Frontiers in Computer Science

TYPE Original Research
PUBLISHED 15 February 2024
DOI 10.3389/fcomp.2024.1355975

OPEN ACCESS

EDITED BY
Zhao Ren,
University of Bremen, Germany

REVIEWED BY
Yi Chang,
Imperial College London, United Kingdom
Xin Jing,
University of Augsburg, Germany

*CORRESPONDENCE
Sheng Li
✉ sheng.li@nict.go.jp

RECEIVED 14 December 2023
ACCEPTED 29 January 2024
PUBLISHED 15 February 2024

CITATION
Li S, Li J and Cao Y (2024) Phantom in the opera: adversarial music attack for robot dialogue system.
Front. Comput. Sci. 6:1355975.
doi: 10.3389/fcomp.2024.1355975

Phantom in the opera: adversarial music attack for robot dialogue system

Sheng Li[1]*, Jiyi Li[2] and Yang Cao[3]

[1]National Institute of Information and Communications Technology, Kyoto, Japan, [2]University of Yamanashi, Kofu, Japan, [3]Hokkaido University, Sapporo, Japan

This study explores the vulnerability of robot dialogue systems' automatic speech recognition (ASR) module to adversarial music attacks. Specifically, we explore music as a natural camouflage for such attacks. We propose a novel method to hide ghost speech commands in a music clip by slightly perturbing its raw waveform. We apply our attack on an industry-popular ASR model, namely the time-delay neural network (TDNN), widely used for speech and speaker recognition. Our experiment demonstrates that adversarial music crafted by our attack can easily mislead industry-level TDNN models into picking up ghost commands with high success rates. However, it sounds no different from the original music to the human ear. This reveals a serious threat by adversarial music to robot dialogue systems, calling for effective defenses against such stealthy attacks.

KEYWORDS

an industry-popular ASR model, namely the time-delay neural network (TDNN), widely used for speech and speaker adversarial music attack, deep neural network, robot dialogue system

1 Introduction

Recently, Android robots have attracted considerable attention from researchers and the public. Their distinguishing feature is their realistic human-like appearance that also requires their behaviors to be human-like (Glas et al., 2016). Their use as an interface for natural conversation makes them an attractive option for use in daily life scenarios (Inoue et al., 2019).

The automatic speech recognition (ASR) module is crucial in these recent robot dialogue systems as the most natural human-machine interface. Owing to their superior representation learning capabilities, deep neural networks (DNNs) have been widely used to achieve state-of-the-art performance in several applications, including ASR. For example, the recently proposed time-delay neural network (TDNN; Waibel et al., 1989; Peddinti et al., 2015; Sun et al., 2017; Myer and Tomar, 2018) has demonstrated much better performance on ASR and speaker recognition tasks (Chen et al., 2023; Wang et al., 2023) than the traditional models. However, DNNs are known to be vulnerable to adversarial attacks. This vulnerability of DNNs has been extensively studied in the image domain for tasks, such as image classification and object detection, but has been rarely explored in the audio domain except for a few seminal works (Carlini and Wagner, 2018; Qin et al., 2019; Yakura and Jun, 2019).

In this study, we test the robustness of the ASR modules of a robot dialogue system to adversarial music attacks. Music can be exploited as a natural camouflage to hide ghost voice commands and mislead the ASR systems into picking up these commands without leaving any obvious traces. Therefore, we propose a novel adversarial music attack to slightly perturb normal music so that it can carry ghost commands into the perturbation. We demonstrate the effectiveness of our proposed attack on an industrial-level ASR system trained on a benchmark speech dataset. It can hide diverse commands in different types of music while achieving high attack success rates. This indicates a serious threat of ghost voice commands to modern artificial intelligence systems and devices such as smart home devices and smart cars.

The rest of this paper is organized as follows. Section 2 briefly reviews the related studies. Section 3 describes our proposed method. Section 4 presents the details of the implementation and experimental evaluations. The conclusions and future studies are described in Section 5.

2 Related studies

In this section, we briefly review the ASR models of a robot dialogue system and study the existing research on audio adversarial attacks.

2.1 Robot dialogue system

As described in the previous section, the state-of-the-art robot dialogue system is highly realistic and displays human-like conversational behaviors. In the typical architecture of the spoken dialogue system (Glas et al., 2016), the robot captures the user's speech through a microphone array so that the user can speak naturally without needing to hold a microphone. Speaker identification and human behavioral sensing are implemented through Kinect and a small camera. A real voice artist trains her text-to-speech synthesized voice, allowing her to generate realistic-sounding backchannels, laughs, and fillers.

The most crucial module of the robot dialogue system is the ASR system because all behaviors are conditioned on the ASR results. This study uses a setting similar to the ASR system to simulate the adversarial music attack on the robot dialogue system.

2.2 Automatic speech recognition (ASR)

ASR is a technique to transcribe voice to text and is one of the core techniques for man-to-machine and machine-to-machine communications. In recent years, ASR techniques have been extensively used in information retrieval and speech-to-text services, such as the speech assistant of Apple Siri, Amazon Alexa/Echo home management service, Google Homesmart search and service assistant, and the Microsoft Cortana personal assistant. In these applications, ASR serves as an efficient and smart interface, and its performance is essential for the functioning of these services.

In a nutshell, ASR maps a spoken audio sequence to a word sequence. Under the statistical framework, the problem is formulated as maximizing the posterior probability of a word sequence when observing an audio sequence. The traditional models are hybrid models, such as the Gaussian-mixture model in combination with the hidden Markov model (GMM-HMM; Rabiner, 1988) or a deep neural network with the hidden Markov model (DNN-HMM; Dahl et al., 2012). These hybrid models consist of two independently optimized components: the acoustic and language models. Modern ASR models follow an end-to-end framework that integrates the two components (e.g., acoustic model and language model) into a single trainable network (Graves et al., 2006; Graves and Jaitly, 2014; Chan et al., 2016; Vaswani et al., 2017; Watanabe et al., 2018). The output words or characters can be treated as labels in these models. We note that deploying ASR systems in real-world applications is still challenging due to the complex and noisy physical world conditions.

2.3 Audio adversarial attacks

Adversarial examples, which (or attacks) can be crafted by adding small, carefully engineered perturbations into clean examples, have attracted enormous interest in the field of computer vision. In the context of images, adversarial examples appear the "same" as their original versions and yet can mislead DNNs with high success rates. In a white-box setting where the attacker knows the model parameters, adversarial examples can be easily generated using gradient-based methods, such as the fast gradient sign method (FGSM; Goodfellow et al., 2014) and projected gradient descent (Madry et al., 2017). These attack methods are mostly developed for images and often require certain modifications to other media, such as texts and audio.

Several studies have crafted audio adversarial examples with the intent to mislead ASR systems. These include the genetic algorithm (Alzantot et al., 2017) and optimization-based (Cisse et al., 2017) audio attacks. These early attempts are all *untargeted* attacks that mislead an ASR model to translate adversarial audio into incorrect transcripts. The DolphinAttack (Zhang et al., 2017) is a targeted attack that can mislead the ASR models into recognizing and converting an inaudible adversarial ultrasound signal to a specific transcript that is of interest to the attacker. However, DolphinAttack requires special ultrasound hardware to generate the ultrasound. Voice commands can also be disguised as noise that sounds meaningless to humans (Carlini et al., 2016; Abdullah et al., 2019). One common weakness of early audio attacks is that the generated adversarial noise is suspicious, potentially exposing the attacker, and can be easily detected by the liveness detection methods (Abdullah et al., 2019).

Consequently, Carlini and Wagner (2018) proposed targeted attacks[1] against an end-to-end DeepSpeech model (Graves and Jaitly, 2014) by directly perturbing the audio waveform in a white-box setting. Alternatively, the CommanderSong (Yuan et al., 2018) attack injects the voice command into a song to mount a targeted attack. However, the adversarial speeches or songs crafted by these two methods often contain obvious distortions and thus sound obviously different from the normal audio to human ears.

To solve these problems, Schönherr et al. (2019) proposed the *psychoacoustic hiding* technique to hide voice signals that are below a certain threshold of human perception. However, this attack is

[1] Here, the target means with a given command in the white-box settings. In the latter sections, the target means the system and models being attacked.

ineffective when played over the air (Qin et al., 2019), and the tested ASR system is a traditional DNN-HMM model. Similarly, Qin et al. (2019) introduced *frequency masking* to hide adversarial commands into regions of the audio that are not perceptible by human hearing. However, this method was only tested on laboratory-level ASR models. Moreover, the frequency masking technique can only be applied to audio regions that have sufficient energy. The attacker must wait for strong signals to insert the ghost command. This significantly limits its flexibility in noisy real-world scenarios. Except for the CommanderSong (Yuan et al., 2018), most existing audio attacks are based on speeches. In this study, we propose a stealthier music attack to trick the ASR system into picking up hidden ghost commands.

3 Proposed adversarial music attack

Unlike existing audio attacks, we focus on music's unique characteristics to ensure that the inserted perturbations are much more flexible over time and are not limited to high-energy regions. Music often has a more complex structure than speech and thus provides a more natural camouflage for adversarial voice commands.

Considering one pop ballad shown in Figure 1 as an example, it is observed that it contains more resonances and a clear overtone structure (as marked out in the figure), as the singer follows traditional vocal singing skills. However, for other music types, such as J-POP, the singers of rock, pop, and heavy-metal songs tend to pay more attention to emotional expressions, and therefore their singing produces fewer overtones. Hence, we need to study different types of music to determine if there is a universal standard for hiding ghost voice commands.

We denote the ASR model (which is also the attacker's target model) as $f(\cdot)$, which maps the audio input to transcription. Given a normal audio input x, a target model $f(\cdot)$, an adversarial attack's problem is finding an adversarial example x' that can mislead the model into making incorrect predictions. In this study, we consider TDNN as our target model. However, our attack is not restricted to a particular DNN architecture.

3.1 Adversarial music example generation stage

The ASR model $f(.)$ is trained to map the audio input to the target transcription. The proposed adversarial music attack is illustrated in Figure 2 and consists of the following steps.

1. We prepare the music audio, attack command texts, and train a TDNN-HMM model with human speech data. The details of model training are described in Section 4.1.
2. We apply forced-alignment (Young et al., 2009) on the music to derive the initial target label to force the attack to be inserted into the most likely mask regions based on its energy and the contextual continuities determined by the silence and non-silence states in the acoustic model. This step corresponds to Figure 2 (a). The purpose is to find the time boundaries for each command word.
3. We iteratively perturb the input examples toward a targeted adversarial label for every frame. This corresponds to step (b) in Figure 2. In particular, the frame-level (tied-triphone-) state sequence is used as the adversarial target label y_{adv}=[s_1, s_2, s_3, ...s_m], where s_i ($1 \leq$ i $\leq m$) is the (tied-triphone-) state id, and m is the frame number of the sentence. The duration of each frame is 10 ms. For the adversarial perturbation, we first compute the actual network output $\hat{y}$:

$$\hat{y} = f(x_i). \tag{1}$$

We then define the difference between the network output $\hat{y}$ and the target label y_{adv} by a cross-entropy loss function $Ł(\hat{y}, y_{\text{adv}})$:

$$Ł(\hat{y}, y_{\text{adv}}) = -\sum y_{\text{adv}} \log \hat{y}, \tag{2}$$

The gradient of the above adversarial loss is computed and back-propagated to the input music x:

$$\nabla x = \frac{\partial Ł(\hat{y}, y_{\text{adv}})}{\partial x}. \tag{3}$$

The input music x is then iteratively perturbed according to the gradient and step size α:

$$x_{t+1} = x_t - \alpha \cdot \nabla x_t, \tag{4}$$

where t is the current perturbation step, we note that, unlike conventional model training, here the model parameters remain unchanged during the attack, and only the input music waveform x is iteratively perturbed for a total number of T (e.g., 1,000) iterations.

In previous studies (Carlini and Wagner, 2018; Yakura and Jun, 2019), an L_p norm-based clipping was also applied to constrain the perturbation within a small L_p norm sphere around the original example x. However, the human perception in the audio space is too complex to be modeled by standard L_p norms. Therefore, we do not use such a constraint in our attack. Instead, we ensure the imperceptibility of the adversarial noise by perturbing only the masking regions. This is guaranteed by the masking operation described in step 1.

To constrain the perturbation, we intend to locate the position on the timeline at the frame level. Music structures like overtones are relatively simple, and it's easy to hide attacks in the low-frequency domain (<8,000 Hz) using the frequency masking method (as shown in Figure 2).

3.2 Attack stage

As introduced in Section 2.1, the most crucial module of the whole robot dialogue system is the ASR system because all the behaviors are conditioned by the keywords extracted from the ASR result. In a typical case, the dialogue manager $g(\cdot)$ is a system that maps the output of the ASR model $\hat{y}$ directly to the robot's response to human speech. Proposed adversarial perturbation steps

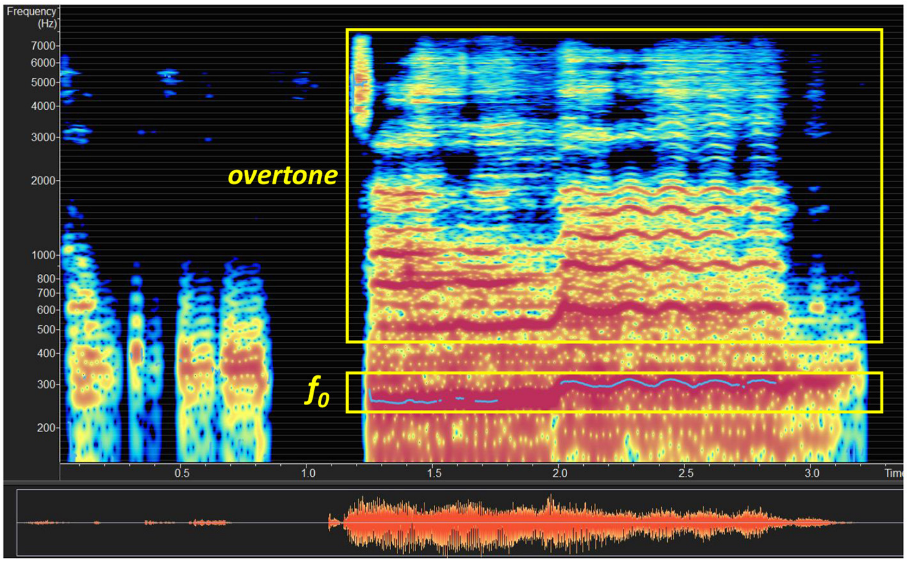

FIGURE 1
Musical structure of a pop ballad song.

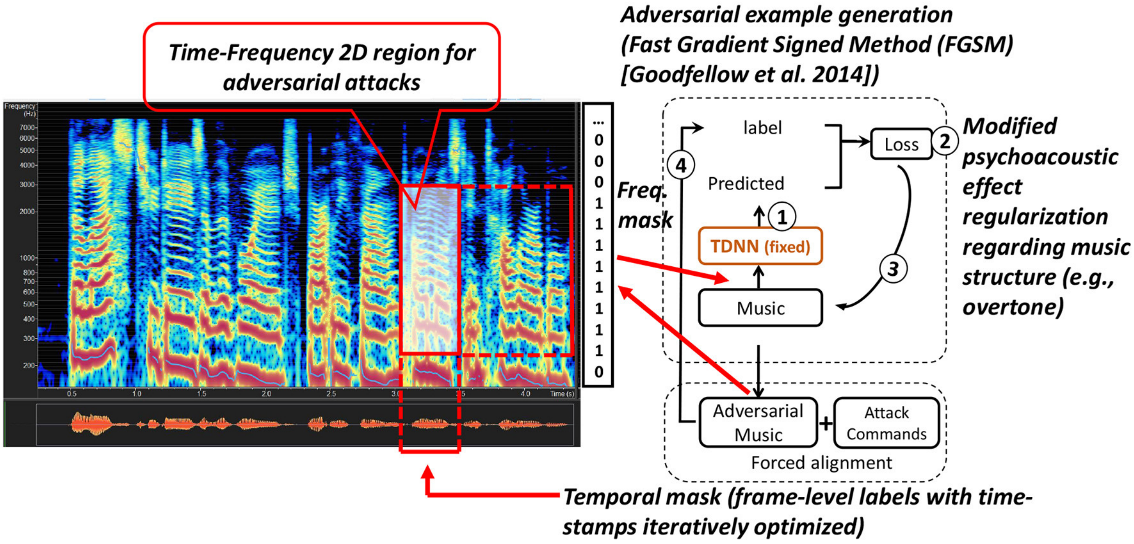
Time-Frequency 2D region for adversarial attacks
Adversarial example generation (Fast Gradient Signed Method (FGSM) [Goodfellow et al. 2014])
...
0
0
0
1
1
1
1
1
1
1
1
0
Freq. mask
label
Predicted
Loss
TDNN (fixed)
Music
1
2
3
4
Modified psychoacoustic effect regularization regarding music structure (e.g., overtone)
Adversarial Music
+
Attack Commands
Forced alignment
Temporal mask (frame-level labels with time-stamps iteratively optimized)

FIGURE 2
An overview of the generation of adversarial music with details.

will mislead the model to output the adversarial target label y_{adv}, and subsequently also mislead the dialogue manager system to produce an incorrect response $g(y_{\text{adv}})$, meeting the goal of this study.

4 Experiments

4.1 Acoustic model training

We train an acoustic model using 460 h of Librispeech data (train-clean-100 and train-clean-360; Panayotov et al., 2015). We first train a GMM-HMM model using the MFCC feature with linear discriminant analysis (LDA), a maximum likelihood linear transform, and fMLLR-based speaker adaptive training. Then, we train a TDNN model with four hidden layers, each layer comprising 2,048 hidden neurons. The output layer has 3,456 neurons corresponding to the tied-triphone-states of the GMM-HMM model. We use the GMM-HMM model to derive state alignment as the training label.

Instead of using the MFCC/Fbank feature to train the TDNN model, we use the 256-dimension raw waveform feature (16,000 kHz, 16 bits, mono-channel). All of these features are mean normalized (CMN) per speaker. It is difficult to obtain a good reconstruction from MFCCs for two reasons: (1) MFCCs discard a large amount of information by a low-rank linear projection of the mel spectrum. (2) The phase information is lost, even though there are several methods for estimating it (e.g., Griffin and Lim, 1984).

At TDNN hidden layers, we splice the frames with offsets $\{-2, -1, 0, 1, 2\}$, $\{-1, 2\}$, $\{-3, 3\}$, and $\{-7, 2\}$. The TDNN model is parallel trained using natural stochastic gradient descent based on the cross-entropy loss criterion (Povey et al., 2015). The HMM model is derived from the GMM-HMM model. All of these models were implemented using the Kaldi toolkit (Povey and et al., 2011).

We use two evaluation sets from Librispeech (dev-clean and test-clean) for testing. The word error rates (WER%) of our TDNN-HMM model are between ∼6 and 7%, according to a 4 g word language model trained from the transcription of the entire Librispeech training data set.

Because our attack target is an ASR system in the communication robot, we must consider the impact of the far field on the ASR system. We generate simulated noisy and reverberant data according to Ko et al. (2015). The recipes[2] and noisy data set[3] are publicly available. The Librispeech clean data are added with the additive noise and convolved with the room impulse responses (RIRs). There are 325 real condition RIRs, and their reverberation times range roughly from 0.2 to 1.5 s. We also add 60,000 simulated RIRs generated from the three large rooms according to different speaker positions. Their reverberation times range roughly from 0.2 to 1.8 s. We also add additive noise with the signal-to-noise ratio (SNR) ranging from −5 to 20 dB. Then, we retrain our acoustic model based on the simulated noisy data. The music clips played on the air require weighted prediction error (WPE) dereverberation (Delcroix et al., 2014). We use the online WPE recipe[4] provided in the Reverb2014 website.[5]

4.2 Command and music

We select 22 commands that may cause some danger from the Google-home command sets.[6] The selected 22 commands cover a wide range of real-life scenarios such as time scheduling, phone calls, media, and activation of other devices. We chose these commands for testing because they are the most frequently used commands in the current spoken dialogue systems. We further divided the 22 commands into five groups according to command length (e.g., the number of words in the command), as summarized in Table 1.

The number of commands for demonstrating this kind of attack is sufficient enough. Since the ASR model is tried at the phone level, the dictionary can be extended to as large as librispeech's dictionary size. In the experiment, we mainly test the influence of the command lengths. For the same reason, The music clip covers the general category. In the experiment, we mainly verify the influences of music types.

TABLE 1 Command text example groups ("OK" serves as a wake-up word).

Length	Command example
4	OK, ring my phone.
5	OK, open the front door.
6	OK, delete all of my reminders.
7	OK, book a room at San Francisco.
9	OK, set an alarm for every morning at three.

TABLE 2 Seven types of music.

Types	No. of clips	Average duration (s)
Anime	5	7
Ballade	1	4
Bigband	2	9
Heavy-metal	4	7
House-music	1	5
Pops	1	9
Rock	2	7

In our experiment, we introduce the commands into seven types of music, including anime, ballade, big band, heavy-metal, house music, pops, and rock, as listed in Table 2. These types of music are selected from NHK background music. These music clips show very diverse styles.

4.3 Attack error rate for ghost commands

We divided the commands into five groups according to the number of words in each command, as shown in Figure 3.

We first evaluate our adversarial music examples by WER. Then, we divided our commands into two parts: the wake-up part and the command part. The wake-up part aims to wake up our robot-like human dialogues where it is necessary to call a person's name at the beginning of the conversation. The command part is the meaning of our sentences that directly determines the robot's reaction. We tested the word error rate of these two parts to evaluate the attack success rate of our adversarial examples and show the results in Figure 3. A lower error rate means that the output of the ASR model is closer to the adversarial target model, i.e., our music attack is more effective. It is observed that the command length has a considerable influence on the attack's error rate. As the command length increases, the error rate of the attack decreases, implying that the proposed music attack becomes increasingly effective. The slight rebound in the last test group may be caused by insufficient test data.

Furthermore, Table 3 shows three types of attack error rates. They are the WER and the two different types of pre-set keywords, namely, the wake-up word error rate (WWER) and command word

[2] https://github.com/kaldi-asr/kaldi/blob/master/egs/reverb

[3] http://www.openslr.org/28/

[4] https://github.com/fgnt/nara_wpe

[5] https://reverb2014.dereverberation.com

[6] https://www.cnet.com/how-to/every-important-google-home-and-google-assistant-command-you-can-give

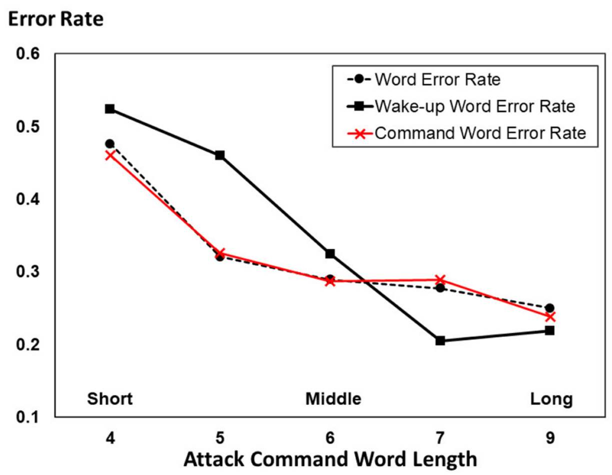

FIGURE 3
Attack error rate of commands with different lengths.

TABLE 3 Attack error rate of different types of adversarial music to the ASR system.

Type	WER (%)	WWER (%)	CWER (%)
Anime	23.15	24.77	24.77
Ballade	21.43	85.71	23.81
Big Band	17.08	17.95	5.13
Heavy Metal	51.15	34.09	62.50
House Music	20.76	7.69	17.95
Pops	18.90	9.52	14.29
Rock	41.35	82.93	63.41

The best and second-best results are marked individually in gray and light-gray cells.

error rate (CWER) results, and they are shown for different types of adversarial music. It is observed that the music type can also significantly impact the attack success rates. Intuitively, it may be expected that music with a high-energy spectrum will make it easier to hide ghost commands with a perturbation in the spectrum. However, the obtained result is opposite to our expectations. It is found that the rock and heavy-metal music adversarial examples have the least impact on the model, which is quite counter-intuitive because these types of music often contain a high-energy spectrum. This is because the forced alignment process tends to insert the commands into the musical regions closer to the human voice. In our experiment, pop, and big band music have more regions with human-like energy and continuities (regardless of duration), making them better for misleading the target model.

4.4 Audio quality for adversarial music examples

To better identify the type of music more suitable for adversarial music generation, we performed a few evaluations using subjective and objective parameters. We aimed to find a type of music with a low error rate in the keyword-spotting task and a generated adversarial music example that is almost the same as the original music. Such type of music was found to be most suitable as a target for adversarial music generation.

TABLE 4 MOS of different types of music.

Type	Control	Short	Middle	Long
Rock (5s)	3.67	2.00	2.33	2.50
Rock (9s)	3.41	2.16	2.42	2.08
Heavy-Metal	3.58	2.08	2.08	2.00
Big Band	3.33	1.67	1.92	1.58
Pops	3.17	2.42	2.08	2.25

1: clearly quite different, 2: slightly different, 3: maybe similar, 4: clearly the same. The best and second best results are marked in gray and light-gray, respectively.

For subjective evaluation, based on the results of machine evaluation, we chose three different lengths of commands (short is a four-word command, middle is a six-word command, and long is a seven-word command) combined with five different types of music to generate our human listening adversarial music test set. In our experiment, we invited 12 listeners. We used a four-level mean opinion score (MOS) to judge the difference between the attack music and original music, where a score of four meant that the attack music and original music were identical, and one meant that the attack music was significantly different from the original music pair. We also added two original music pairs as a control group to ensure the objectivity of the score.

An examination of the data presented in Table 4 shows that most of the attack music samples have scores that are approximately one lower than the control group, indicating that the music attack examples used in this study are the same as the original music. However, the best results were obtained for pop music and rock music, which generated the highest scores in this listening test. According to the results of different command lengths, we find that the attack music example for the middle-length command has the least difference from the original music.

Table 5 shows the objective evaluation of the signal distortion between the attack and original music. Four criteria from the speech enhancement field are used, namely, perceptual evaluation of sleep quality (PESQ) and segment-SNR (SSNR), a composite measure for signal distortion (CSIG), a composite measure for noise distortion (CBAK), and a composite measure for overall speech quality (COVRL). Almost identical to the subjective test results, 5-s rock music and pop music perform best in the objective evaluation.

We also directly compared the spectrum of the adversarial examples generated by different music types. These music attack examples are uploaded.[7] Figure 4 shows the original big band music spectrum (left part) and the adversarial example spectrum (right part). It is observed that the music has a considerable

[7] halspeech.github.io/demo_journal.html

TABLE 5 Enhancement scores of different types of examples (length type: S, short; M, middle; L, long; the best and second best results are marked in gray and light-gray cells, respectively).

Type	Len	CSIG	CBAK	COVL	SSNR	PESQ
Rock (5s)	S	5.00	4.93	4.48	24.33	3.71
	M	5.00	4.99	4.57	24.43	3.82
	L	5.00	5.00	4.77	24.45	4.07
Rock (9s)	S	3.56	4.41	3.01	26.58	2.34
	M	3.53	4.40	2.96	26.67	2.29
	L	3.45	4.33	2.88	26.21	2.21
Heavy-Metal	S	4.67	4.47	3.86	23.16	2.95
	M	4.65	4.45	3.83	23.13	2.91
	L	4.51	4.35	3.66	23.01	2.72
Big Band	S	4.71	4.72	4.10	23.59	3.36
	M	3.92	4.08	2.98	24.32	1.93
	L	4.58	4.62	3.95	23.14	3.22
Pops	S	5.00	4.58	4.57	18.47	3.75
	M	4.99	4.38	4.23	18.29	3.35
	L	5.00	4.52	4.46	18.32	3.64

Perceptual evaluation of sleep quality (PESQ) and segment-SNR (SSNR), a composite measure for signal distortion (CSIG), a composite measure for noise distortion (CBAK), and a composite measure for overall speech quality (COVRL).

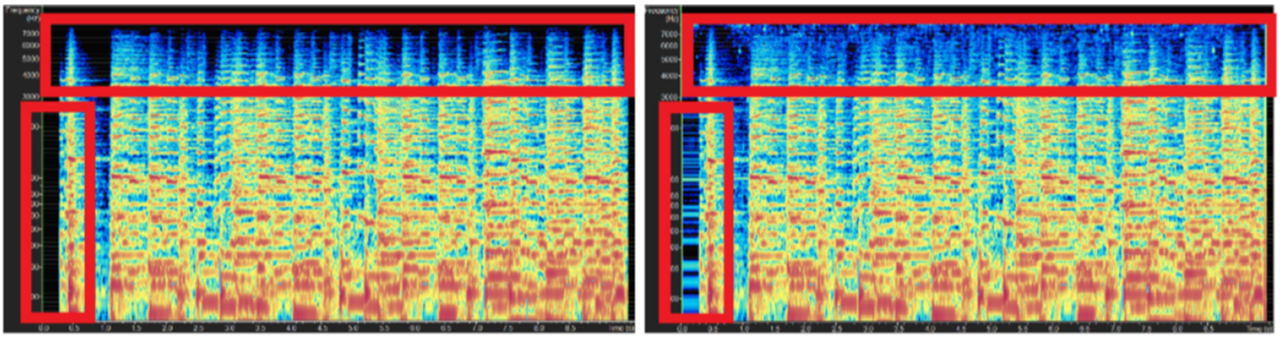

FIGURE 4
Spectrograms of the original big band music **(left)** and generated big band music with attacks in low-frequency <8 kHz **(right)**.

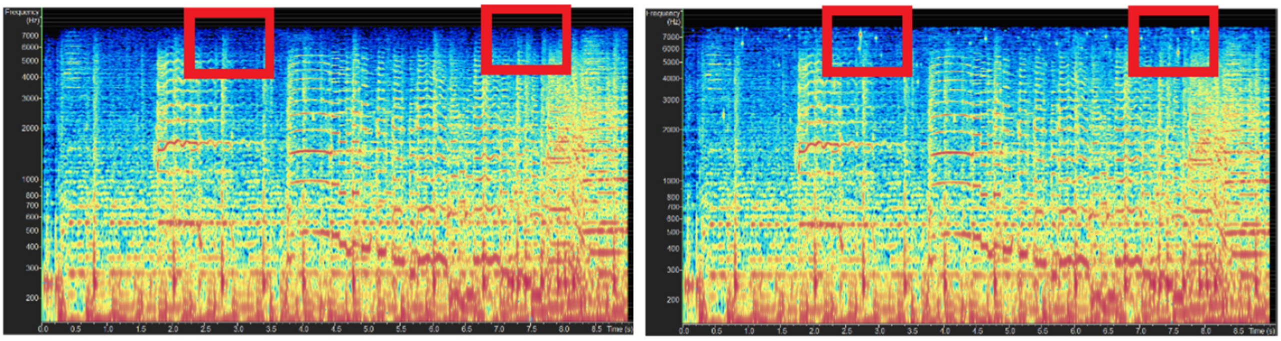

FIGURE 5
Spectrograms of the original pop music **(left)** and generate pop music with attacks in low-frequency <8 kHz **(right)**.

change in the energy-sparse part of the spectrum highlighted in red boxes. However, in the case of pop music shown in Figure 5, only a few sparse parts are observed in the spectrum, so the spectrum only changed slightly (also highlighted in red boxes). This might be the reason why big band music has the best ASR results but poor results in subjective and objective music quality experiments.

Based on the result of keyword-spotting error rates experiments, we noticed that pop and big band music demonstrate the lowest error rate because they are similar to the human voice regarding continuity and accentuation. In the subjective and objective music quality experiments, we discovered that pop and rock adversarial music cannot be easily distinguished from the original music because the spectrum of these types of music is very dense. Based on the above two points, pop-like music is identified as the best choice for making adversarial music examples to attack the TDNN-based ASR system.

The experiments show that the proposed adversarial music attack examples have a high attack success rate on the TDNN-based ASR model. Moreover, we verified the following features to be optimal for generating inaudible adversarial music. First, the use of a middle-length command increases the attack success rate. Second, the optimal regions for hiding our attacks should match the energy and contextual nature of the acoustic model. Third, the music spectrum range should be wide enough to hide the attack perturbation and to make it inaudible.

5 Conclusion

To the best of our knowledge, our study was the first to systematically investigate the design of adversarial examples for music to mislead the industry-level ASR model for robot dialogue systems. We first verified that the ghost words can be easily hidden in the music. Then, we proposed our FGSM-based method to generate the adversarial music attack examples. The experiments show that the proposed adversarial music attack examples have a high attack success rate on the TDNN-based ASR model. In the future, these discoveries will guide the use of adversarial music examples to develop more robust audio-based human-machine interfaces effectively. Also, the white-box method with knowledge of music proves our ideas in an explainable way, and we will extend this discovery to more black-box scenarios.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

SL: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing. JL: Formal analysis, Funding acquisition, Supervision, Validation, Writing - review & editing. YC: Formal analysis, Funding acquisition, Supervision, Validation, Writing - review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by JSPS KAKENHI Grant Numbers JP23K11227, JP22H03595, and JP23H03402.

Acknowledgments

Many thanks to Mr. Heran Zhang, Dr. Xingjun Ma, and Dr. Yi Zhao for their kind suggestions and help during this work.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abdullah, H., Garcia, W., Peeters, C., Traynor, P., Butler, K. R., and Wilson, J. (2019). Practical hidden voice attacks against speech and speaker recognition systems. *arXiv:1904.05734*. doi: 10.14722/ndss.2019.23362

Alzantot, M., Balaji, B., and Srivastava, M. B. (2017). "Did you hear that? Adversarial examples against automatic speech recognition," in *NIPS 2017 Machine Deception Workshop*.

Carlini, N., Mishra, P., Vaidya, T., Zhang, Y., Sherr, M., Shields, C., et al. (2016). "Hidden voice commands," in *25th $USENIX$ Security Symposium* (*$USENIX$ Security 16*), 513–530.

Carlini, N., and Wagner, D. A. (2018). Audio adversarial examples: targeted attacks on speech-to-text. *abs/1801.01944*. doi: 10.1109/SPW.2018.00009

Chan, W., Jaitly, N., Le, Q., and Vinyals, O. (2016). "Listen, attend and spell: a neural network for large vocabulary conversational speech recognition," in *Proceedings of IEEE-ICASSP*. doi: 10.1109/ICASSP.2016.7472621

Chen, Z., Han, B., Xiang, X., Huang, H., Liu, B., and Qian, Y. (2023). "Build a SRE challenge system: lessons from VoxSRC 2022 and CNSRC 2022," in *Proceedings of INTERSPEECH*, 3202–3206. doi: 10.21437/Interspeech.2023-1217

Cisse, M. M., Adi, Y., Neverova, N., and Keshet, J. (2017). "Houdini: fooling deep structured visual and speech recognition models with adversarial examples," in *Advances in Neural Information Processing Systems (NIPS)*, 6977–6987.

Dahl, G., Yu, D., Deng, L., and Acero, A. (2012). Context dependent pre-trained deep neural networks for large vocabulary speech recognition. *IEEE Trans. ASLP* 20, 30–42. doi: 10.1109/TASL.2011.2134090

Delcroix, M., Yoshioka, T., Ogawa, A., Kubo, Y., Fujimoto, M., Nobutaka, I., et al. (2014). "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the reverb challenge," in *Proceedings of REVERB Challenge Workshop*.

Glas, D. F., Minato, T., Ishi, C. T., Kawahara, T., and Ishiguro, H. (2016). "Erica: the Erato intelligent conversational android," in *International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 22–29. doi: 10.1109/ROMAN.2016.7745086

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Graves, A., Fernandez, S., Gomez, F., and Shmidhuber, J. (2006). "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of ICML*. doi: 10.1145/1143844.1143891

Graves, A., and Jaitly, N. (2014). "Towards end-to-end speech recognition with recurrent neural networks," in *Proceedings of ICML*.

Griffin, D., and Lim, J. (1984). Signal estimation from modified short-time Fourier transform. *IEEE Trans. ASSP* 32, 236–243. doi: 10.1109/TASSP.1984.1164317

Inoue, K., Hara, K., Lala, D., Nakamura, S., Takanashi, K., and Kawahara, T. (2019). "A job interview dialogue system with autonomous android Erica," in *Int'l Workshop Spoken Dialogue Systems (IWSDS)*.

Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). "Audio augmentation for speech recognition," in *Proceedings of INTERSPEECH*. doi: 10.21437/Interspeech.2015-711

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Myer, S., and Tomar, V. S. (2018). Efficient keyword spotting using time delay neural networks. *arXiv preprint arXiv:1807.04353*. doi: 10.21437/Interspeech.2018-1979

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). "Librispeech: an ASR corpus based on public domain audio books," in *Proceedings of IEEE-ICASSP*. doi: 10.1109/ICASSP.2015.7178964

Peddinti, V., Povey, D., and Khudanpur, S. (2015). "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proceedings of INTERSPEECH*. doi: 10.21437/Interspeech.2015-647

Povey, D., et al. (2011). "The Kaldi speech recognition toolkit," in *Proceedings of IEEE-ASRU*.

Povey, D., Zhang, X., and Khudanpur, S. (2015). "Parallel training of deep neural networks with natural gradient and parameter averaging," in *Proceedings of ICLR Workshop*.

Qin, Y., Carlini, N., Cottrell, G., Goodfellow, I., and Raffel, C. (2019). "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, *Vol. 97*, 5231–5240.

Rabiner, L. (1988). A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE* 77, 257–286. doi: 10.1109/5.18626

Schönherr, L., Kohls, K., Zeiler, S., Holz, T., and Kolossa, D. (2019). Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. doi: 10.14722/ndss.2019.23288

Sun, M., Snyder, D., Gao, Y., Nagaraja, V. K., Rodehorst, M., Panchapagesan, S., et al. (2017). "Compressed time delay neural network for small-footprint keyword spotting," in *Interspeech*, 3607–3611. doi: 10.21437/Interspeech.2017-480

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., et al. (2017). "Attention is all you need," in *31st Conference on Neural Information Processing Systems (NIPS 2017)* (Long Beach, CA).

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. (1989). Phoneme recognition using time-delay neural networks. *IEEE/ACM Trans. ASLP* 37, 328–339. doi: 10.1109/29.21701

Wang, H., Liang, C., Wang, S., Chen, Z., Zhang, B., Xiang, X., et al. (2023). "Wespeaker: a research and production oriented speaker embedding learning toolkit," in *Proceedings of IEEE-ICASSP*, 1–5. doi: 10.1109/ICASSP49357.2023.10096626

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., et al. (2018). "Espnet: end-to-end speech processing toolkit," in *Proceedings of INTERSPEECH*. doi: 10.21437/Interspeech.2018-1456

Yakura, H., and Jun, S. (2019). "Robust audio adversarial example for a physical attack," in *International Joint Conferences on Artificial Intelligence Organization (IJCAI)*. doi: 10.24963/ijcai.2019/741

Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Liu, X., et al. (2009). "The HTK book version 3.4.1," in *Tutorial Books*.

Yuan, X., Chen, Y., Zhao, Y., Long, Y., Liu, X., Chen, K., et al. (2018). "Commandersong: a systematic approach for practical adversarial voice recognition," in *USENIX*, 49–64.

Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T., and Xu, W. (2017). "Dolphinattack: inaudible voice commands," in *ACM CCS*, 103–117. doi: 10.1145/3133956.3134052