IEEE Signal Processing Magazine* 26, 2 (2009), 95–103.

[7] Jiangyi Deng, Yanjiao Chen, and Wenyuan Xu. 2022. FenceSitter: Black-box, Content-Agnostic, and Synchronization-Free Enrollment-Phase Attacks on Speaker Recognition Systems. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security* (Los Angeles, CA, USA) *(CCS '22)*. Association for Computing Machinery, New York, NY, USA, 755–767. https://doi.org/10.1145/3548606.3559357

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[9] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Proc. Interspeech 2020*. International Speech Communication Association, Shanghai, China, 3830–3834. https://doi.org/10.21437/Interspeech.2020-2650

[10] Jane Edwards, Henry J Jackson, and Philippa E Pattison. 2002. Emotion recognition via facial expression and affective prosody in schizophrenia: a methodological review. *Clinical psychology review* 22, 6 (2002), 789–832.

[11] Yansong Gao, Yeonjae Kim, Bao Gia Doan, Zhi Zhang, Gongxuan Zhang, Surya Nepal, Damith C. Ranasinghe, and Hyoungshick Kim. 2022. Design and Evaluation of a Multi-Domain Trojan Detection Method on Deep Neural Networks. *IEEE Transactions on Dependable and Secure Computing* 19, 4 (2022), 2349–2364. https://doi.org/10.1109/TDSC.2021.3055844

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

[13] Hanqing Guo, Xun Chen, Junfeng Guo, Li Xiao, and Qiben Yan. 2023. MASTERKEY: Practical Backdoor Attack Against Speaker Verification Systems. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. Association for Computing Machinery, New York, NY, USA, Article 48, 15 pages. https://doi.org/10.1145/3570361.3613261

[14] Wei Guo, Benedetta Tondi, and Mauro Barni. 2022. An Overview of Backdoor Attacks Against Deep Neural Networks and Possible Defences. *IEEE Open Journal of Signal Processing* 3 (2022), 261–287. https://doi.org/10.1109/OJSP.2022.3190213

[15] Sanghyun Hong, Nicholas Carlini, and Alexey Kurakin. 2022. Handcrafted backdoors in deep neural networks. *Advances in Neural Information Processing Systems* 35 (2022), 8068–8080.

[16] INTERPOL. 2024. Speaker Identification Integrated Project (SIIP). https://www.interpol.int/Who-we-are/Legal-framework/Information-communications-and-technology-ICT-law-projects/Speaker-Identification-Integrated-Project-SIIP. Accessed: 2024-06-08.

[17] Muhammad Mohsin Kabir, M. F. Mridha, Jungpil Shin, Israt Jahan, and Abu Quwsar Ohi. 2021. A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities. *IEEE Access* 9 (2021), 79236–79263. https://doi.org/10.1109/ACCESS.2021.3084299

[18] Stefanos Koffas, Luca Pajola, Stjepan Picek, and Mauro Conti. 2023. Going in Style: Audio Backdoors Through Stylistic Transformations. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Institute of Electrical and Electronics Engineers (IEEE), Rhodes Island, Greece, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096332

[19] Stefanos Koffas, Jing Xu, Mauro Conti, and Stjepan Picek. 2022. 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) *(WiseML '22)*. Association for Computing Machinery, New York, NY, USA, 57–62. https://doi.org/10.1145/3522783.3529523

[20] Xinfeng Li, Junning Ze, Chen Yan, Yushi Cheng, Xiaoyu Ji, and Wenyuan Xu. 2023. Enrollment-Stage Backdoor Attacks on Speaker Recognition Systems via Adversarial Ultrasound. *IEEE Internet of Things Journal* PP (01 2023), 1–1. https://doi.org/10.1109/JIOT.2023.3328253

[21] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2018. Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks. arXiv:1805.12185 [cs.CR]

[22] Qiang Liu, Tongqing Zhou, Zhiping Cai, and Yonghao Tang. 2022. Opportunistic backdoor attacks: Exploring human-imperceptible vulnerabilities on speech recognition systems. In *Proceedings of the 30th ACM International Conference on Multimedia*. Association for Computing Machinery (ACM), New York, NY, United States, 2390–2398.

[23] Steven R. Livingstone and Frank A. Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE* 13, 5 (05 2018), 1–35. https://doi.org/10.1371/journal.pone.0196391

[24] Yuxiao Luo, Jianwei Tai, Xiaoqi Jia, and Shengzhi Zhang. 2022. Practical backdoor attack against speaker recognition system. In *Information Security Practice and Experience*. Springer, Taipei, Taiwan, 468–484.

[25] Dan Meng, Xue Wang, and Jun Wang. 2023. Backdoor Attack Against Automatic Speaker Verification Models in Federated Learning. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Institute of Electrical and Electronics Engineers (IEEE), Rhodes Island, Greece, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10094675

[26] Geoffrey Stewart Morrison, Farhan Hyder Sahito, Gaëlle Jardine, Djordje Djokic, Sophie Clavet, Sabine Berghs, and Caroline Goemans Dorny. 2016. INTERPOL survey of the use of speaker identification by law enforcement agencies. *Forensic Science International* 263 (2016), 92–100. https://doi.org/10.1016/j.forsciint.2016.03.044

[27] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A General-Purpose Speech Toolkit. arXiv:2106.04624 [eess.AS] arXiv:2106.04624.

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[29] Mickael Rouvier and Pierre-Michel Bousquet. 2021. Studying Squeeze-and-Excitation Used in CNN for Speaker Verification. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 1110–1115. https://doi.org/10.1109/ASRU51503.2021.9687936

[30] Cong Shi, Tianfang Zhang, Zhuohang Li, Huy Phan, Tianming Zhao, Yan Wang, Jian Liu, Bo Yuan, and Yingying Chen. 2022. Audio-domain position-independent backdoor attack via unnoticeable triggers. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking* (Sydney, NSW, Australia) *(MobiCom '22)*. Association for Computing Machinery, New York, NY, USA, 583–595. https://doi.org/10.1145/3495243.3560531

[31] Nirupam Shome, Anisha Sarkar, Arit Kumar Ghosh, Rabul Hussain Laskar, and Richik Kashyap. 2023. Speaker Recognition through Deep Learning Techniques: A Comprehensive Review and Research Challenges. *Periodica Polytechnica Electrical Engineering and Computer Science* 67, 3 (2023), 300–336. https://doi.org/10.3311/PPee.20971

[32] Nilu Singh, Raees Ahmad Khan, and Raj Shree. 2012. Applications of Speaker Recognition. *Procedia Engineering* 38 (2012), 3122–3126. https://api.semanticscholar.org/CorpusID:109086245

[33] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-Vectors: Robust DNN Embeddings for Speaker Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Institute of Electrical and Electronics Engineers (IEEE), Calgary, Alberta, Canada, 5329–5333. https://doi.org/10.1109/ICASSP.2018.8461375

[34] Yu Tang, Lijuan Sun, and Xiaolong Xu. 2024. SilentTrig: An imperceptible backdoor attack against speaker identification with hidden triggers. *Pattern Recognition Letters* 177 (2024), 103–109. https://doi.org/10.1016/j.patrec.2023.12.002

[35] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. 2018. Clean-label backdoor attacks. (2018).

[36] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. 2014. Deep neural networks for small footprint text-dependent speaker verification. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Institute of Electrical and Electronics Engineers (IEEE), Florence, Italy, 4052–4056. https://doi.org/10.1109/ICASSP.2014.6854363

[37] Jianbin Ye, Xiaoyuan Liu, Zheng You, Guowei Li, and Bo Liu. 2022. DriNet: Dynamic Backdoor Attack against Automatic Speech Recognization Models. *Applied Sciences* 12, 12 (2022). https://doi.org/10.3390/app12125786

[38] Junning Ze, Xinfeng Li, Yushi Cheng, Xiaoyu Ji, and Wenyuan Xu. 2023. UltraBD: Backdoor Attack against Automatic Speaker Verification Systems via Adversarial Ultrasound. In *2022 IEEE 28th International Conference on Parallel and Distributed Systems (ICPADS)*. 193–200. https://doi.org/10.1109/ICPADS56603.2022.00033

[39] Tongqing Zhai, Yiming Li, Ziqi Zhang, Baoyuan Wu, Yong Jiang, and Shu-Tao Xia. 2021. Backdoor Attack against Speaker Verification. arXiv:2010.11607 [cs.CR]

[40] Tianfang Zhang, Huy Phan, Zijie Tang, Cong Shi, Yan Wang, Bo Yuan, and Yingying Chen. 2024. Inaudible Backdoor Attack via Stealthy Frequency Trigger Injection in Audio Spectrogram. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking* (Washington D.C., DC, USA) *(ACM MobiCom '24)*. Association for Computing Machinery, New York, NY, USA, 31–45. https://doi.org/10.1145/3636534.3649345

[41] Haodong Zhao, Wei Du, Junjie Guo, and Gongshen Liu. 2023. A Universal Identity Backdoor Attack against Speaker Verification based on Siamese Network. arXiv:2303.16031 [cs.CR]

[42] Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 2021. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 920–924.