INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Exploring Targeted Universal Adversarial Perturbations to End-to-end ASR Models

Zhiyun Lu, Wei Han, Yu Zhang, Liangliang Cao

Google Inc., USA
{zhiyunlu,weihan,ngyuzh,llcao}@google.com

Abstract

Although end-to-end automatic speech recognition (e2e ASR) models are widely deployed in many applications, there have been very few studies to understand models' robustness against adversarial perturbations. In this paper, we explore whether a targeted universal perturbation vector exists for e2e ASR models. Our goal is to find perturbations that can mislead the models to predict the given targeted transcript such as "thank you" or empty string on any input utterance. We study two different attacks, namely additive and prepending perturbations, and their performances on the state-of-the-art LAS, CTC and RNN-T models. We find that LAS is the most vulnerable to perturbations among the three models. RNN-T is more robust against additive perturbations, especially on long utterances. And CTC is robust against both additive and prepending perturbations. To attack RNN-T, we find prepending perturbation is more effective than the additive perturbation, and can mislead the models to predict the same short target on utterances of arbitrary length.

Index Terms: end-to-end ASR, adversarial examples, model robustness

1. Introduction

Adversarial example [1, 2, 3] is a carefully crafted input that fools the network into wrong predictions by applying a worst-case perturbation to a test sample. Adversarial perturbations are extensively studied in vision [1, 4] and language [5, 6]. In audio domain, [7, 8, 9] studies adversarial examples for sound/speech classification and speaker identification problems, where the output is a single class instead of a label sequence.

Adversarial example for automatic speech recognition (ASR) models [10, 11, 12, 13, 14, 15] is substantially different from those in the image domain [15], due to the sequential predictions. For example, an *untargeted* adversarial attacks can easily yield high word error rate (WER) to ASR by introducing spelling errors [14], or by inserting random words to the transcript, while preserving much of the ground truth's text or semantics. This is very different from the attack to image classification models, where a high error rate guarantees that users can not relate the mis-classification with the ground truth. To this end, we study *targeted* attack in this work. In addition to high WER, the ASR model should predict a specific mis-transcription target. Targeted attack is much more challenging [13, 16] than untargeted attack.

A large portion of previous works focus on input-dependent perturbations [13, 14], where for each utterance we solve a separate optimization to compute the perturbation. On the other hand, a *universal* (audio-agnostic) perturbation is one perturbation vector, learnt from training set, which can generalize and cause *any* audio to be mis-transcribed with high probability. Universal perturbation can attack unseen utterances in real-time without any new optimization. As it is more efficient and practical, we focus on universal perturbations. Figure 1 illustrates targeted universal perturbation: for any utterance, a universal perturbation is applied to the utterance. The ASR model is fooled to predict a specific transcript, e.g. "thank you", regardless of the input. Both prepending and additive perturbations are studied in the experiments.

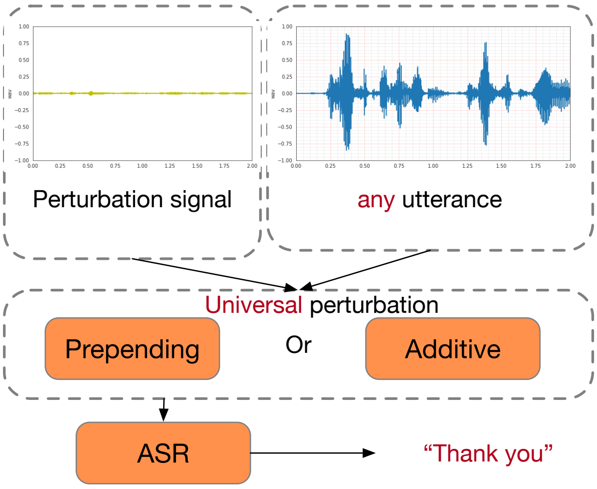


Figure 1: *System overview: We apply a universal perturbation to the utterance,which causes the ASR model to output the same transcript,* e.g. *"thank you", for* ***any*** *utterance. The input utterance can be of different lengths and arbitrary content. We study both prepending and additive perturbations.*

One challenge of adversarial attack to ASR models lies in the fast evolution of models. In the past five years, various e2e ASR models have been deployed on both servers and mobile devices, including CTC [17], LAS [18] and RNN-T [19], and RNN-T models have pushed the-state-of-the-art [20] on the Librispeech dataset [21]. In contrast, most of the studies on ASR adversarial attacks are limited to a single type of model. For example, [14] focuses on LAS model, and [13, 22] on CTC.

This paper aims to conduct a comprehensive study on the targeted universal adversarial perturbations against different ASR models. There are two main discoveries from our research. Firstly, we show that universal perturbation exists for e2e ASR models on some given targets. The learnt universal perturbation can fool the state-of-the-art ASR models to generate the same mis-transcription target on Librispeech [21] test sets[1]. On short sentences composed of frequent words, the success rate of the perturbation can be over 99%. For long targets, the success rate will be much lower. Secondly, we find dif-

[1] We encourage the reader to listen to the perturbation examples on this page: zhiyun.github.io/Targeted-Universal-Adversarial-Perturbations-to-End-to-end-ASR-Models/.

http://dx.doi.org/10.21437/Interspeech.2021-1668

ferent e2e models behave differently against the perturbations. We compared two perturbations, additive perturbation, studied in [14, 22], and prepending perturbation, a new type of perturbation for speech data. We find LAS model is vulnerable to both additive and prepending perturbations, while RNN-T is more robust against additive perturbation but becomes vulnerable to prepending perturbations. For a state-of-the-art RNN-T model trained from Librispeech, we can append a 4 seconds long audio clip to utterances of arbitrary lengths, and the ASR model would always generate the same wrong predictions. Lastly, CTC is robust against both perturbations. Note that our work is not concerned about the imperceptibility of the perturbation, as motivated in [23].

This paper is organized as follows: § 2 introduces the learning framework of the target universal perturbation. In § 3 we describe the data as well as the evaluation metrics of the experimental study. Extensive experiments on different e2e models are discussed in § 4, and § 5 concludes the paper.

2. Method

We formally define the targeted universal perturbation problem in § 2.1, and discuss the perturbation function in § 2.2. Lastly we provide the architecture details of the e2e models used in the empirical study in § 2.3.

2.1. Targeted universal adversarial perturbation

ASR is a model the predicts a text sequence given an input utterance. Define $\boldsymbol{x} = [x_1, \dots, x_T]$ as the input sequence, and $\boldsymbol{\delta} = [\delta_1, \dots, \delta_S]$ as the perturbation sequence, where S and T are the sequence lengths. We denote the perturbed input as $\mathcal{T}(\boldsymbol{\delta}, \boldsymbol{x})$, which applies perturbation $\boldsymbol{\delta}$ to $\boldsymbol{x}$ through operation $\mathcal{T}$. We will discuss the form of $\mathcal{T}(\boldsymbol{\delta}, \boldsymbol{x})$ in next section. The text sequence is denoted as y. For an input-output pair $(\boldsymbol{x}, y)$, we denote the loss of the ASR model as $\ell(\boldsymbol{x}, y)$. Here we ignore the dependency on model parameters for brevity since the parameters are fixed in the attack problem. Note ℓ is different for different models: cross-entropy for attention model, RNN-T loss for RNN-T model, and CTC loss for CTC model.

For targeted perturbation, we specify a mis-transcription y'. Our goal is to find a universal perturbation $\boldsymbol{\delta}$ that can attack the ASR system to output y' for any utterances $\boldsymbol{x}$, drawn from some distribution or dataset $\mathcal{D}$. Mathematically a sufficient condition is that $\ell(\mathcal{T}(\boldsymbol{\delta}, \boldsymbol{x}), y')$ is small across all samples $\forall \boldsymbol{x} \in \mathcal{D}$. Therefore the targeted audio-agnostic adversarial attack can be formulated as the following optimization problem

$$\min_{\boldsymbol{\delta}} \sum_{x \in \mathcal{D}} \ell(\mathcal{T}(\boldsymbol{\delta}, \boldsymbol{x}), y'), \quad \text{s.t. } \|\boldsymbol{\delta}\|_\infty < \epsilon, \tag{1}$$

where ϵ is the maximum allowed L_∞ norm of the perturbation, following [14, 13, 22]. Since we are not concerned about imperceptibility in this work, we choose ϵ to be 2^{15}, the signal range for a 16-bit PCM format audio, for most of the experiments. While more sophisticated loss functions specific to architectures are possible, like the one discussed in [13], we try to keep a simple and unified framework that works for all models in this study. We solve the minimization problem by stochastic gradient descent.

2.2. Perturbation $\mathcal{T}(\boldsymbol{\delta}, \boldsymbol{x})$

While perturbation on the frequency domain is possible, we focus on generating adversarial perturbations in the audio-domain.

Table 1: *Number of parameters and WERs on Librispeech test sets of e2e ASR models. The last column shows the WER of the test sets with 4 seconds of silence prepended to each utterance. It serves as an baseline for the prepending perturbation.*

model	# params (M)	WER clean / other	WER pp-4s clean / other
LAS	10.3	4.9 / 7.9	11.9 / 13.8
RNN-T	10.4	2.6 / 6.4	4.5 / 9.3
CTC	8.9	3.6 / 8.7	3.6 / 8.8

Additive perturbation. This is the perturbation studied in previous works [7, 13, 14]. It is a direct adaptation of image perturbation [4] to audio domain. $\boldsymbol{\delta}$ is truncated or padded with 0 to the same length of $\boldsymbol{x}$ and then added to $\boldsymbol{x}$. Formally,

$$\mathcal{T}(\boldsymbol{\delta}, \boldsymbol{x}) = [x_1 + \bar{\delta}_1, x_2 + \bar{\delta}_2, \dots, x_T + \bar{\delta}_T], \tag{2}$$

where $\bar{\delta}_t = \delta_t$ when $t \leq S$, and 0 otherwise.

Prepending perturbation. We propose a new adversarial perturbation, observing that speech is of sequential nature and that the input is of variable length. Prepending perturbation is defined as

$$\mathcal{T}(\boldsymbol{\delta}, \boldsymbol{x}) = [\delta_1, \dots, \delta_S, x_1, \dots, x_T]. \tag{3}$$

where $\boldsymbol{\delta}$ is concatenated before $\boldsymbol{x}$. Note that prepending perturbation does not contaminate the utterance at all except that the model runs for a few steps with $\boldsymbol{\delta}$ as input before seeing $\boldsymbol{x}$. This perturbation attacks e2e models' stability over time, while additive perturbation measures the networks' additive stability [1].

2.3. Models

We focus on the e2e ASR models, and study three most popular architectures: RNN-T [19], attention model [18], and CTC [17]. For RNN-T model, we use the state-of-the-art Conformer (small) model introduced in [20]. It has an encoder of 16 layers conformer blocks of 144 dimensions, and a single LSTM layer decoder with 1024 hidden units and a 640 units projection layer. The joint network has 640 units. For attention model, we use the same encoder and decoder as in the RNN-T model, and with a multi-head attention of 4 heads and hidden size 288. For CTC model, the encoder is the same as RNN-T, but the decoder is a simple projection layer of dimension 1024. For all three models, we extracted 80-channel filterbanks features computed from a 25ms window with a stride of 10ms. The tokenizer is a 1k word-piece model built from LibriSpeech960h. We use SpecAugment [24] during training. Please refer to [20] for more details about the training hyper-parameters. All models and the attack are implemented and trained with Lingvo toolkit [25]. The number of parameters in the model and the WERs on Librispeech test-clean and test-other sets are given in Table 1. To provide a baseline for the prepending perturbation, we report the WER of the test sets with 4 seconds of silence (numerical 0) prepended to each utterance in the last column. Note that there is a large degradation of WER for LAS, mostly dominated with deletion error. The deletion error on test-clean and test-other are 9.5%/7.7% respectively.

3. Experimental Setup

3.1. Dataset

LibriSpeech960h dataset [21] is used to train the perturbations as well as the e2e ASR models (see § 2.3) in our experiments. It is a corpus of 16KHz English speech from audiobooks. The training set has 281,241 utterances. We report the attack results on Librispeech test-clean and test-other sets, which contain 2620 and 2939 utterances respectively.

Throughout our experiments, the perturbation is of 4 seconds long, for both additive and prepending types. For the attack optimization[2], we use Adam [26] optimizer. The learning rate is 1.0 for the first 50k steps, and exponentially decays after 50k steps. The batch size is 1024. We set the the perturbation max norm constraint ϵ in Eq. (1) to be 2^{15}, which is the signal range for a 16-bit PCM format audio, if not otherwise specified. Our main focus is to study the feasibility of universal perturbation, similar as in [23], and imperceptibility is not our concern.

3.2. Evaluation metrics

To evaluate the performance of the targeted attack, following existing works [13, 14] we report, (1) Success rate (sentence-level accuracy): N_s/N where N_s is the number of audios that are transcribed as y', and N is the total number of audios in the test set; (2) Distortion: We quantify the relative loudness of the perturbation $\boldsymbol{\delta}$ with respect to the original audio $\boldsymbol{x}$ in decibels (dB): $D(\boldsymbol{\delta}, \boldsymbol{x}) = \text{dB}(\boldsymbol{\delta}) - \text{dB}(\boldsymbol{x})$ where $\text{dB}(\boldsymbol{x}) = 20\log_{10}(\max_t(x_t))$. The success rate is the higher the better, while the distortion is the lower the better.

4. Experimental Results

There are three variables that determine a targeted attack: the perturbation (see § 2.2), the model architecture (see § 2.3), and the mis-transcription target y'. In the experiment study, we try to answer the following questions: (a) which ASR model is vulnerable to targeted attacks; (b) what perturbation $\mathcal{T}$ is effective to attack ASR models; (c) what mis-transcription y' is more likely to be attacked. To answer (a), we attack three models under additive perturbation and prepending perturbation in § 4.1 and § 4.2 respectively. Contrasting results in § 4.1 and § 4.2 answers question (b). We vary different transcripts in § 4.3 for question (c). Lastly we experiment with different max-norm constraints in § 4.4 and discuss baselines in § 4.5.

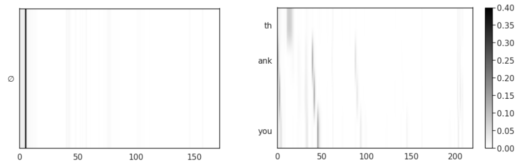
4.1. Attack with additive perturbation

Table 2 shows the attack results of 4 second additive perturbation across three models. For both targets empty string (∅) and "thank you", LAS model achieves almost 100% success with relatively small perturbation. The failure of LAS comes from that the attention is fooled to only focus on the beginning of the speech where the perturbation is applied and ignores the rest of the speech, thus generates the wrong target transcript. See Fig. 2 for an illustration.

However, when the decoder is time-synchronous, *i.e.* RNN-T and CTC, the attack is less successful. Despite a very large perturbation magnitude, the success rate of RNN-T is less than 20% for empty target ∅, and less than 3% for "thank you" target. The success rate of CTC is less than 6% for ∅, and 0% for "thank you". The reason is that it is challenging to attack arbitrary length utterance by adding a fixed length perturbation. When the utterance is longer than the added perturbation, it is hard to alter the output of those clean speech frames for RNN-T and CTC decoders. To verify this, we plot the attack success rate across different utterance lengths for RNN-T model in Fig. 3 (a). The target is ∅. We can see that the success rate drops significantly when $\boldsymbol{x}$ is longer than 4 seconds, which is the length of the perturbation.

Table 2: *Attack results of 4 seconds additive perturbation on different models. LAS model can be attacked with almost 100% accuracy. RNN-T and CTC models are more robust.*

target y'	model	dB ↓	success (%) ↑
∅	LAS	-7.69 / -6.82	99.35 / 99.66
	RNN-T	6.00 / 6.87	14.66 / 19.02
	CTC	6.00 / 6.87	3.28 / 4.80
"thank you"	LAS	-0.55 / 0.32	98.74 / 99.52
	RNN-T	6.00 / 6.87	1.85 / 2.75
	CTC	6.00 / 6.87	0.00 / 0.00



(a) $y' = \varnothing$ (b) $y' =$ "thank you"

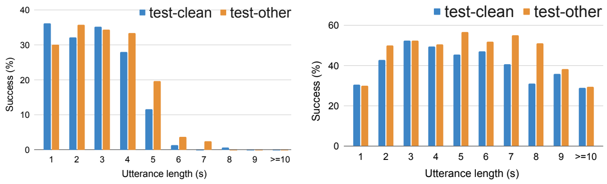
Figure 2: *Heatmap of the attention probability of a perturbed input example on LAS. The x-axis is the encoder frame, and the y-axis is the output token. The attention only focuses on the beginning of the utterance. It quickly emits the target y' and EOS, and skips the rest of the frames. The attention is the culprit for LAS vulnerability under adversarial perturbations.*

4.2. Attack with prepending perturbation

Before we present the attack results, please see the last column of Table 1 for the WER of a simple baseline, which pads 4 seconds of silence (numerical 0) before the utterance. We can see that CTC model is stable with respect to the silence padding. There are some WER degration on RNN-T and LAS, but silence padding does not fail the ASR for most of the time.

In Table 4, we show the attack results of prepending perturbation. Again for both targets ∅ and "thank you", LAS model achieves almost 100% success. The perturbation magnitude of prepending perturbation is smaller than that of additive perturbation, when we compare the first row of Table 4 and Table 2.

On RNN-T model, prepending perturbation achieves higher success rate with smaller perturbation magnitude, compared to additive perturbation in Table 2. And if we plot out the attack success rate across different utterance lengths, it is almost



(a) additive perturbation (b) prepending perturbation

Figure 3: *Attack success rate across different length of utterances. The model is RNN-T and the target is $y' = \varnothing$. The perturbations are 4 seconds long. For additive perturbation, the success rate drops significantly when the length of the utterance is longer than the perturbation. For prepending perturbation, the success rate is relatively uniform across different lengths.*

[2] We did not encounter small gradient issue after properly tuning the batch size, thus the vanilla SGD is preferred over the Fast Gradient Sign Method [2] variant for simplicity.

Table 3: *Example ASR output of untargeted attack of 4 seconds additive perturbation on LAS. The bold part shows that untargeted attack can keep the same text from ground truth in the prediction, despite high WER.*

Hyp:	and but martha and his wife's rather planted with	**himself**	for the first of the	**better**	deserts of	**anything else he said**
Ref:	i really liked that account of	**himself**		**better**	than	**anything else he said**
Err:	*ins/del*	*correct*	*ins*	*correct*	*sub*	*correct*

Table 4: *Attack results of 4 seconds prepending perturbation on different models. Prepending perturbation is more effective than additive perturbation in terms of success rate and distortion on both LAS and RNN-T.*

target y'	model	dB ↓	success (%) ↑
∅	LAS	-30.86 / -29.99	99.54 / 99.39
	RNN-T	-3.51 / -2.64	41.18 / 47.43
	CTC	-4.72 / -3.85	1.26 / 0.99
"thank you"	LAS	-23.52 / -22.65	99.92 / 99.93
	RNN-T	-0.24 / 0.63	35.65 / 37.80
	CTC	-0.27 / 0.60	0.00 / 0.07

Table 5: *Comparing different target with prepending perturbation on LAS. The shorter the transcript, and the more frequent the words, the easier the attack can succeed.*

target y'	dB ↓	success (%) ↑
∅	-30.86 / -29.99	99.54 / 99.39
"hello"	-23.58 / -22.71	99.77 / 100.00
"thank you"	-23.52 / -22.65	99.92 / 99.93
"carpe diem"	-11.25 / -10.38	95.69 / 92.04
"to be or not to be"	-18.11 / -17.24	66.87 / 71.76

uniform across different lengths as seen in Fig. 3 (b). We hypothesize that prepending perturbation leverages a different failure mode than the additive instability in RNN-T model. By prepending the perturbation, we let the decoder runs into bad RNN states and fail the alignment in RNN-T. It makes the attack agnostic to the length of the utterance.

Lastly, CTC model is robust to the prepending attack due to the conditional independence modeling assumption.

4.3. Attack results on more mis-transcript targets

In Table 5, we discuss the results of different targets y' with prepending perturbation on LAS. We choose LAS as the model and prepending perturbation because they are relatively "easier" to learn, according to results from Table 2 and 4. As can be seen from Table 5, the general trend is that the success rate drops when y' gets longer, like ''to be or no to be", and when y' consists of less frequent words, like "carpe diem" in Latin. It is worth mentioning that the learnt perturbation contains phonetic similar sound as the ones of the target y'. When the target y' is long, it is hard to be attacked. For example, "to be or no to be that is the question" achieves 1.57% and 2.01% attack accuracy on test-clean and test-other sets, and "the gods had condemned sisyphus to ceaselessly rolling a rock to the top of a mountain whence the stone would fall back of its own weight" achieves 0% accuracy on both test sets. Similar observation is discussed for per-utterance targeted attack [13].

4.4. Attack results under different max-norm constraints

Lastly, we vary the max norm constraint ϵ and report the attack results. In Table 6, we can see that the success rate drops when the ϵ gets smaller.

Table 6: *Attack results under different max-norm constraints. The perturbation is 4 second prepending perturbation, and the model is RNN-T. The success rate drops when ϵ become smaller.*

$\|\boldsymbol{\delta}\|_\infty$	dB ↓	success (%) ↑
2^{15}	-3.51 / -2.64	41.18 / 47.43
4000	-12.25 / -11.38	28.63 / 31.30
2000	-18.25 / -17.38	3.32 / 3.74

4.5. Comparing with other adversarial attacks

In this section, we briefly compare our study with two existing works [14, 22]. Note that this paper studies a different problem from theirs, since the perturbation is input-agnostic as well as targeted. In addition, we experiment with both additive and prepending perturbations, while [14, 22] are limited to additive perturbations. However, it would be intuitive to compare different attacks and discuss why their method cannot solve the universal targeted perturbation problem.

Targeted per-utterance perturbation. [14] solves $\min_{\boldsymbol{\delta}} \ell(\mathcal{T}(\boldsymbol{\delta}, \boldsymbol{x}), y')$, s.t. $\|\boldsymbol{\delta}\|_\infty < \epsilon$, for a $(\boldsymbol{x}, y')$ utterance-text pair[3]. We randomly pick 5 utterances from the Librispeech100h clean training set to learn 5 additive perturbations on LAS, which is the same length as its corresponding training utterance. The target is empty string. However, for the perturbation trained from a single utterance, the success rates on test-clean is $0.7\% \pm 0.7\%$ and test-other $2.7\% \pm 3.7\%$, averaging over the 5 perturbations. It shows that the perturbation optimized for one utterance can hardly generalize to other unseen utterances.

Untargeted universal perturbation. [22] studies universal perturbation for *untargeted* attack. The loss is $\max_{\boldsymbol{\delta}} \sum_{x \in \mathcal{D}} \ell(\mathcal{T}(\boldsymbol{\delta}, \boldsymbol{x}), y)$, s.t. $\|\boldsymbol{\delta}\|_\infty < \epsilon$, where y is the ground-truth transcription. The goal is to find a perturbation that maximally deviates the predictions from the ground-truth, but the error is not specified. We implement their universal attack on Librispeech960h. The WERs in percentage (%) on test-clean/test-other for 3 models are: LAS 76.0/100.5, RNN-T 87.5/100.6, and CTC 106.1/121.2. Table. 3 shows an example of such attack. Despite high WER, the predictions on the untargeted perturbation oftentimes preserves some of the ground truth's texts. In contrast, our targeted attack will keep no words from the original ground-truth.

5. Conclusion

In this work, we study the problem of targeted universal adversarial perturbations. We compare the performance of both additive and prepending perturbations against three state-of-the-art e2e ASR models, and suggest that targeted universal attacks exist for both LAS and RNN-T, but not for CTC models. Our future work is to develop more robust e2e models based on the lessons learnt from this study.

[3] For simplicity, here we omit the irrelevant optimization over the imperceptibility and robustness of the perturbation.

6. References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” *arXiv preprint arXiv:1312.6199*, 2013.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.

[3] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *2017 ieee symposium on security and privacy (sp)*. IEEE, 2017, pp. 39–57.

[4] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1765–1773.

[5] M. Alzantot, Y. Sharma, A. Elgohary, B.-J. Ho, M. Srivastava, and K.-W. Chang, “Generating natural language adversarial examples,” *arXiv preprint arXiv:1804.07998*, 2018.

[6] R. Jia and P. Liang, “Adversarial examples for evaluating reading comprehension systems,” *arXiv preprint arXiv:1707.07328*, 2017.

[7] S. Abdoli, L. G. Hafemann, J. Rony, I. B. Ayed, P. Cardinal, and A. L. Koerich, “Universal adversarial audio perturbations,” *arXiv preprint arXiv:1908.03173*, 2019.

[8] J. Vadillo and R. Santana, “Universal adversarial examples in speech command classification,” *arXiv preprint arXiv:1911.10182*, 2019.

[9] Y. Xie, C. Shi, Z. Li, J. Liu, Y. Chen, and B. Yuan, “Real-time, universal, and robust adversarial attacks against speaker recognition systems,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1738–1742.

[10] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou, “Hidden voice commands,” in *25th {USENIX} Security Symposium ({USENIX} Security 16)*, 2016, pp. 513–530.

[11] H. Abdullah, W. Garcia, C. Peeters, P. Traynor, K. R. Butler, and J. Wilson, “Practical hidden voice attacks against speech and speaker recognition systems,” *arXiv preprint arXiv:1904.05734*, 2019.

[12] M. Alzantot, B. Balaji, and M. Srivastava, “Did you hear that? adversarial examples against automatic speech recognition,” *arXiv preprint arXiv:1801.00554*, 2018.

[13] N. Carlini and D. Wagner, “Audio adversarial examples: Targeted attacks on speech-to-text,” in *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2018, pp. 1–7.

[14] Y. Qin, N. Carlini, G. Cottrell, I. Goodfellow, and C. Raffel, “Imperceptible, robust, and targeted adversarial examples for automatic speech recognition,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 5231–5240.

[15] H. Abdullah, K. Warren, V. Bindschaedler, N. Papernot, and P. Traynor, “Sok: The faults in our asrs: An overview of attacks against automatic speech recognition and speaker identification systems,” *arXiv e-prints*, pp. arXiv–2007, 2020.

[16] M. Cisse, Y. Adi, N. Neverova, and J. Keshet, “Houdini: Fooling deep structured prediction models,” *arXiv preprint arXiv:1707.05373*, 2017.

[17] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[18] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, “Listen, attend and spell,” *arXiv preprint arXiv:1508.01211*, 2015.

[19] A. Graves, “Sequence transduction with recurrent neural networks,” *arXiv preprint arXiv:1211.3711*, 2012.

[20] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, “Conformer: Convolution-augmented transformer for speech recognition,” *arXiv preprint arXiv:2005.08100*, 2020.

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

[22] P. Neekhara, S. Hussain, P. Pandey, S. Dubnov, J. McAuley, and F. Koushanfar, “Universal adversarial perturbations for speech recognition systems,” *arXiv preprint arXiv:1905.03828*, 2019.

[23] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, “Adversarial patch,” *arXiv preprint arXiv:1712.09665*, 2017.

[24] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “Specaugment: A simple data augmentation method for automatic speech recognition,” *arXiv preprint arXiv:1904.08779*, 2019.

[25] J. Shen, P. Nguyen, Y. Wu, Z. Chen, M. X. Chen, Y. Jia, A. Kannan, T. Sainath, Y. Cao, C.-C. Chiu *et al.*, “Lingvo: a modular and scalable framework for sequence-to-sequence modeling,” *arXiv preprint arXiv:1902.08295*, 2019.

[26] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.