INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Adversarial Audio: A New Information Hiding Method

Yehao Kong, Jiliang Zhang[*]

Cyberspace Security Research Center, Peng Cheng Laboratory, Shenzhen, China
College of Computer Science and Electronic Engineering, Hunan University, China
zhangjiliang@hnu.edu.cn

Abstract

Audio is an important medium in people's daily life, hidden information can be embedded into audio for covert communication. Current audio information hiding techniques can be roughly classified into time domain-based and transform domain-based techniques. Time domain-based techniques have large hiding capacity but low imperceptibility. Transform domain-based techniques have better imperceptibility, but the hiding capacity is poor. This paper proposes a new audio information hiding technique which shows high hiding capacity and good imperceptibility. The proposed audio information hiding method takes the original audio signal as input and obtains the audio signal embedded with hidden information (called stego audio) through the training of our private DNN-based automatic speech recognition (ASR) model. The experimental results show that the proposed audio information hiding technique has a high hiding capacity of 48 cps with good imperceptibility and high security.

Index Terms: Information hiding, Imperceptibility, Automatic Speech Recognition, DNN

1. Introduction

With the rapid development of communication-related technologies, multimedia information is generated in large quantities and brings great convenience to people. However, multimedia information services pose a potential threat to the legitimate rights of the information owner. As a different technology from traditional cryptography, information hiding technique [1] is considered to be able to provide technical protection for the rights of multimedia information.

As an important medium in people's daily life communication, the audio has a good imperceptibility in the transmission of information and provides a lot of redundant space for embedding hidden information, making the research of audio information hiding techniques valuable.

Traditional audio information hiding techniques can be roughly divided into two classes: time domain-based and transform domain-based techniques. Time domain technique directly embeds the information into the carrier in the time domain. In general, it has large hiding capacity but low imperceptibility. The commonly used time domain-based techniques include the least significant bit (LSB) [2], echo hiding [3] and spread spectrum [4] techniques. Transform domain techniques modify the parameters in the transform domain to hide information, which have better imperceptibility but poor hiding capacity. Commonly used transform domain-based techniques include phase coding [5], discrete cosine transform (DCT) [6] and discrete wavelet transform (DWT) [7] techniques.

In order to maintain the hiding capacity and imperceptibility, this paper proposes to embed the hidden information to the audio signal by the private ASR model based on DNN. Experiment results show that our proposed information hiding technique has good hiding capacity, imperceptibility and security. The contributions are as follows.

- **Novel hiding approach.** We propose a new audio information hiding technique based on the adversarial perturbations, which embeds and extracts the hidden information by the DNN-based ASR model.
- **High hiding capacity.** The proposed technique embeds the hidden information with a hiding capacity of 48 character per second (cps).
- **Good imperceptibility.** The value of perceptual evaluation of speech quality (PESQ) is 3.598 on average. Human can barely perceive the perturbation.
- **High security.** Four public ASR models such as Google and IBM are used to test the stego audios and experimental results show that these models are unable to extract the hidden information.

2. Related Works

2.1. Adversarial Examples

Deep learning, especially neural networks has shown great advantages in the fields of image recognition, speech processing and autonomous driving, etc. In particular, the recognition ability of image recognition models has exceeded the accuracy of human eye. However, recent researches have shown that DNN models are vulnerable to adversarial examples [8, 9]. Adversarial example (AE) is carefully designed by attackers to fool deep learning models. The difference between the AEs and real examples is almost indistinguishable by the human eye, but it can cause the model to be misclassified.

The majority of AE researches focused on generating AEs against image recognition models [8, 10, 11, 12, 13, 14, 15]. Recently, Zhang et al. [9] compared the existing adversarial examples generation methods in detail and elaborated the opportunities and challenges of adversarial examples. However, recent researches have shown that speech recognition models are also vulnerable to AEs [16, 17, 18]. Carlini et al. [17] proposed a method for generating audio AEs against a white-box ASR model. It can produce very strong audio AEs, resulting in a misclassification rate of up to 100%. Yuan et al. [19] proposed CommanderSong to generate audio AEs by using ASR model Kaldi. However, CommanderSong can only apply to some special cases, which will not work for the complex end-to-end deep learning speech recognizing system like DeepSpeech.

The vulnerability to AEs threatens the security of DNN-based ASR models. However, we don't focus on attacking the ASR models in this paper. On the contray, we introduce this characteristic into the field of audio information hiding for se-

[*] Corresponding author

Copyright © 2020 ISCA

http://dx.doi.org/10.21437/Interspeech.2020-1294

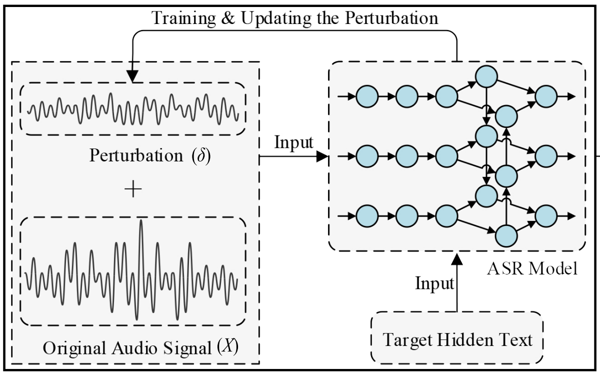
Training & Updating the Perturbation
Perturbation (δ)
+
Original Audio Signal (X)
Input
ASR Model
Input
Target Hidden Text
Output
(X+δ)
Stego Audio Signal (X')

Figure 1: *The process of embedding the hidden information into audio signals.*

curity and trust communications, which transforms the AE from treatening security to protecting security.

3. The Proposed Method

The paper proposes a new technique based on the adversarial examples for audio information hiding, which embeds and extracts the hidden information by the private DNN-based ASR model. The technique is described in detail below.

3.1. Embedding Method

The process is shown in Fig. 1. First, input the original audio signal *X* and the hidden text *t* into the ASR model. In training phase, the slight perturbation δ that needs to be added to the *X* is constantly updated according to the result of the loss function. Finally, the generated stego audio signal $X + \delta$ can be recognized as the hidden text t with a small perturbation δ. For example, input an audio signal that is “Good morning”, and through training the model, the output of ASR is finally recognized as “see you at 5pm”.

In order to recognize the audio as the hidden information, the CTC-loss is selected as the loss function of our method, which can output a probability for any text given an audio signal. The detail principle of CTC-loss can be found in [20].

In the meantime, for better imperceptibility, the perturbation δ added to the original audio should be quieter than the original audio signal *X*, hence its value should be smaller than the original one. To make a lower computational complexity, the L infinite norm $\|\delta\|_\infty$ is used to represent the magnitude of perturbation.

Converting the overall goal into an optimization problem is to minimize the $\|\delta\|_\infty$ in the case where the model *C(·)* recognizes the speech *(X+δ)* as the target text *t* (i.e., $C(X + \delta) = t$), that is,

$$\begin{aligned} &\min\ \|\delta\|_\infty \\ &s.t.\ C(X + \delta) = t. \end{aligned} \tag{1}$$

Therefore, the optimization problem becomes two minimize problems: 1) minimize the perturbation $\|\delta\|_\infty$; 2) minimize the loss function $l(X + \delta, t)$, which indicates the magnitude of the CTC loss between the recognition result of $X+\delta$ and the target text t. For facilitating the application of the gradient optimizer, we separate them into two steps that keep iterating.

1. Calculate the δ that satisfies $C(X + \delta) = t$ by applying gradient descent optimization to $l(X + \delta, t)$;
2. Reduce the range of δ and clip it into the range.

The two steps keep iterating until reaching the set threshold of iteration times. For the step 1, the gradient optimization of δ is performed by using Adam Optimizer to make the recognition result of $X + \delta$ close to the target text t gradually. For the step 2, a threshold τ is set for δ to ensure the maximum fluctuation range of δ will not exceed the threshold. The two steps can be integrated to an iterative function Eq. (2):

$$\begin{cases} \delta_0 = 0,\ X_0 = X + \delta_0 \\ \delta_{N+1} = clip_{\delta,\tau}(\nabla_\delta l(X_N,\ t)),\ X_{N+1} = X + \delta_{N+1} \end{cases} \tag{2}$$

Algorithm 1 Information Embedding Algorithm

```
Input: Original audio signal $X$, Hidden text $t$
Output: Stego audio signal $X'$
Initialize: $\delta$−an initial zero array with the same shape of $X$,
    $\tau$−the threshold of $\delta$, $N$−the max iteration times
$X' = X + \delta$
for $i = 0, 1, 2, \ldots, N$ do
    //Calculate the loss
    $L = l(X', t)$
    //Update $\delta$
    $\delta \leftarrow$AdamOptimizer.$minimize(L, \delta)$
    $\delta = clip(\delta, -\tau, \tau)$
    $X' = X + \delta$
    if $C(X') == t$ then
        //Update the threshold $\tau$
        if $max(\delta) \leq \tau$ then
            $\tau = max(\delta)$
        end if
        $\tau = 0.8 \cdot \tau$
        //Save the last best result
        $temp = X'$
    end if
end for
return $temp$
```

The detailed algorithm is shown in Algorithm 1. The function $clip(\delta, -\tau, \tau)$ sets the values of δ larger than τ become τ, and values smaller than $-\tau$ become $-\tau$. The minimization process is repeated until reaching the number of iterations we set. Finally, the last best result will be returned.

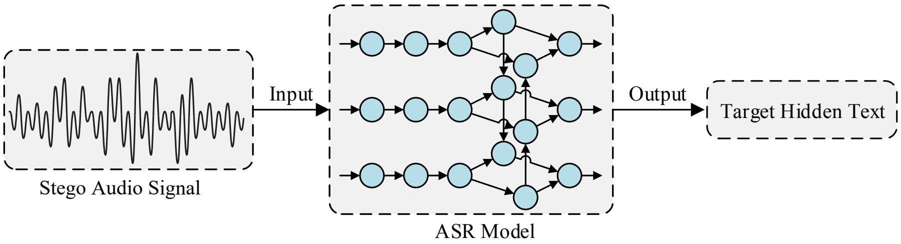

Figure 2: *The process of extracting the hidden information from stego audio signals.*

3.2. Extracting Method

Compared with traditional audio information hiding methods, our proposed method does not need to use any complicated algorithm to process the stego audio signal. As shown in Fig. 2, the hidden text can be obtained by simply inputting the stego audio signal into the private ASR model to recognize. In order to ensure that other public ASR models cannot identify the hidden text, four state-of-art ASR models are used to extract the hidden text from the stego audios. The experimental results show that in addition to the private ASR model, all other public models cannot get any content related to the hidden text, the test results can be seen in Section 4.3.

4. Experiments and Results

Table 1: *The specific hidden information in different groups.*

Group	Audio Range	Hidden Information
G1	A00-A09	be quiet
G2	A10-A19	sing louder
G3	A20-A29	close the door
G4	A30-A39	the key is one one nine
G5	A40-A49	call the police
G6	A50-A59	happy birthday to you
G7	A60-A69	be careful
G8	A70-A79	bob is the spy
G9	A80-A89	help me
G10	A90-A99	see you at five pm

Hiding capacity, imperceptibility, security and robustness are the main performance indicators of audio information hiding techniques [21, 22]. To evaluate the performance of the proposed audio information hiding technique, 100 test audios (A00 - A99) are selected from the Mozilla common voice dataset [23], which are wav files with a length of 3 seconds, sampled at the rate of 16 kHz and quantized with 16 bits, to embed the hidden information. We divide these audios into 10 groups G1-G10. The specific information to be hidden in different groups is shown in Table 1. The stego audios are generated with tensorflow and DeepSpeech v0.1.0 version. The initial parameters we set in the experiments are as follows. The iteration times N is 500, the initial δ is an array of 0 with the same shape of the audio signal, and the initial τ is set to 3000.

The hiding capacity and imperceptibility are compared with a recently proposed spread spectrum-based audio information hiding method [4]. We have implemented this audio information hiding method in MATLAB using the original configuration in the [4].

4.1. Hiding Capacity Analysis

Hiding capacity, also known as the hiding rate, is the amount that hidden information can be embedded in the carrier signal per second. Character per second (cps) is used as the unit of hiding capacity here.

Since the DeepSpeech model divides the audio signal into 50 frames per second when extracting speech features, which indicates that up to 50 characters can be recognized per second. Thus, the theoretical maximum capacity of this information hiding method is 50 cps. We conduct a hiding capacity test on the ten groups. The experimental results are shown in Table 2, and the average hiding capacity is 48.0 cps. In the meantime, the hiding capacity of method in [4] is a fixed value of 84 bps. As 1 character equals to 8 bits, the capacity of [4] is 10.5 cps. Therefore, our proposed method has a higher hiding capacity.

4.2. Imperceptibility Analysis

In this paper, perceptual evaluation of speech quality (PESQ) is used to perform imperceptible analysis of audio signals. PESQ is an objective mean opinion score (MOS) value evaluation method provided by ITU-T Recommendation P.862, which uses the stego audio to compare with the original audio. In general, the score is between 1.0 and 4.5. The worse the speech quality, the lower the score.

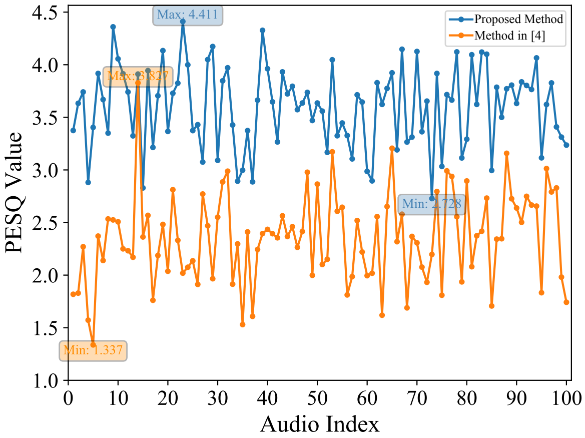

Figure 3: *The PESQ value of stego audios for the proposed method and the method in [4].*

The tested PESQ value of the 100 stego audios are shown in Fig. 3, the average PESQ value of our proposed method is 3.598, while the method in [4] is 2.351, which means that our method has good imperceptibility.

Table 2: *The hiding capacity of stego audios*

Proposed Method											Method in [4]
Group	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	Avg	84 bps = 10.5 cps
Capacity(cps)	47.9	48.2	48.0	46.6	48.6	48.8	48.8	47.6	46.8	48.6	48.0	

Table 3: *The extraction success rate of different ASR models*

Group	Model Internal Security		Model External Security		
	DeepSpeech v0.1.0	DeepSpeech v0.2.0	Google Cloud	IBM Watson	iFlytek
G1-G10	100%	0%	0%	0%	0%

4.3. Security Analysis

The security of audio information hiding refers to the ability that the hidden information cannot be extracted by the attacker. In this paper, the original model (DeepSpeech v0.1.0) is used as the private model.

4.3.1. Model Internal Security Analysis

The model internal security analysis is to find out if the model will output the same result while the weights of model are different. We evaluate the security by comparing the extraction success rate from the private model to its upgraded version model DeepSpeech v0.2.0. The two DNN models have different neuron weights while holding the same neural network structure. The extracting results are shown in Table 3. The extraction success rate of DeepSpeech v0.1.0 is 100% while the v0.2.0 is 0%.

4.3.2. Model External Security Analysis

The model external security analysis is to find out if the model will output the same result while the whole model structure and parameters are different. We evaluate the model external security by comparing the extraction success rate from the private model to other ASR models. Three public commercialized ASR platform services Google Cloud [24], IBM Watson [25] and iFlytek [26] Speech-to-Text are selected to extract the hidden information in different groups.The extraction success rates are shown in Table 3. The extraction success rates of the 3 state-of-the-art ASRs are 0%.

From the above results, it can be seen that only the private model can extract the hidden information successfully. Even the same model cannot extract hidden information after the model parameters are updated (i.e., DeepSpeech v0.2.0). In addition, according to the specific extraction information during the experiment, any content related to the hidden text cannot be obtained at all. Therefore, the security of this audio information hiding method is high.

4.4. Robustness Analysis

Table 4: *The extraction success rate after 4 signal processing methods*

Gaussian Noise	Resampling	Lowpass Filtering	Echo Interference
0%	46%	1%	1%

The robustness of audio information hiding refers to the ability that the hidden text can be completely extracted after suffering some modification or transformation. To test the robustness, the stego audios are processed as follows:

1. Gaussian white noise: A Gaussian white noise with a signal-noise ratio (SNR) of 20 dB is added to the stego audio signals;
2. Resampling attack: Resample the stego audio signals by 2 times the original sampling rate, and then restored to the original sampling rate;
3. Low-pass filtering: The Butterworth low-pass filter with a 2-order cutoff frequency of 6 kHz is processed for stego audio signals;
4. Echo interference: Add an echo with a 50% attenuation rate and a delay of 30ms in the stego audio signals.

As shown in Table 4, except for the resampling attack, the stego audio signals have almost lost the hidden text after being processed by these methods. The experimental results show that the robustness of our proposed hiding technique is not good. Therefore, in order to enable the receiving end to extract the hidden text successfully, the stego audio signals can only be transmitted in a lossless propagation, for example, to upload the audio file.

5. Conclusions

The paper proposes a novel technique for audio information hiding based on AEs, which takes the original audio signal as input and obtains the stego audio through the training process of the private ASR model. According to experimental results, the generated stego audio signal has a hiding capacity of 48.0 cps with good imperceptibility and high security.

However, our proposed new audio information hiding technique is not robust enough. At current stage, the stego audio signals can only be transmitted in a lossless propagation. We expect to provide a new solution for audio information hiding, and gradually address the shortcoming in further research.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No.61874042, the Hunan Natural Science Foundation for Distinguished Young Scholars under Grant No.2020JJ2010, the Hu-Xiang Youth Talent Program under Grant No.2018RS3041, the Key Research and Development Program of Hunan Province under Grant No.2019GK2082 and the CCF-Tencent Open Fund.

7. References

[1] R. Anderson, Ed., *Information Hiding*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 1996, vol. 1174.

[2] S. Jadhav and A. M. Rawate, "A new audio steganography with enhanced security based on location selection scheme," *International Journal of Performability Engineering*, vol. 12, no. 5, pp. 451–458, 2016.

[3] G. Hua, J. Goh, and V. L. L. Thing, "Time-Spread Echo-Based Audio Watermarking With Optimized Imperceptibility and Robustness," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 227–239, feb 2015.

[4] Y. Xiang, I. Natgunanathan, D. Peng, G. Hua, and B. Liu, "Spread Spectrum Audio Watermarking Using Multiple Orthogonal PN Sequences and Variable Embedding Strengths and Polarities," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 529–539, mar 2018.

[5] N. M. NGO and M. UNOKI, "Method of Audio Watermarking Based on Adaptive Phase Modulation," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 1, pp. 92–101, 2016.

[6] M. Jeyhoon, M. Asgari, L. Ehsan, and S. Z. Jalilzadeh, "Blind audio watermarking algorithm based on DCT, linear regression and standard deviation," *Multimedia Tools and Applications*, vol. 76, no. 3, pp. 3343–3359, feb 2017.

[7] D. Avci, T. Tuncer, and E. Avci, "A new information hiding method for audio signals," in *2018 6th International Symposium on Digital Forensic and Security (ISDFS)*. IEEE, mar 2018, pp. 1–4.

[8] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations (ICLR)*, apr 2014.

[9] J. Zhang and C. Li, "Adversarial Examples: Opportunities and Challenges," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 7, pp. 2578–2593, july 2020.

[10] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *3nd International Conference on Learning Representations (ICLR)*, may 2015.

[11] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *5nd International Conference on Learning Representations (ICLR)*, april 2017.

[12] N. Papernot, P. Mcdaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," *Proceedings - 2016 IEEE European Symposium on Security and Privacy, EURO S and P 2016*, pp. 372–387, 2016.

[13] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2016, pp. 2574–2582.

[14] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," *Proceedings - IEEE Symposium on Security and Privacy*, pp. 39–57, 2017.

[15] X. Liu, J. Zhang, Y. Lin, and H. Li, "ATMPA: Attacking machine learning-based malware visualization detection methods via adversarial examples," in *Proceedings of the International Symposium on Quality of Service*. New York, NY, USA: ACM, jun 2019, pp. 1–10.

[16] D. Iter, J. Huang, and M. Jermann, "Generating adversarial examples for speech recognition," 2017.

[17] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *Proceedings - 2018 IEEE Symposium on Security and Privacy Workshops, SPW 2018*, 2018, pp. 1–7.

[18] R. Taori, A. Kamsetty, B. Chu, and N. Vemuri, "Targeted adversarial examples for black box audio systems," in *IEEE Deep Learning and Security Workshop*, may 2019.

[19] X. Yuan, Y. Chen, Y. Zhao, Y. Long, X. Liu, K. Chen, S. Zhang, H. Huang, X. Wang, and C. A. Gunter, "Commandersong: A systematic approach for practical adversarial voice recognition," in *Proceedings of the 27th USENIX Conference on Security Symposium*, ser. SEC'18. Berkeley, CA, USA: USENIX Association, 2018, pp. 49–64.

[20] A. Hannun, "Sequence modeling with ctc," *Distill*, 2017.

[21] F. Petitcolas, R. Anderson, and M. Kuhn, "Information hiding-a survey," *Proceedings of the IEEE*, vol. 87, no. 7, pp. 1062–1078, jul 1999.

[22] G. Simmons, "The history of subliminal channels," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 4, pp. 452–462, may 1998.

[23] Mozilla, "Common Voice." [Online]. Available: https://voice.mozilla.org/datasets

[24] Google, "Google Cloud Speech-to-Text." [Online]. Available: https://cloud.google.com/speech-to-text/

[25] IBM, "IBM Watson Speech-to-Text." [Online]. Available: https://speech-to-text-demo.ng.bluemix.net/

[26] IFlytek, "iFlytek Speech-to-Text." [Online]. Available: https://www.iflyrec.com/html/addMachineOrder.html