Time Shuffle: A Transferability-Booster for Multiple Audio Adversarial Tasks

JiaCheng Deng [1], Dengpan Ye[*] [1,2], Yuhong Liu[†] [1], Zhaolin Wei [1], Ziyi Liu [1], Haoran Duan [1]

[1]Wuhan University, School of Cyber Science and Engineering, Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, Wuhan, 430072, China
[2]Cyberspace Institute of Advanced Technology, Guangzhou University, Guangdong, 510006, China
{dengjiacheng,yedp,2024282210189,wzl_chunzhen,ziyi_liu,hrduan}@whu.edu.cn

Abstract

Existing audio adversarial attack methods suffer from poor transferability, primarily due to insufficient exploration of model decision mechanisms and overreliance on heuristic-driven algorithm design. This paper aims to alleviate this gap. Specifically, through observations across three mainstream audio tasks (Automatic Speech Recognition, Speaker Verification, and Keyword Spotting), we reveal that these models primarily rely on local temporal features—inputs with time shuffled retain 83.7% of original accuracy. The SHAP-based visualization further validated that time shuffle leads to a significant shift in the salient regions of the model, but the samples can still be correctly identified, indicating the presence of redundant features that can affect decision-making. Inspired by these findings, we propose Time-Shuffle (TS) adversarial attack (including segments-based TS and phoneme-level-based TS-p). This method divides audio or phonemes into segments, randomly shuffles them, and computes gradients on the shuffled structure. By forcing perturbations to exploit transferable local temporal features and reduce overfitting to source-specific patterns, TS/TS-p inherently enhances transferability. As a model-agnostic framework, TS/TS-p can seamlessly integrates with existing attack methods. Comprehensive experiments demonstrate that TS-p achieved SOTA and boosts transferability by about 23%/14.7%/6.3% on ASR/ASV/KWS.

Introduction

Speech interaction technology, as a core application of artificial intelligence, has been deeply integrated into key scenarios such as intelligent assistants, voice payments, and security authentication, becoming one of the most natural and essential interfaces for human-machine interaction. Not only facilitates more natural and convenient interactions, but it also carries critical user interests, including privacy.

Voice systems driven by deep neural networks (DNNs), while powerful, exhibit significant vulnerabilities. Researchers found that injecting imperceptible perturbations into audio inputs—known as adversarial examples (Carlini and Wagner 2018)—can deceive systems into incorrect output. Attacks are categorised as white-box (attackers access model structure/parameters) or black-box (no model access), with the latter more relevant to real-world use. For example, attackers can make voice recognition misinterpret "pay 100 dollars" as "pay 1000 dollars" (causing financial loss) or forge voice features to impersonate users (compromising privacy). Such threats hinder large-scale deployment of intelligent voice tech

Existing research has extensively studied adversarial example transferability in computer vision (Jia et al. 2022) and NLP (Wang et al. 2024), but speech domain explorations remain limited. Much current work focuses on improving transferability via multi-model ensembles (Fang et al. 2024; Chen, Li, and Chen 2024)—effective yet costly, and failing to boost intrinsic transferability. Single-model-based methods to improve transferability could deepen understanding of audio adversarial examples' nature. Research on single-model-based methods to improve transferability can advance the fundamental understanding of the nature of audio adversarial examples.

Some use speech data augmentation to boost adversarial transferability: noise injection during computation (Kim, Park, and Lee 2023), random time interval masking (Bui et al. 2024), audio scaling for dynamic input (Xie et al. 2019), and voice activity detection to filter regions for localized attacks (Gao et al. 2024). While these approaches improve transferability to a certain extent, they fail to incorporate speech-specific characteristics. Others explore model-level tweaks (Chen et al. 2025), like fine-tuning to improve transferability. Some study task-specific transferability, e.g., using multi-speaker speech to boost speaker recognition attacks (Chen et al. 2022). Yet existing methods mainly improve transferability via ensembles, model changes, or traditional speech enhancement—lacking in-depth exploration of speech models' operations.

The operating mechanism of the model is mainly determined by the feature patterns learned by the model. From the perspective of transferability, the features learned by the model can be divided into two categories: one is non-transferable features, which overfit the surrogate model or source voice and thus fail to generalize to black-box models; the other is transferable features, which capture class commonalities and influence model decisions across different models. Theoretically, reducing the perturbations' overfitting to the source voice and surrogate model while en-

[*]Corresponding Author
[†]Equal Contribution Author.

Input	WER↓ of ASR Task		
	wav2vec	conformer	crdnn
Normal	0.020	0.020	0.029
TS-3	–	–	–
TS-5	–	–	–
TS-p3	0.20(81%)	0.24(77%)	0.47(54%)
TS-p5	0.20(81%)	0.30(71%)	0.59(42%)
Input	Cosine Similarity ↑ of ASV Task		
	hubert	ecapa-tdnn	resnet
Normal	1.000	1.000	1.000
TS-3	0.84(84%)	0.99(99%)	0.98(98%)
TS-5	0.83(83%)	0.98(98%)	0.97(97%)
TS-p3	0.73(73%)	0.79(79%)	0.76(76%)
TS-p5	0.71(71%)	0.72(72%)	0.66(66%)
Input	Accuracy ↑ of KWS Task		
	bc_resnet	wav2_kws	kwt
Normal	0.955	0.863	0.997
TS-3	0.88(92%)	0.81(94%)	0.98 (98%)
TS-5	0.72(75%)	0.73(76%)	0.98(98%)
TS-p3	0.72(75%)	0.71(85%)	0.94(94%)
TS-p5	0.59(62%)	0.62(72%)	0.89(89%)

Table 1: The performance of the model on audio that has undergone time shuffle (TS) and phoneme-level time shuffle (TS-p), where 3/5 represents being divided into 3/5 segments. Normal means this input is the original audio. The relative performance ratio (%) is in brackets. Metrics: Word Error Rate (WER), Cosine Similarity (CS), and Predict Accuracy (Accuracy). The next line of each task represents the model name (*e.g.*, wav2vec, hubert and bc_resnet).

hancing the learning of transferable features can improve the transferability of adversarial examples.

To this end, we first explored the decision mechanisms of existing speech models. We analyzed three mainstream speech tasks: speaker verification (ASV), keyword spotting (KWS), and automatic speech recognition (ASR). As shown in Table 1, when predicting time-shuffled samples, the performance decrease of most models was limited. Based on this observation, we conclude that the decision mechanisms of existing models are dominated by local temporal features. Further, by comparing the SHAP values before and after time shuffle, we observed a significant change in the contribution area of the spectrogram, but the model still predicted correctly, indicating the presence of redundant features in the audio. Consequently, existing speech adversarial attacks, which directly generate perturbations based on the source audio or model, may overfit the source sample and fail to utilize all features influencing model decisions, leading to low transferability. We propose a novel audio adversarial attack that employs multiple time shuffle operations to force adversarial perturbations to focus on local temporal features and redundant decision features, thereby enhancing the transferability of audio adversarial examples.

In summary, we highlight our key contributions as follows:

- We reveal the common decision-making mechanisms of existing speech models across three mainstream tasks, *i.e.*, their reliance on local temporal features.
- Using SHAP analysis, we demonstrate the presence of redundant features in the original data that influence model decisions.
- We propose a TS and TS-p adversarial attack, which can be integrated with existing approaches. The empirical evaluation of three main tasks and nine models shows that TS-p achieves SOTA and can improve transferability by about 23%/14.7%/6.3% on ASR/ASV/KWS.

Related Work

Automatic Acoustic System

Automatic Acoustic Systems serve as the core backbone of speech interaction technology. We primarily study three typical tasks: Automatic Speech Recognition (ASR) (Gulati et al. 2020; Baevski et al. 2020; Ravanelli et al. 2021), Speaker Verification (ASV) (Desplanques, Thienpondt, and Demuynck 2020; Hsu et al. 2021; Villalba et al. 2020), and Keyword Spotting (KWS) (Berg, O'Connor, and Cruz 2021; Kim et al. 2021; Seo, Oh, and Jung 2021). ASR systems transform continuous speech signals into acoustic features (*e.g.*, Mel-spectrograms) via front-end signal processing (*e.g.*, frame segmentation, windowing, and Fourier transform), further extract high-level features using acoustic models, and finally decode these features into text sequences through language model decoding, finding widespread applications in scenarios such as intelligent assistants and speech-to-text transcription. ASV systems typically start with preprocessing to extract speaker-related vocal tract features (*e.g.*, d-vectors or x-vectors), then discriminate between different speaker identities via metric learning (*e.g.*, cosine similarity or PLDA scoring), primarily used for security authentication and personalized services. KWS systems, on the other hand, employ sliding window-based speech segment extraction to capture time-frequency features, utilize classification models to determine the presence of target keywords, and serve as a critical entry point for device wake-up or triggering subsequent interactions.

Audio Adversarial Attack

Audio adversarial attacks inject imperceptible small perturbations into audio signals to mislead automatic acoustic systems into generating erroneous outputs, primarily serving to evaluate the security robustness of acoustic systems (*e.g.*, forging speech to deceive authentication, altering automatic speech recognition outcomes). These attacks are categorized into two scenarios based on attack conditions: white-box attacks (Carlini and Wagner 2018), where attackers have full access to model parameters and directly generate tailored perturbations; and black-box attacks, where attackers only have access to the model's input-output interface. Black-box attacks further subdivide into two types: transfer-based attacks (Fang et al. 2024; Ge et al. 2023), which generate adversarial examples via surrogate models and exploit the cross-model generalization capability of perturbations to indirectly attack the target; and query-based attacks (Tong et al. 2023), which iteratively adjust perturbations based on

interaction feedback with the target. Among these, transfer-based attacks are more practically valuable due to their reduced need for target interaction and lower implementation cost. This work focuses on enhancing the black-box transferability of audio adversarial examples to strengthen their practical attack capability.

SHapley Additive exPlanations

SHAP (SHapley Additive exPlanations) (Lundberg and Lee 2017) is a game-theoretic model interpretability analysis tool whose core function is to quantify the contribution values of input features to a model's prediction results (*i.e.*, SHAP values), thereby uncovering the intrinsic decision logic and critical features of the model. Its capabilities are primarily manifested in three aspects: First, evaluating global feature importance by computing the mean SHAP values across all samples to identify core features that most influence model predictions; second, explaining local decision rationales by calculating SHAP values for individual samples to decompose their prediction results into marginal contributions of each feature, clarifying the decision logic for specific inputs; third, detecting model potential biases by analyzing the distribution of feature contributions to identify issues such as over-reliance on specific features (*e.g.*, speaker identity in speech) or fairness concerns.

Methodology

We first reveal the inherent decision-making logic of three mainstream audio tasks(KWS/ASV/ASR). Then we use SHAP analysis to identify critical features. Finally, based on the above analyses, we designed Time Shuffle to enhance transferability.

Preliminaries

This section defines the mathematical description of audio adversarial attacks and then introduces the task objectives and corresponding loss functions of three typical automatic acoustic systems: KWS, ASV, and ASR.

Adversarial Attack. Consider an acoustic model F with parameters θ, given a audio input $x \in R^{1\times D}$ with ground-truth label y. Attackers aim to inject a tiny perturbation p such that $F(x+p) \neq y$. Attackers typically generate p by maximizing the loss function, which can be formalized as:

$$x^{adv} = \underset{\|p<\epsilon\|_\infty}{\arg\max} L(x+p, y; \theta), \tag{1}$$

where ϵ bounds the magnitude of p, and $L(\cdot)$ denotes the corresponding loss function (*e.g.*, CTC-Loss, cross-entropy, cosine similarity). Next, we introduces a representative attack method, I-FGSM (Carlini and Wagner 2018), which iteratively applies small perturbations to update perturbation p^t and generate adversarial examples $x + p^T$:

$$p^t = p^{t-1} + \alpha \cdot sign(\nabla_{p^{t-1}} L(x + p^{t-1}, y; \theta). \tag{2}$$

where $sign$ is sign function and T is the number of iterations. While effective in improving white-box attack success rates, this method suffers from poor transferability. Most existing gradient-based adversarial attacks extend based on this framework.

Next, we will formalize the three types of automatic acoustic systems involved and provide corresponding loss functions.

Keyword Spotting. The goal of keyword spotting (KWS) is to classify input speech commands into one of a predefined set of limited classes. Given an input x and its corresponding label y, along with a KWS model F^{kws}. Attackers intend to cause misclassification, and thus the loss function can be formalized as:

$$L(x+p, y; \theta) = \mathcal{H}\left(F^{kws}(x+p), y\right), \tag{3}$$

where $\mathcal{H}(\cdot)$ denotes the cross-entropy function.

Automatic Speaker Verification. ASV aims to confirm whether a speaker possesses legitimate access rights, typically involving enrollment and verification phases. In the enrollment phase, a speaker's voice x and identity y are used to extract and store speaker embeddings $F^{asv}(x)$ via an ASV model. During verification, the system compares the extracted embedding of a test utterance $x+p$ with the stored embedding. Attackers seek to bypass verification by minimizing the cosine similarity between the adversarial embedding and the genuine embedding, leading to the loss function:

$$L(x+p, y; \theta) = 1 - \frac{F^{asv}(x+p) \cdot F^{asv}(x)}{\|F^{asv}(x+p)\|_2 \, \|F^{asv}(x)\|_2}, \tag{4}$$

where $\cdot$ represents the dot product, and $\|\cdot\|_2$ denotes the L2 norm.

Automatic Speech Recognition. ASR converts spoken language into text. Given an input x and its ground-truth transcription y, an ASR model F^{asr} outputs a probability distribution over possible text sequences. Attackers aim to cause the ASR model to fail in correct recognition, typically employing the Connectionist Temporal Classification (CTC) loss to quantify the discrepancy between predicted outcomes and the target transcription. This loss aligns audio frames with text labels without requiring explicit frame-level annotations, formulated as:

$$L(x+p, y; \theta) = \text{CTC-Loss}\left(F^{asr}(x+p), y\right). \tag{5}$$

Model Temporal Sensitivity Analysis

To reveal the decision-making mechanisms of existing audio models, this study conducts systematic experiments by introducing a controlled temporal shuffle. As shown in Table 1, when time shuffle and phoneme-level time shuffle with n-segments (TS-n/TS-pn) are applied to audio inputs, the majority of models across three mainstream tasks—ASR, ASV, and KWS—still maintain over 70% of their original performance. Detailed analyses are as follows:

Global Temporal Insensitivity. In KWS and ASV tasks, the TS-3 (splitting voice into 3 segments and randomly rearranging them) causes only a 1$\sim$17% drop in model performance. Even under more severe TS-5, it still maintains an average accuracy of 87.3%. This indicates that the decision-making mechanisms of KWS/ASV do not rely on absolute temporal positions.

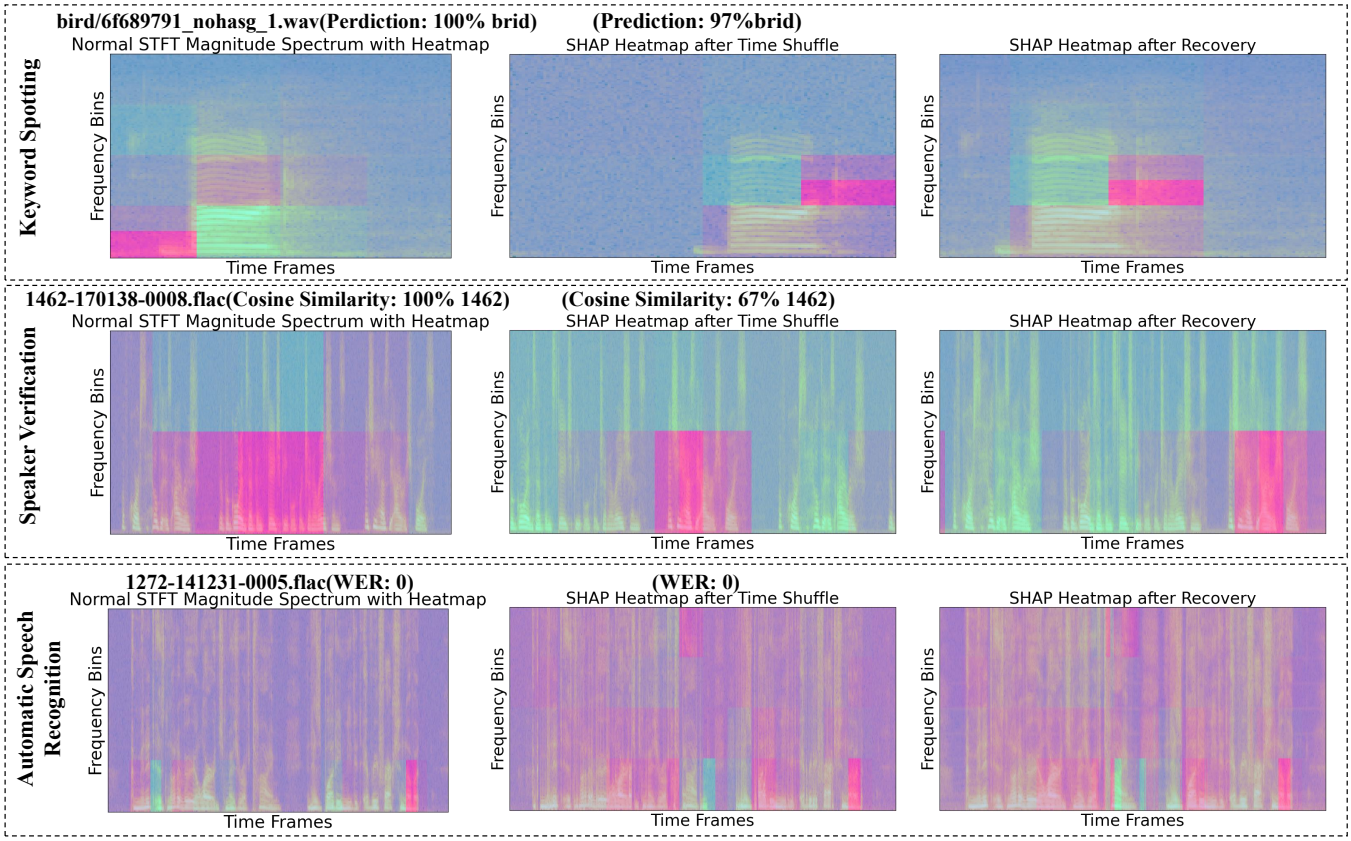


Figure 1: SHAP analysis across three mainstream acoustic tasks, where purple denotes high attribution significance and blue signifies low attribution significance.

Phoneme-Level Time Shuffle. Given the high dependency of ASR systems on phoneme sequences, we further designs phoneme-level time shuffle (TS-pn), which shuffles the temporal order within phonemes while preserving the natural boundaries of phonemes (*e.g.*, spectral transitions between phonemes). Observations reveal that under the TS-p3 setting, performance of ASV and KWS drops to 75%$\sim$98%, whereas the average performance of ASR models (wav2vec/conformer) remains at 79% (excluding crdnn). In the TS-p5 setting, performance in all three tasks is significantly decreased (with an average of 69.5% for ASR/ASV/KWS). Notably, audio processed with TS-p loses linguistic intelligibility (due to disrupted intra-phoneme continuity), yet models still achieve 69.5%(TS-p5) and 77.1%(TS-p3) performance. This suggests that models can compensate for phoneme-level misalignment through sub-phoneme-level features.

In summary, the decision-making mechanisms of existing models primarily rely on sub-phoneme-level local temporal features.

SHAP Analysis

Building upon the dependency of audio models on local temporal features identified in the above section, this study further conducts a SHAP-based interpretability analysis on three representative models: bc_resnet(KWS), ecapa-tdnn(ASV), and wav2vec(ASR).

As shown in Figure 1, time-shuffling operations induce significant shifts in the model's contribution distribution. For KWS models, we input audio labeled as "Bird" (phonetically transcribed as /br/, /I/, /d/) into the original model. The SHAP heatmap reveals that contributions to classification primarily stem from the consonant cluster /br/ and vowel /I/, with minimal contributions from the terminal consonant /d/. This aligns with intuitive expectations, as other categories (*e.g.*, "bed") share the same terminal consonant /d/. After TS, the contribution regions shift toward the vowel /I/ and the terminal consonant /d/. This observation suggests the model compensates for temporal perturbations by strengthening reliance on discriminative transitions between /I/ and /d/ rather than /br/ and /I/. For ASV tasks (second row), TS induces a redistribution of attention from mid-audio segments to the terminal regions. Similarly, in ASR tasks subjected to TS-p3 processing, significant shifts in SHAP's spatiotemporal distributions are observed. Collectively, these findings validate the universality of this phenomenon across diverse audio tasks and indicate that acoustic models exploit redundant local features to compensate for global temporal perturbations.

Notably, despite the drastic transformation of SHAP value

distributions, all models maintain correct prediction performance (see the label in Figure 1). This phenomenon fundamentally reveals the existence of redundant features in audio data - multiple subsets of features can independently support accurate model decisions. These findings not only deepen our understanding of the robustness of audio models, but also lay a theoretical foundation for designing adversarial attack strategies that utilize this redundancy.

Time Shuffle Adversarial Attack (TS Attack)

Motivation. Existing audio adversarial attacks suffer from limited transferability due to overfitting to the source sample and the surrogate model. Through systematic analysis of decision dynamics in acoustic models, we identify two critical factors motivating the proposed TS attack:

1. **Locality Dominance in Decision-Making:** As shown in Table 1, time-shuffled inputs (TS-3/TS-p3) retain 94.1/77.1% original accuracy. Experiments reveal that key discriminative features in audio tasks are concentrated within sub-phonemic temporal features. This locality characteristic implies that adversarial perturbations can bypass source-specific temporal dependencies by focusing on transferable sub-phonemic features.
2. **Redundancy-Driven Transferability Gap:** SHAP analysis (Figure 1) uncovers a decoupling between feature importance and predictive performance: time-shuffling redistributes SHAP value distributions while maintaining high accuracy. This reveals the existence of *redundant features*—multiple feature subsets can drive correct predictions. By disrupting the original temporal feature, the TS attack exploits this redundancy to amplify transferable perturbation.

Algorithm Design. To address the above challenges, we propose a novel adversarial perturbation generation framework: Define the time-shuffle operator $\tau(\cdot)$ parameterized by segment number n and permutation seed $\mathcal{S}$:

$$\tau(x, n, \mathcal{S}) = \text{Concat}(\{x_{seg} \mid seg \in \mathcal{P}_n(\mathcal{S})\})$$

where $\mathcal{P}_n(\mathcal{S})$ denotes a random permutation of n segments derived from input audio x. Two variants are implemented: - $\tau^{seg}(x, n, \mathcal{S})$: Segment-level shuffling (arbitrary segmentation and shuffling). - $\tau^{pho}(x, n, \mathcal{S})$: Phoneme-level shuffling (retain phoneme order and shuffle internally).

Gradient Aggregation Strategy To mitigate performance degradation caused by time shuffle, we introduce a *multi-gradient accumulation* mechanism:

$$g = \nabla_p L\left(x + p, y; \theta\right) + w \cdot \sum_{m=1}^{M} \nabla_p L\left(\tau(x + p, n, \mathcal{S}), y; \theta\right)$$

where p represents adversarial perturbation, w is weight, and M controls the number of shuffle operation. This strategy ensures: (1) Captures invariant features across multiple time shuffle operations. (2) Accumulate gradients to reduce the potential variance caused by time shuffle. The specific algorithm steps can be found in Algorithm 1.

Algorithm 1: Time Shuffle Adversarial Attack

Parameter: Input audio x with ground-truth label y. The max number of iterations T, allowed perturbation magnitude ϵ, step size α, the number of segments n, the times of shuffle M,the weight w.

Output: an audio adversarial example.

Initialize step size $\alpha \leftarrow \epsilon/T$, perturbation $p_0 \leftarrow 0$.
for $t = 1...T$ **do**
  $g \leftarrow \nabla_{p_t} L(x + p_t, y; \theta)$;
  for $m = 1...M$ **do**
    Generate time shuffled audio $\tau^{seg}(x + p_t, n, \mathcal{S})$ or $\tau^{pho}(x + p_t, n, \mathcal{S})$;
    $g \leftarrow g + w \cdot \nabla_{p_t} L(\tau(x + p_t, n, \mathcal{S}), y; \theta)$;
  end for
  $p_t \leftarrow \text{Clip}_{\epsilon}\{p_t + \alpha \cdot sign(g)\}$
end for
return $\text{Clip}_{-1,1}\{x + p_T\}$

Experiments

Experimental Setting

Acoustic Models: We set out to assess the efficacy of adversarial attacks across three prominent acoustic systems: Automatic Speaker Verification (ASV), Keyword Spotting (KWS), and Automatic Speech Recognition (ASR). For the KWS task, we selected three representative models: bc-resnet (Kim et al. 2021), kwt (Berg, O'Connor, and Cruz 2021), and wav2-kws (Seo, Oh, and Jung 2021)——all trained on the Google Speech Command V2 (Warden 2018) dataset. For the ASR, we examined conformer (Gulati et al. 2020), wav2vec (Baevski et al. 2020), and crdnn-Trans, which were trained on LibriSpeech (Panayotov et al. 2015). For ASV evaluation, we selected ecapa-tdnn (Desplanques, Thienpondt, and Demuynck 2020), hubert (Hsu et al. 2021), and resnet-tdnn(Zeinali et al. 2019), all trained using data from the VoxCeleb (Nagrani et al. 2020) dataset.

Dataset: The datasets employed for the three tasks are detailed as follows: we selected 350 utterances with 35 classes from Google Speech Command V2 for KWS, 300 utterances from LibriSpeech for ASR and ASV.

Metric: We employ untargeted attacks to assess the transferability of the proposed method via three tasks: KWS, ASR, and ASV. For the KWS task, we use the Transfer Attack Success Rate (TASR), defined as the proportion where $F(x') \neq y$; for the ASR task, the Word Error Rate (WER) serves as the evaluation metric; and for the ASV task, the cosine similarity (CS) is adopted, which measures the cosine similarity between the adversarial embedding $F(x^{adv})$ and the source embedding $F(x)$.

Comparison method: We compare the proposed method with I-FGSM (Carlini and Wagner 2018), NIA (Kim, Park, and Lee 2023), TI-FGSM (Dong et al. 2019), DI-FGSM (Xie et al. 2019), RMA (Bui et al. 2024), ACG (Yamamura et al. 2022), and STA-MDCT(Yao et al. 2024). For all FGSM-series attacks, the maximum perturbation is set to $\epsilon = 0.03$, the number of iterations is fixed at 30, and the step

Attack Method	WER ↑ of ASR			CS ↓ of ASV			TASR ↑ of KWS		
	wav2vec*	conformer	crdnn	hubert*	ecapa-tdnn	resnet	bc_resnet*	wav_kws	kwt
Clean	0.020	0.020	0.029	1.000	1.000	1.000	0.045	0.137	0.003
I-FGSM	1.438	0.124	0.158	0.052	0.647	0.671	**1.000**	0.430	0.503
TI-FGSM	1.367	0.115	0.146	0.050	0.648	0.672	**1.000**	0.490	0.500
DI-FGSM	0.552	0.219	0.210	0.585	0.736	0.755	**1.000**	0.450	0.460
NIA	0.357	0.169	0.193	0.447	0.587	0.634	0.951	0.454	0.411
RMA	0.441	0.108	0.132	0.356	0.729	0.760	0.976	0.463	0.496
ACG	0.927	0.167	**0.302**	0.170	0.668	0.703	0.940	0.496	0.523
STA-MDCT	0.842	0.224	0.214	0.324	0.604	0.637	0.977	0.537	0.485
TS-p3	**1.697**	**0.245**	0.263	**0.020**	**0.586**	**0.613**	**1.000**	**0.561**	**0.586**
TS-3	–	–	–	0.067	0.631	0.656	**1.000**	0.520	0.566

Table 2: Comparison of transferability of different adversarial attack methods. (* is a surrogate model.)

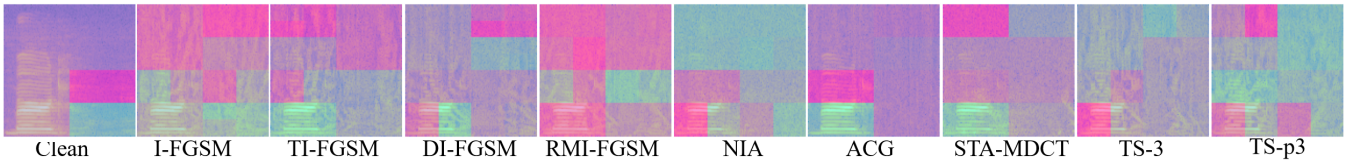


Figure 2: Comparison of SHAP heatmaps for different adversarial attacks.

size is 0.001. The noise introduced in NIA and STA-MDCT is ϵ times the Gaussian disturbance. The hyperparameters of TS are as follows: $n = 3$, $m = 3$, and $w = 1/n$.

Comparative Analysis Against Existing Attacks

This section systematically compares the proposed Time-Shuffle (TS) attack with seven audio adversarial methods across three tasks (ASR, ASV, KWS) and nine models. The experimental results are shown in Table 2. For the ASR task, TS-p3 achieved the best performance (1.697) on the white-box model wav2vec, exceeding the second-best method I-FGSM by 0.259. On the black-box models conformer and crdnn, TS-p3 achieved the best and second-best results (with the second-best result lower than ACG), respectively, and attained the highest average black-box transferability of 0.254. For the ASV task, TS-p achieved the lowest cosine similarity on the white-box model hubert and also achieved the best solution on the black-box models ecapa-tdnn and resnet. Specifically, the average cosine similarity of TS-p is 0.599, which is lower than those of I-FGSM (0.659), TI-FGSM (0.66), DI-FGSM (0.745), NIA (0.610), RMA (0.744), ACG (0.685), STA-MDCT (0.620), and TS (0.643). TS-p3 ranks fourth in transferability on the ASV task, lower than TS-p, NIA, and STA-MDCT. This means that the phoneme-based TS algorithm has better transferability than the segment-based TS. We also observed that DI-FGSM performs well on ASR but poorly on ASV, which may be due to gradient deviation caused by scaled audio significantly altering the frequency band characteristics. Similarly, for the KWS task, TS-p also achieved the SOTA attack success rate on both white-box and black-box models, while TS ranked second. The reason TS performs better than ASV in KWS is that the audio length of ASV is much greater than that of KWS, limiting the effectiveness of shuffling with the same number of segments.

Figure 2 presents the SHAP heatmaps of different adversarial samples. A successful adversarial attack can divert the model's attention. Visually, in this example, both TS-3 and TS-p3 exhibit distinct SHAP value distributions that significantly deviate from the original sample's feature importance pattern.

In general, the TS-p method achieves SOTA results in adversarial attacks against various acoustic models (except crdnn). The performance of TS is lower than that of TS-p, but TS performs better on the KWS task than on the ASV task, further indicating that the acoustic model's decision-making relies more on local temporal features.

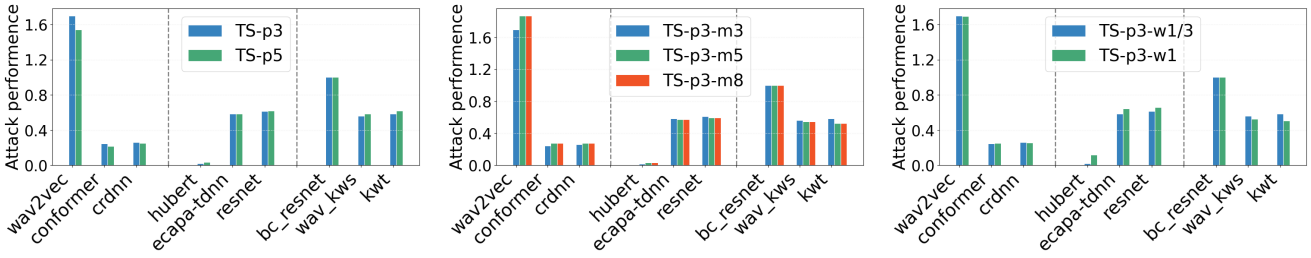
Considering that TS-p has better performance, we mainly perform ablation experiments on TS-p in the main paper. Please refer to the appendix for detailed experiments on TS.

The Boost Effect of TS-p

As an effective input transformation attack, TS should not only exhibit stronger transferability but also be compatible with other input transformation methods to generate more transferable adversarial examples. We combine TS-p with other attack methods and denote the combined approach as TSp+*(where * represents the integrated attack method). In this experiment, we fix the parameters n=3, m=3, and $w = 1/3$. As shown in Table 3, most attack methods (except for STA-MDCT) demonstrated improved performance after integrating TS-p. Specifically, the average white box/black box transferability of the TS-p series in ASR, ASV, and KWS tasks reached 1.051/0.213, 0.195/0.61, and 0.93/0.50, respectively. Compared with the original results of 0.831/0.173, 0.295/0.660, and 0.981/0.47,

Attack Method	WER ↑ of ASR			CS ↓ of ASV			TASR ↑ of KWS		
	wav2vec*	conformer	crdnn	hubert*	ecapa-tdnn	resnet	bcresnet*	wavkws	kwt
TSp+I-FGSM	$1.69_{\uparrow.25}$	$0.24_{\uparrow.12}$	$0.26_{\uparrow.10}$	$0.02_{\downarrow.03}$	$0.58_{\downarrow.06}$	$0.61_{\downarrow.05}$	1.00	$0.56_{\uparrow.13}$	$0.58_{\uparrow.08}$
TSp+TI-FGSM	$1.71_{\uparrow.34}$	$0.25_{\uparrow.13}$	$0.26_{\uparrow.12}$	$0.03_{\downarrow.01}$	$0.58_{\downarrow.06}$	$0.61_{\downarrow.05}$	1.00	$0.50_{\uparrow.00}$	$0.48_{\downarrow.02}$
TSp+DI-FGSM	$1.36_{\uparrow.81}$	$0.25_{\uparrow.04}$	$0.27_{\uparrow.06}$	$0.07_{\downarrow.50}$	$0.63_{\downarrow.10}$	$0.66_{\downarrow.08}$	1.00	$0.54_{\uparrow.09}$	$0.47_{\uparrow.01}$
TSp+NIA	$0.37_{\uparrow.01}$	$0.18_{\uparrow.01}$	$0.23_{\uparrow.04}$	$0.38_{\downarrow.06}$	$0.51_{\downarrow.07}$	$0.57_{\downarrow.06}$	$0.98_{\uparrow.03}$	$0.53_{\uparrow.07}$	$0.48_{\uparrow.07}$
TSp+RMA	$1.15_{\uparrow.71}$	$0.28_{\uparrow.12}$	$0.26_{\uparrow.10}$	$0.14_{\downarrow.21}$	$0.56_{\downarrow.16}$	$0.60_{\downarrow.15}$	$1.00_{\uparrow.02}$	$0.54_{\uparrow.08}$	$0.52_{\uparrow.02}$
TSp+STAMDCT	$0.03_{\downarrow.80}$	$0.03_{\downarrow.21}$	$0.05_{\downarrow.15}$	$0.53_{\uparrow.21}$	$0.69_{\uparrow.08}$	$0.72_{\uparrow.08}$	$0.64_{\downarrow.32}$	$0.44_{\downarrow.09}$	$0.43_{\downarrow.05}$

Table 3: The transferability of different adversarial attack algorithms after introducing TS-p.



(a) Comparing different n in TS-p algorithm (b) Comparing different M in TS-p algorithm (c) Comparing different w in TS-p algorithm

Figure 3: The impact of different parameters n, M and w on the TS-p algorithm.

the transferability increased by 26.4%/23%, 14.1%/14.7%, and 5.4%/6.3%, respectively. This indicates that TS-p has strong integrative capability, further validating its high effectiveness in mainstream acoustic tasks and its excellent compatibility with other input transformation-based attacks.

Ablation Study

In order to gain a deeper understanding of the performance improvement of TS-p, we further analyze the impact of hyperparameters n, M, and w on transferability in this section.

The effectiveness on the number of segments n. The parameter n represents the number of segments involved in the shuffling, where a higher n indicates a finer-grained shuffling process.

As shown in Figure 3.(a), as n increases from 3 to 5, TS-p exhibits a slight decrease in WER for the ASR task and a slight increase in CS for the ASV task, while TASR shows a slight increase for KWS. Based on the results in Table 1, this discrepancy may arise from KWS tasks prioritizing finer-grained processing compared to ASR and ASV. However, regardless of the scenario, TS-p still reaches SOTA performance, which further supports the finding that existing acoustic models emphasize a fine-grained local temporal feature.

The effectiveness on the times M **of the shuffle operation.** The parameter M denotes the number of shuffles. In this experiment, we followed the settings of Algorithm 1, setting $w = 1/M$ and $n = 3$.

As shown in Figure 3.(b), with increasing M, the WER of TS-p on the ASR task shows a significant improvement, and the CS on the ASV task also shows a slight improvement (a decrease indicates improvement). On the KWS task, TASR exhibited a slight decrease (especially in the Black-box model). Based on the results from the n-variation experiment, KWS focuses more on fine-grained local features (at the TS-p5 level), whereas ASV and ASR focus more on the TS-p3 level. Therefore, sampling more gradients at the TS-p3 level enhances adversarial attack performance on ASV and ASR, while slightly reducing KWS performance.

The effectiveness on weight w**.** The parameter w denotes the weight of gradient fusion, where an increase in w corresponds to the incorporation of more gradients from shuffled samples . In this experiment, we fixed $n = 3$ and $M = 3$. As shown in Figure 3.(c), when $w = 1$, there is a significant decrease in transferability across ASR, ASV, and KWS tasks. Specifically, average black-box transferability decreases by 0.007, 0.004, and 0.115 for ASR, ASV, and KWS tasks, respectively—still exceeding the performance of most competitive methods. This reduction may stem from excessive gradient incorporation from the shuffled sample, which induces model decision-making confusion and reduces enhancement efficacy.

Conclusion

This paper experimentally demonstrates the dependence of the acoustic models' decision-making on local temporal features. Through SHAP visualization analysis, we observe that time shuffle shifts the model's attention toward redundant features. Building on these findings, we propose a time shuffle adversarial attack, encompassing phoneme-based TS-p and segment-based TS. Extensive experiments validate that the TS series demonstrates competitive performance across diverse speech adversarial tasks. Notably, TS-p enhances transferability more effectively than TS, which further corroborates the critical role of local temporal features in the decision of acoustic models.

Acknowledgments

This research was supported by National Natural Science Foundation of ChinaNSFC (No.62472325)

References

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460.

Berg, A.; O'Connor, M.; and Cruz, M. T. 2021. Keyword transformer: A self-attention model for keyword spotting.

Bui, M.; Doan, T.-P.; Hong, K.; and Jung, S. 2024. Boosting Black-Box Transferability of Weak Audio Adversarial Attacks with Random Masking. In *International Conference on Information Security Applications*, 96–108.

Carlini, N.; and Wagner, D. 2018. Audio adversarial examples: Targeted attacks on speech-to-text. In *IEEE security and privacy workshops (SPW)*, 1–7.

Chen, G.; Zhao, Z.; Song, F.; Chen, S.; Fan, L.; and Liu, Y. 2022. AS2T: Arbitrary source-to-target adversarial attack on speaker recognition systems. *IEEE Transactions on Dependable and Secure Computing*.

Chen, J.; Feng, Z.; Zeng, R.; Pu, Y.; Zhou, C.; Jiang, Y.; Gan, Y.; Li, J.; and Ji, S. 2025. Enhancing Adversarial Transferability with Adversarial Weight Tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2061–2069.

Chen, Z.; Li, J.; and Chen, C. 2024. Ensemble Adversarial Defenses and Attacks in Speaker Verification Systems. *IEEE Internet of Things Journal*.

Desplanques, B.; Thienpondt, J.; and Demuynck, K. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification.

Dong, Y.; Pang, T.; Su, H.; and Zhu, J. 2019. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4312–4321.

Fang, Z.; Wang, T.; Zhao, L.; Zhang, S.; Li, B.; Ge, Y.; Li, Q.; Shen, C.; and Wang, Q. 2024. Zero-query adversarial attack on black-box automatic speech recognition systems. In *ACM SIGSAC Conference on Computer and Communications Security*, 630–644.

Gao, X.; Li, Z.; Chen, Y.; Liu, C.; and Li, H. 2024. Transferable adversarial attacks against asr. *IEEE Signal Processing Letters*.

Ge, Y.; Zhao, L.; Wang, Q.; Duan, Y.; and Du, M. 2023. Advddos: Zero-query adversarial attacks against commercial speech recognition systems. *IEEE Transactions on Information Forensics and Security*, 18: 3647–3661.

Gulati, A.; Qin, J.; Chiu, C.-C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. 2020. Conformer: Convolution-augmented transformer for speech recognition.

Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29: 3451–3460.

Jia, S.; Yin, B.; Yao, T.; Ding, S.; Shen, C.; Yang, X.; and Ma, C. 2022. Adv-attribute: Inconspicuous and transferable adversarial attack on face recognition. *Advances in Neural Information Processing Systems*, 35: 34136–34147.

Kim, B.; Chang, S.; Lee, J.; and Sung, D. 2021. Broadcasted Residual Learning for Efficient Keyword Spotting. In *Proc. Interspeech*, 4538–4542.

Kim, H.; Park, J.; and Lee, J. 2023. Generating transferable adversarial examples for speech classification. *Pattern Recognition*, 137: 109286.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Nagrani, A.; Chung, J. S.; Xie, W.; and Zisserman, A. 2020. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60: 101027.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: an asr corpus based on public domain audio books. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5206–5210.

Ravanelli, M.; Parcollet, T.; Plantinga, P.; Rouhe, A.; Cornell, S.; Lugosch, L.; Subakan, C.; Dawalatabad, N.; Heba, A.; Zhong, J.; et al. 2021. SpeechBrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

Seo, D.; Oh, H.-S.; and Jung, Y. 2021. Wav2kws: Transfer learning from speech representations for keyword spotting. *IEEE Access*, 9: 80682–80691.

Tong, C.; Zheng, X.; Li, J.; Ma, X.; Gao, L.; and Xiang, Y. 2023. Query-efficient black-box adversarial attacks on automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 3981–3992.

Villalba, J.; Chen, N.; Snyder, D.; Garcia-Romero, D.; McCree, A.; Sell, G.; Borgstrom, J.; García-Perera, L. P.; Richardson, F.; Dehak, R.; Torres-Carrasquillo, P. A.; and Dehak, N. 2020. State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations. *Computer Speech & Language*, 60: 101026.

Wang, Z.; Wang, W.; Chen, Q.; Wang, Q.; and Nguyen, A. 2024. Generating valid and natural adversarial examples with large language models. In *International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 1716–1721.

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

Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; and Yuille, A. L. 2019. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2730–2739.

Yamamura, K.; Sato, H.; Tateiwa, N.; Hata, N.; Mitsutake, T.; Oe, I.; Ishikura, H.; and Fujisawa, K. 2022. Diversified adversarial attacks based on conjugate gradient method. In *International Conference on Machine Learning*, 24872–24894. PMLR.

Yao, J.; Luo, H.; Qi, J.; and Zhang, X.-L. 2024. Interpretable spectrum transformation attacks to speaker recognition systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 1531–1545.

Zeinali, H.; Wang, S.; Silnova, A.; Matějka, P.; and Plchot, O. 2019. But system description to voxceleb speaker recognition challenge 2019. *arXiv preprint arXiv:1910.12592*.