Transferable Adversarial Perturbations between Self-Supervised Speech Recognition Models

Raphael Olivier [1] **Hadi Abdullah** [2 3] **Bhiksha Raj** [1]

Abstract

A targeted adversarial attack produces audio samples that can force an Automatic Speech Recognition (ASR) system to output attacker-chosen text. To exploit ASR models in real-world, black-box settings, an adversary can leverage the *transferability* property, i.e. that an adversarial sample produced for a proxy ASR can also fool a different remote ASR. Recent work has shown that transferability against large ASR models is extremely difficult. In this work, we show that modern ASR architectures, specifically ones based on Self-Supervised Learning, are uniquely affected by transferability. We successfully demonstrate this phenomenon by evaluating state-of-the-art self-supervised ASR models like Wav2Vec2, HuBERT, Data2Vec and WavLM. We show that with relatively low-level additive noise achieving a 30dB Signal-Noise Ratio, we can achieve target transferability with up to 80% accuracy. We then use an ablation study to show that Self-Supervised learning is a major cause of that phenomenon. Our results present a dual interest: they show that modern ASR architectures are uniquely vulnerable to adversarial security threats, and they help understanding the specificities of SSL training paradigms.

1. Introduction

Adversarial audio algorithms are designed to force Automatic Speech Recognition (ASR) models to produce incorrect outputs. They do so by introducing small amounts of imperceptible, carefully crafted noise to benign audio samples that can force the ASR model to produce incorrect transcripts. Specifically, *targeted* adversarial attacks (Carlini & Wagner, 2018; Qin et al., 2019) are designed to force ASR models to output any target sentence of the attacker's choice. However, these attacks have limited effectiveness as they make unreasonable assumptions (e.g., white-box access to the model weights), which are unlikely to be satisfied in real-world settings.

An attacker could hypothetically bypass this limitation by using the *transferability* property of adversarial samples: they generate adversarial samples for a white-box proxy model; then pass these to a different remote black-box model, as we illustrate in Figure 1. Transferability has been successfully demonstrated in other machine learning domains, like computer vision (Papernot et al., 2016). This is a sample text in black. Yet for ASR, recent work has shown that transferability is close to non-existent between large models (Abdullah et al., 2021b), even between *identically* trained models (i.e., same training hyper-parameters, even including the random initialization seed). These findings were demonstrated on older ASR architectures, specifically on LSTM-based DeepSpeech2 models (et al., 2016). However, robustness properties sometimes vary considerably between different ASR architectures (Lu et al., 2021; Olivier & Raj, 2022), and it is worth studying adversarial transferability on more recent families of models.

In this work, we study the robustness of modern transformer-based ASR architectures. We show that, in contrast with previously evaluated architectures, many state-of-the-art ASR models are in fact vulnerable to the transferability property. Specifically, our core finding can be formulated as follows:

Pretraining transformer-based ASR models with Self-Supervised Learning (SSL) makes them vulnerable to transferable adversarial attacks.

SSL is an increasingly popular learning paradigm in ASR (Figure 2), used to boost model performance by leveraging large amounts of unlabeled data. We demonstrate that it hurdles robustness by making the following contributions:

- First, we show that most public SSL-pretrained ASR

[*]Equal contribution [1]Language Techologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA [2]Visa Research, Atlanta, GA, USA [3]Florida Institute for Cybersecurity, University of Florida, Gainesville, FL, USA. Correspondence to: Raphael Olivier <rolivier@cs.cmu.edu>.

2^{nd} *AdvML Frontiers workshop at* 40^{th} *International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023.

models are vulnerable to transferability. We generate 85 adversarial samples for the proxy HuBERT and Wav2Vec2 models (Section 3). We show that these samples are effective against a wide panel of public transformer-based ASRs. *This includes ASRs trained on different data than our proxies*.Our attacks can achieve a Signal-Noise Ratio of 30dB - not quite as imperceptible as the strongest white-box attacks (Qin et al., 2019), yet surprising for transferable attacks.

- Second, we show that SSL-pretraining is the reason for this vulnerability to transferability. We do so using an ablation study on Wav2Vec2-type models, either pretrained or trained from scratch, and similar in all other aspects. We use each model as a proxy to generate adversarial examples, which we attack all other models with. We show that SSL pretraining in both proxy and private models indeed contributes to the transferability of adversarial attacks and that factors such as the amount of unlabeled data play an important role.

Our work has important implications for ASR security: we show that SSL, a line of work gathering attention in the speech community, is a source of vulnerability in adversarial settings. Formerly innocuous attacks with unreasonable assumptions are now effective against many modern models. As it is likely that SSL will be used to train ASR systems in production, our results pave the way for practical, targeted attacks in the real world. By no means do these results imply that this line of work should be aborted, but they emphasize the pressing need to focus on robustness alongside performance.

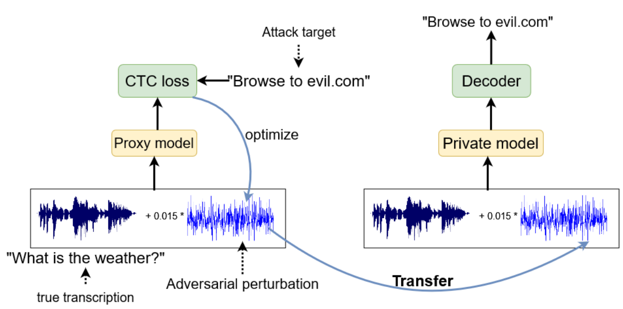


Figure 1. Diagram illustrating the transferability of an adversarial attack between a proxy and a private model

2. Background

2.1. SSL pretraining for ASR models

We describe in this Section the principles of SSL-pretrained ASR models, whose robustness to attacks we evaluate in this work. These models usually follow the neural architecture of Wav2Vec2 (Baevski et al., 2020). Raw audio inputs are

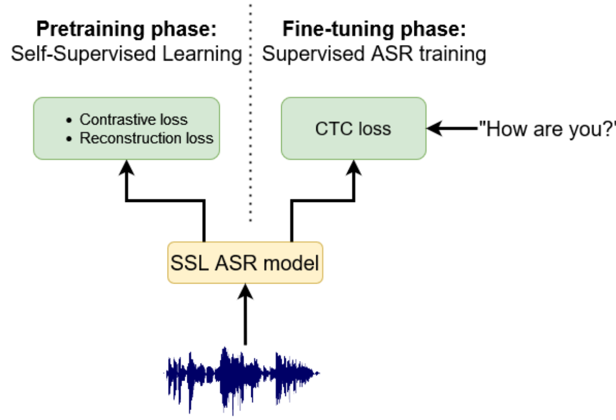


Figure 2. Diagram illustrating the training procedure of SSL ASR models

fed directly to a CNN. A Transformer encodes CNN outputs into contextualized representations, that a final feed-forward network projects in a character output space. The model is fine-tuned with CTC loss (Graves et al., 2006).

A number of different models follow this architecture, including Wav2Vec2, HuBERT (Hsu et al., 2021), Data2Vec (Baevski et al., 2022), UniSpeech-SAT (Wang et al., 2021; Chen et al., 2021b) or WavLM (Chen et al., 2021a). These networks only have very minor differences in their architectures. Base models have 12 transformer hidden layers and 90M parameters. Large models have 24 layers and 300M parameters. Finally, XLarge models have 48 layers for a total of 1B parameters.

While the networks are similar, the training pipelines of these models differ substantially. All models are pretrained on large amounts of unlabeled data, then fine-tuned for ASR on varying quantities of labeled data. The pretraining involves SSL objectives, such as Quantization and Contrastive Learning (Wav2Vec2), offline clustering and masked predictions (HuBERT), or masked prediction of contextualized labels (Data2Vec). Unispeech combines SSL and CTC pretraining with multitask learning. WavLM adds denoising objectives and scales to even greater amounts of unlabeled data.

SSL pretraining is helpful in many regards: it makes the same network easy to fine-tune for multiple downstream tasks with little labeled data and has improved state-of-the-art results in ASR benchmarks, especially in low-resource settings. As we demonstrate, it is also a source of vulnerabilities.

2.2. Adversarial attacks

Adversarial examples are inputs modified imperceptibly by an attacker to fool machine learning models (Szegedy et al., 2014; Goodfellow et al., 2014; Carlini & Wagner, 2016; Madry et al., 2018). While most works have focused on image classification, several proposed attacks for other tasks such as ASR (Cisse et al., 2017; Carlini & Wagner, 2018; Qin et al., 2019).

The attack we use is based on the Carlini&Wagner ASR attack (Carlini & Wagner, 2018), although slightly simplified. Given an input x, a target transcription y_t, and an ASR model f trained with loss L, our attack finds an additive perturbation δ optimizing the following objective:

$$\arg\min_{\delta} L(f(x+\delta), y_t) + c\left\|\delta\right\|_2^2 \ s.t. \ \left\|\delta\right\|_\infty < \epsilon \quad (1)$$

which we optimize using L_∞ Projected Gradient Descent. While the CW attack typically uses a large initial ϵ, then gradually reduces it as it finds successful perturbations, we fix a single value of ϵ and optimize for a fixed number of iterations. We find that this scheme, closer to the PGD algorithm (Madry et al., 2018), greatly improves attack transferability. However we keep using CW's L_2 regularization term $c\left\|\delta\right\|_2^2$.

We also find that applying regularization such as dropout during attack optimization greatly helps to generate transferable perturbations. This effect is analyzed more in detail in Appendix C.3. Throughout the rest of the paper, we run all attack optimization steps using the default dropout, layer drop, etc. that the proxy model used during training (typically a dropout of 0.1).

3. A Transferable attack on ASR models

In our core experiment, we fool multiple state-of-the-art SSL-pretrained ASR models with targeted and transferred adversarial attacks. We generate a small set of targeted audio adversarial examples using fixed proxy models. We then transfer those same examples on a large number of models available in the HuggingFace Transformers library. Table 1 specifies how much unlabeled and labeled data these models were trained on. We provide full experimental details in Appendix A.

3.1. Generating adversarial examples on proxies

We describe our procedure to generate adversarial examples. To maximize the transferability success rate of our perturbations we improve the base attack in Section 2.2 in several key ways:

- To limit attack overfitting on our proxy, we combine the losses of *two* proxy models: Wav2Vec2 and HuBERT (Large). Both models were pretrained on the entire LV60k dataset and finetuned on 960h of LibriSpeech. As these models have respectively a contrastive and predictive objective, they are a representative sample of SSL-pretrained ASR models. The sum of their losses is used as the optimization objective in Equation 1.
- We use 10000 optimization steps, which is considerable (for comparison (Carlini & Wagner, 2018) use 4000) and can also lead to the adversarial noise overfitting the proxy models. To mitigate this effect we use a third model, the Data2Vec BASE network trained on LibriSpeech, as a stopping criterion for the attack. At each attack iteration, we feed our adversarial example to Data2Vec, and keep track of the best-performing perturbation (in terms of WER). We return that best perturbation at the end of the attack.

  Because this procedure is computationally expensive, we only apply it to a subset A of 85 utterances of less than 7 seconds. We sample them randomly in the LibriSpeech test-clean set. We select attack targets at random: we sample a completely disjoint subset B of utterances in the LibriSpeech test-other set. To each utterance in A we assign as target the transcription of the sentence in B whose length is closest to its own. This ensures that a very long target isn't assigned to a very short utterance or vice versa.

3.2. Transferring adversarial examples on ASR

We evaluate all SSL-pretrained models mentioned in Section 2.1, along with several others for comparison: the massively multilingual speech recognizer or M-CTC (Lugosch et al., 2022) trained with pseudo-labeling, The Whisper large model (Radford et al., 2022) trained for ASR and Speech translation on 680khrs of (unreleased) multilingual data, and models trained from scratch for ASR: the Speech-to-text model from Fairseq (Wang et al., 2020), the CRDNN and Transformer from SpeechBrain (Ravanelli et al., 2021)

3.3. Metrics

We evaluate ASR performance with the **Word-Error-Rate** (WER) between model outputs and ground truth.

When evaluating the success of adversarial examples, we can also use the Word-Error-Rate. Between the prediction and the attack target y_t, a low WER indicates a successful attack. We define the **targeted attack success rate** as

$$\text{TASR} = \max(1 - \text{WER}(f(x+\delta), y_t), 0) \quad (2)$$

It is also interesting to look at the results of the attack in terms of *denial-of-service*, i.e. the attack's ability to stop the model from predicting the correct transcription y. Here a high WER indicates a successful attack. We define the **untargeted attack success rate** as

$$\text{UASR} = \min(\text{WER}(f(x+\delta), y), 1) \quad (3)$$

Model	Unlabeled data	Labeled data	Clean WER	Attack success rate	
				targeted	untargeted
Wav2Vec2-Large	LV60k	LS960	2.0%	88.0%	100%
HuBERT-Large	LV60k	LS960	1.9%	87.2%	100%
Data2Vec-Base	LS960	LS960	2.5%	63.4%	100%
Wav2Vec2-Base	LS960	LS960	2.6%	**55.7%**	**100%**
Wav2Vec2-Base	LS960	LS100	3.4%	**53.9%**	**100%**
Wav2Vec2-Large	LS960	LS960	2.3%	**50.7%**	**100%**
Data2Vec-Large	LS960	LS960	1.9%	**66%**	**100%**
HuBERT-XLarge	LV60k	LS960	1.8%	**80.9%**	**100%**
UniSpeech-Sat-Base	LS960	LS100	3.5%	**50.4%**	**100%**
WavLM-Base	LV60k+VoxPopuli+GS	LS100	2.9%	**21.7%**	**100%**
Wav2Vec2-Large	LV60k+CV+SB+FSH	LS960	3.3%	**67.3%**	**100%**
Wav2Vec2-Large	LV60k+CV+SB+FSH	SB	6.3%	**41.5%**	**100%**
Wav2Vec2-Large	CV-multi	CV-multi	15.6%	**17.7%**	**100%**
Wav2Vec2-Large	CV-en	CV-en	7.69%	**19.7%**	**100%**
Wav2Vec2-Large	CV-fr	CV-fr	100%	0%	100%
M-CTC-Large	None	CV (en)	21.7%	7.5%	76.4%
Speech2Text	None	LS960	3.5%	7.3%	63.3%
SB CRDNN	None	LS960	2.9%	5.9%	86.4%
SB Transformer	None	LS960	2.3%	6.5%	90.6%
Whisper Large	None	680khrs	2.3%	0%	43.8%

Table 1. Results of the transferred attack on different ASR models (SNR $= 30$dB). The first three lines correspond to the proxies used to generate the adversarial examples. On all other models, the adversarial examples are transferred. We report for each model how much data was used for SSL pretraining and ASR finetuning. We also report its Word-Error-Rate on the LibriSpeech test-clean set, and the targeted and untargeted word-level attack success rate (see Section 3.3)

Finally, we control the amount of noise in our adversarial examples with the Signal-Noise Ratio (SNR), defined as

$$\mathrm{SNR}(\delta, x) = 10 \log(\frac{\|x\|_2^2}{\|\delta\|_2^2}) \tag{4}$$

for an input x and a perturbation δ. When generating adversarial examples we adjust the L_∞ bound ϵ (equation 1) to achieve a target SNR.

3.4. Results

We report the results of our adversarial examples in Table 1 for $\epsilon = 0.015$, corresponding to a Signal-Noise Ratio of 30dB on average. In Appendix C.1 we also report results for a larger ϵ value.

On 12 out of 17 models, we observe that the attack achieves total denial-of-service: the untargeted success rate is 100%. Moreover, on the first 6 models (proxies aside), the targeted attack success rate ranges between 50% and 81%: the target is more than half correctly predicted! These results are in flagrant contradiction with past works on DeepSpeech2-like models, where even the slightest change in training leads to a total absence of targeted transferability between proxy and private model. Our private models vary from the proxies in depth, number of parameters and even training methods, yet we observe important transferability. However, these 6 models have all been pretrained on LibriSpeech or Libri-Light with SSL pretraining, i.e. the same data distribution as our proxies.

The following five models were pretrained on different datasets. One was pretrained on a combination of Libri-Light, VoxPopuli and GigaSpeech; two on Libri-Light, CommonVoice, SwitchBoard and Fisher; and two on CommonVoice either multilingual or English. The transferability success rate on these five models ranges from 18% to 67%, which is significant. Even the CommonVoice models, whose training data has no intersection with Libri-Light, are partially affected.

Although our inputs and attack targets are in English, we apply them to a French-only CommonVoice Wav2vec2. This model, incapable of decoding clean LibriSpeech data, is also unaffected by our targeted perturbation. It therefore seems that, while multilingual models are not robust to our examples, a minimal performance on the original language is required to observe transferability.

The final 5 models for which the targeted transferability

Model \Proxy	None LS960	LS960 LS960	LV LS960	LV-CV-SB-FSH LS960	LV-CV-SB-FSH SB300	CV CV
None LS960	*99.1%*	0%	0%	0%	0%	0%
LS960 LS960	0%	*30.7%*	9.1%	5.5%	0%	0%
LV LS960	0%	0%	*84.3%*	44.1%	18.5%	0%
LV-CV-SB-FSH LS960	0%	0%	55.2%	*76.0%*	*33.9%*	0%
LV-CV-SB-FSH SB300	0%	0%	9.7%	*53.7%*	*62.3%*	0%
CV CV	0%	0%	11.9%	8.4%	7.4%	*91.1%*

Table 2. Word-level success rate of the attack with different proxies and models. Each row corresponds to a different proxy, each column to a different private model. The format is [pretraining-data fine-tuning-data]. All models follow the Wav2Vec2-Large architecture.

rate is null or close to null, are those that were not SSL-pretrained at all (including M-CTC which was pretrained with pseudo-labeling). These four models also partially resist the untargeted attack.

It emerges from these results that some recent ASR models, specifically those pretrained with SSL, can be vulnerable to transferred attacks. These results diverge significantly from previous works like (Abdullah et al., 2021b; 2022a) which showed no transferability between different models. Table 1 hints that SSL plays an large role in transferability. The next section establishes stronger evidence of that hypothesis.

4. Ablation study

In this section, we conduct a thorough ablation study and quantify to what extent SSL pretraining makes ASR models vulnerable to transferred attacks. We also measure the influence of several other factors on transferability. This ablation study requires the generation of many sets of adversarial examples, using varying models as proxy. Since those attacks have an important computational cost, we reduce the number of forward/backward passes in out attack. We run the attack in Section 2.2 with one proxy model at a time and 1000 optimization steps.

We choose particularly relevant sets of models for this ablation study. In Appendix B, we report a full results table of cross transferability between all models we have access to.

4.1. Influence of self-supervised learning

In this section, we compare Wav2Vec2 Large models with varying pretraining data. We consider models pretrained on LibriVox (60khrs), LibriSpeech (960h), CommonVoice english (1087hrs), an ensemble of LibriVox, CommonVoice, Fisher and SwitchBoard ($\sim$ 64khrs), or none at all. We use each model both as a proxy to generate adversarial noise and as a private model for evaluation with other proxies. We train the Wav2Vec2 model with no pretraining data locally, by using the Wav2Vec2 finetuning recipe in Fairseq (Ott et al., 2019) with random initialization. This model achieves 7.6% WER on the LibriSpeech clean test set.

We report the success rate of our attacks in Table 2. These results show unambiguously that SSL pretraining plays a huge role in the transferability of adversarial attacks. Models pretrained on small datasets (LibriSpeech, CommonVoice, none) are ineffective as proxy, with success rates of 0% consistently; while the other models are partially successful, with transferability success rates in the 0-55% range depending on the private model. In addition, the unpretrained model is the only one on which all proxies entirely fail. Within our range of study, we can summarize the results of Table 2 as follows:

- SSL pretraining in the proxy *and* the attacked model are necessary conditions for attack transferability
- Increasing the amount of pretraining data in the proxy increases the transferability success rate

In other words the vulnerability of ASR models to adversarial attacks without gradient access worsened rather than mitigated by increasing amounts of pretraining data. The effect of SSL pretraining on transferability is easily observable with the evolution of the target loss while generating adversarial examples: we plot that loss in Appendix C.3.

4.2. Model size and training hyperparameters

We now extend our ablation study to models pretrained with different SSL paradigms. We report the results in Table 3. We observe that adversarial examples also transfer between models trained with different paradigms. Using the same pretraining method in proxy and attacked model does not appear to be a critical factor in transferability: at equal data, some proxies are consistently more effective than others across all attacked models. The HuBERT Large model (pretrained on 60kh) is the best proxy by a large margin.

In Appendix C, we evaluate the impact of additional factors such as attack radius, regularization and the usage of language models with ASR.

Model \Proxy	W2V2 Base LS960	W2V2 Large LS960	W2V2 Large LV60	D2V Base LS960	D2V Large LS960	HB Large LV60	HB XLarge LV60
W2V2 Base LS960	*84.72%*	8.49%	9.48%	0%	0%	19.17%	8.98%
W2V2 Large LS960	11.0%	*30.69%*	1.84%	0%	0	19.02	5.87
W2V2 Large LV60	0	0	*88.61*	0	0%	39.18%	11.88%
D2V Base LS960	0%	0%	0.92%	*94.13%*	0%	14.29%	0.14%
D2V Large LS960	0%	0%	15.21%	0%	*94.53%*	38.61%	18.6%
HB Large LV60	0%	0%	17.47%	0%	0%	*87.98%*	28.71%
HB XLarge LV60	0%	0%	14.99 %	0%	0%	55.3%	*62.38%*

Table 3. Attack success rate with different proxies and models. Each row corresponds to a different proxy, each column to a different private model. The format is [Model-type Model-size pretraining-data] where model types are Wav2Vec2 (W2V2), Data2Vec (D2V) and HuBERT (HB). Each model was fine-tuned on 960h of LibriSpeech training data.

5. Related work

The transferability of adversarial attacks has been known for many years in Image Classification (Papernot et al., 2016). On ASR it has been limited to simple attack objectives, like preventing WakeWord detection in Alexa (Li et al., 2019) or signal processing-based attacks (Abdullah et al., 2021a; 2022b). When it comes to optimization-based attacks on large ASR models, transferability claims are usually limited and focus on untargeted attacks (Wu et al., 2022). In very specific cases there have been limited claims of targeted, transferable attacks, such as (Yuan et al., 2018); however, this work does not focus on imperceptible attacks with small amounts of noise, but rather attacks embedded in music. When it comes to standard targeted optimization attacks, (Abdullah et al., 2021b) have shown that they display no transferability on DeepSpeech2 models, even when the proxy and the attacked model are trained with identical hyperparameters apart from the initial random seed.

Past ASR adversarial attacks usually focus on a handful of neural architectures, typically DeepSpeech2 (et al., 2016), sometimes Listen Attend and Spell (Chan et al., 2016). Only recently have attacks been extended to multiple recent architectures for a fair comparison between models (Lu et al., 2021; Olivier & Raj, 2022; Wu et al., 2022). Most related to this work is Wu et al. (2022), which focuses on the vulnerability of SSL speech models. They however focus on attacking the base pretrained model with untargeted noise that remains effective on downstream tasks. We study targeted attacks, with a much deeper focus on transferability between different models. Olivier & Raj (2022) have hinted that Wav2Vec2 models are vulnerable to transferred attacks, but only report limited results on two models and do not investigate the cause of that phenomenon. We attribute it to SSL pretraining and back our claims empirically.

Abdullah et al. (2022a) have identified factors that hinder transferability for ASR attacks, such as MFCC features, Recurrent Neural Networks, and large output sizes. Since Wav2Vec2 is a CNN-Transformer model with character outputs: this gives it a better prior than DeepSpeech2 to achieve transferable adversarial attacks. However, according to that paper, this should be far from sufficient to obtain transferable attacks: our results differ for SSL-pretrained models.

6. Conclusion

We have shown that ASR targeted attacks are transferable between SSL-pretrained ASR models. Direct access to their weights is no longer required to fool models to predict outputs of the attacker's choice - and to an extent, knowledge of its training data is not required either. With that in mind, and given the existence of over-the-air attack algorithms, we expect attacks against ASR to become a practical, realistic threat as Wav2Vec2-type models are deployed in production.

In that context, it is paramount to develop adversarial defense mechanisms for ASR models. Fortunately, such defenses already exist, but they come at the cost of a tradeoff in model performance.We illustrate it in appendix D. Further research should be carried out into mitigating that tradeoff and adapting to ASR the most effective defenses in image classification, such as adversarial training.

Beyond those security aspects, our results are interesting in that they demonstrate a unique similarity between speech models representations trained very differently. They could suggest that several SSL paradigms ultimately ”converge” to the same representations, when scaled. We believe that those properties are worth exploring further and illustrate how adversarial perturbations are a powerful analysis tool for understanding neural models.

References

Abdullah, H., Rahman, M. S., Garcia, W., Blue, L., Warren, K., Yadav, A. S., Shrimpton, T., and Traynor, P. Hear “no evil”, see “Kenansville”: Efficient and Transferable

Black-box Attacks on Speech Recognition and Voice Identification Systems. In *IEEE Symposium on Security and Privacy (IEEE S&P)*, 2021a.

Abdullah, H., Warren, K., Bindschaedler, V., Papernot, N., and Traynor, P. SoK: The Faults in our ASRs: An Overview of Attacks against Automatic Speech Recognition and Speaker Identification Systems. In *IEEE Symposium on Security and Privacy (IEEE S&P)*, 2021b.

Abdullah, H., Karlekar, A., Bindschaedler, V., and Traynor, P. Demystifying limited adversarial transferability in automatic speech recognition systems. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=l5aSHXi8jG5.

Abdullah, H., Karlekar, A., Prasad, S., Rahman, M. S., Blue, L., Bauer, L. A., Bindschaedler, V., and Traynor, P. Attacks as defenses: Designing robust audio captchas using attacks on automatic speech recognition systems, 2022b. URL https://arxiv.org/abs/2203.05408.

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, 2018. URL https://arxiv.org/abs/1802.00420.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf.

Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., and Auli, M. data2vec: A general framework for self-supervised learning in speech, vision and language, 2022. URL http://arxiv.org/abs/2202.03555. cite arxiv:2202.03555.

Carlini, N. and Wagner, D. A. Towards evaluating the robustness of neural networks. *CoRR*, 2016.

Carlini, N. and Wagner, D. A. Audio adversarial examples: Targeted attacks on speech-to-text. *CoRR*, 2018.

Chan, W., Jaitly, N., Le, Q., and Vinyals, O. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4960–4964, 2016. doi: 10.1109/ICASSP.2016.7472621.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., and Wei, F. Wavlm: Large-scale self-supervised pre-training for full stack speech processing, 2021a.

Chen, S., Wu, Y., Wang, C., Chen, Z., Chen, Z., Liu, S., Wu, J., Qian, Y., Wei, F., Li, J., and Yu, X. Unispeech-sat: Universal speech representation learning with speaker aware pre-training, 2021b.

Cisse, M. M., Adi, Y., Neverova, N., and Keshet, J. Houdini: Fooling deep structured visual and speech recognition models with adversarial examples. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/d494020ff8ec181ef98ed97ac3f25453-Paper.pdf.

Cohen, J. M., Rosenfeld, E., and Kolter, J. Z. Certified adversarial robustness via randomized smoothing. *CoRR*, 2019.

et al., D. A. Deep speech 2 : End-to-end speech recognition in english and mandarin. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pp. 173–182, New York, New York, USA, 20–22 Jun 2016. PMLR.

et al., T. W. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples, 2014.

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pp. 369–376, New York, NY, USA, 2006. ACM. ISBN 1-59593-383-2. doi: 10.1145/1143844.1143891. URL http://dx.doi.org/10.1145/1143844.1143891.

Hsu, W.-N., Bolte, B., Tsai, Y.-H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, PP:1–1, 10 2021. doi: 10.1109/TASLP.2021.3122291.

Lhoest, Q., Villanova del Moral, A., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., Davison, J., Šaško, M., Chhablani, G., Malik, B., Brandeis, S., Le Scao, T., Sanh, V., Xu, C., Patry, N., McMillan-Major, A., Schmid, P., Gugger, S., Delangue, C., Matussière, T., Debut, L., Bekman, S., Cistac, P., Goehringer, T., Mustar, V., Lagunas, F., Rush, A., and Wolf, T. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 175–184, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-demo.21.

Li, J., Qu, S., Li, X., Szurley, J., Kolter, J. Z., and Metze, F. Adversarial music: Real world audio adversary against wake-word detection system. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/ebbdfea212e3a756a1fded7b35578525-Paper.pdf.

Lu, Z., Han, W., Zhang, Y., and Cao, L. Exploring Targeted Universal Adversarial Perturbations to End-to-End ASR Models. In *Proc. Interspeech 2021*, pp. 3460–3464, 2021. doi: 10.21437/Interspeech.2021-1668.

Lugosch, L., Likhomanenko, T., Synnaeve, G., and Collobert, R. Pseudo-labeling for massively multilingual speech recognition. *ICASSP*, 2022.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *ICLR 2018, Conference Track Proceedings*, 2018.

Olivier, R. and Raj, B. Sequential randomized smoothing for adversarially robust speech recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6372–6386, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.514. URL https://aclanthology.org/2021.emnlp-main.514.

Olivier, R. and Raj, B. Recent improvements of asr models in the face of adversarial attacks. *Interspeech*, 2022. URL https://arxiv.org/abs/2203.16536.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

Papernot, N., McDaniel, P. D., and Goodfellow, I. J. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *CoRR*, 2016.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance pdf.

Qin, Y., Carlini, N., Cottrell, G., Goodfellow, I., and Raffel, C. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5231–5240. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/qin19a.html.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision., 2022. URL https://cdn.openai.com/papers/whisper.pdf.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R. D., and Bengio, Y. SpeechBrain: A general-purpose speech toolkit, 2021. arXiv:2106.04624.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *ICLR*, 2014. URL http://arxiv.org/abs/1312.6199.

Tramer, F., Carlini, N., Brendel, W., and Madry, A. On adaptive attacks to adversarial example defenses. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1633–1645. Curran Associates, Inc., 2020.

Wang, C., Tang, Y., Ma, X., Wu, A., Okhonko, D., and Pino, J. fairseq s2t: Fast speech-to-text modeling with fairseq. In *Proceedings of the 2020 Conference of the Asian Chapter of the Association for Computational Linguistics (AACL): System Demonstrations*, 2020.

Wang, C., Wu, Y., Qian, Y., Kumatani, K., Liu, S., Wei, F., Zeng, M., and Huang, X. Unispeech: Unified speech representation learning with labeled and unlabeled data. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10937–10947. PMLR, 2021. URL http://proceedings.mlr.press/v139/wang21y.html.

Wu, H., Zheng, B., Li, X., Wu, X., Lee, H., and Meng, H. Characterizing the adversarial vulnerability of speech self-supervised learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pp. 3164–3168. IEEE, 2022. doi: 10.1109/ICASSP43922.2022.9747242. URL https://doi.org/10.1109/ICASSP43922.2022.9747242.

Yuan, X., Chen, Y., Zhao, Y., Long, Y., Liu, X., Chen, K., Zhang, S., Huang, H., Wang, X., and Gunter, C. A. Commandersong: A systematic approach for practical adversarial voice recognition. In *27th USENIX Security Symposium (USENIX Security 18)*, pp. 49–64, Baltimore, MD, August 2018. USENIX Association. ISBN 978-1-939133-04-5. URL https://www.usenix.org/conference/usenixsecurity18/presentation/yuan-xuejing.

A. Experimental details for LibriSpeech experiments

A.1. Frameworks

We compute adversarial examples using the robust_speech framework (Olivier & Raj, 2022). This library uses Speechbrain (Ravanelli et al., 2021) to load and train ASR models and offers implementations of various adversarial attack algorithms. Models and attacks are implemented using PyTorch (Paszke et al., 2019).

We use robust_speech for evaluation on SpeechBrain-supported models. In section 3 we export a HuggingFace Dataset (Lhoest et al., 2021), then evaluate models via the HuggingFace Transformers (et al., 2020) library. Finally, we use Fairseq (Ott et al., 2019) for training models from scratch

All of our robust_speech and Fairseq configurations are released alongside this article.

A.2. Attack Hyperparameters

We exploit the Carlini&Wagner attack (see section 2.2) implemented in robust_speech, with the following hyperparameters:

- initial ϵ: 0.015 (and 0.04 in appendix C.1)
- learning rate: 0005
- number of decreasing ϵ values: 1
- Regularization constant c: 10
- optimizer: SGD
- attack iterations: 10000 in section 3.1, 1000 in section 4

A.3. Dataset and targets

Our adversarial dataset in section 3.1 consists of 85 sentences from the LibriSpeech test-clean set. To extract these sentences we take the first 200 sentences in the manifest, then keep only those shorter than 7 seconds. In section 4, we take the first 100 sentences and filter those shorter than 14 seconds.

As attack targets, we use actual LibriSpeech sentences sampled from the test-other set. Our candidate targets are:

- Let me see how can i begin
- Now go I can't keep my eyes open
- So you are not a grave digger then
- He had hardly the strength to stammer
- What can this mean she said to herself
- Not years for she's only five and twenty
- What does not a man undergo for the sake of a cure
- It is easy enough with the child you will carry her out
- Poor little man said the lady you miss your mother don't you
- At last the little lieutenant could bear the anxiety no longer
- Take the meat of one large crab scraping out all of the fat from the shell
- Tis a strange change and I am very sorry for it but I'll swear I know not how to help it
- The bourgeois did not care much about being buried in the Vaugirard it hinted at poverty pere Lachaise if you please

For each sentence we attack, we assign the candidate target with the closest length to the sentence's original target.

A.4. Models

A.4.1. TRAINING WAV2VEC2 MODELS FROM SCRATCH

We use Fairseq to train Base and Large Wav2Vec2 models from scratch. Unfortunately, no configuration or pretrained weights have been released for that purpose, and we resort to using Wav2Vec2 fine-tuning configurations while simply skipping the pretraining step. Despite our attempts to tune training hyperparameters, we do not match the expected performance of a Wav2Vec2 model trained from scratch: (Baevski et al., 2020) report a WER of 3.0% for a large model, while we only get 9.1%.

A.4.2. GENERATING ADVERSARIAL EXAMPLES

Wav2Vec2, HuBERT and Data2Vec models are all supported directly in robust_speech and are therefore those we use for generating adversarial examples. We use the HuggingFace backend of Speechbrain for most pretrained models, and its Fairseq backend for a few (Wav2Vec2-Base models fine-tuned on 10h and 1h, and models trained from scratch). In both cases, the model's original tokenizer cannot be loaded in SpeechBrain directly. Therefore, we fine-tune the final projection layer of each model on 1h of LibriSpeech train-clean data.

The Wav2Vec2 model pretrained and fine-tuned on CommonVoice is a SpeechBrain original model. Similarly, we fine-tune it on 1h of LibriSpeech data as a shift from the CommonVoice output space to the LibriSpeech one. As a result, all our models share the same character output space.

A.4.3. EVALUATING PRETRAINED MODELS

In section 3, we directly evaluate models from HuggingFace Transformers and SpeechBrain on our adversarial dataset, without modification.

B. Full results table for cross-model LibriSpeech and LibriVox attacks

Table 4 completes the ablation study in Section 4 by evaluating all pairwise Proxy-Model combinations in our pool of Wav2Vec2-type models, trained on LibriSpeech and/or LibriLight.

C. Influence of hyperparameters on attack results

C.1. Attack radius

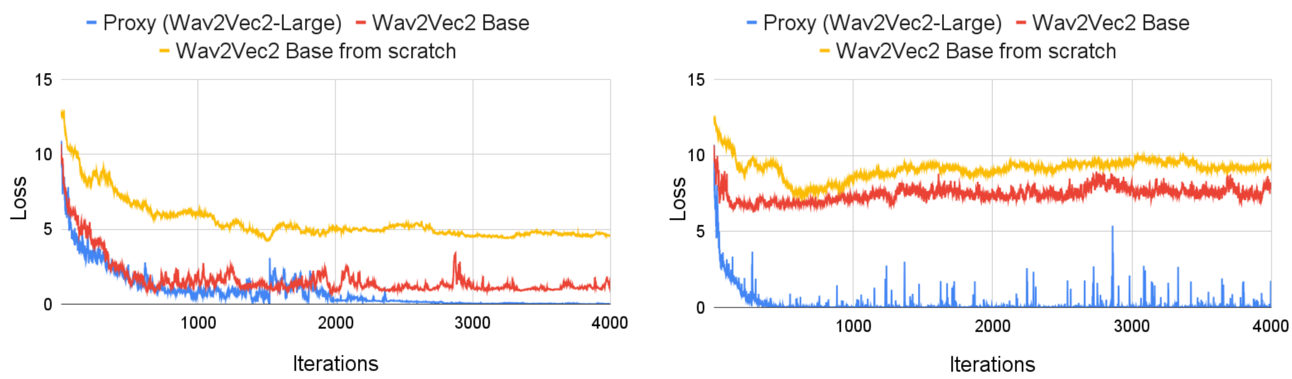
In Table 5 we extend the results of Table 1 by comparing attack results for two different attack radii. These radii are $\epsilon = 0.015$ and $\epsilon = 0.04$, corresponding respectively to Signal-Noise Ratios of 30dB and 22dB respectively. The former is identical to Table 5; the latter is substantially larger, and corresponds to a more easily perceptible noise.

Looking at the white-box attack results on the proxy models the difference is drastic: with larger noise the targeted success rate jumps from 88% to 98%. The transferred attack results on SSL-pretrained models also increase overall, with success increases ranging from 0% (Wav2Vec2-Large) to 20% (Data2Vec-Large) with a median increase of 10%. Crucially however, the targeted success does not increase at all and even decreases for ASR models trained from scratch. This confirms that there is a structural difference between the robustness of ASR models with and without SSL, that cannot be bridged simply by increasing the attack strength.

C.2. Language models

In section 3 we report the results of our adversarial dataset on multiple Wav2Vec2-type models, enhanced with an N-gram language model whenever available. In Table 6 we evaluate the influence of that language model on attack results.

We observe that the attack success rate systematically increases by 8 to 17% when adding a language model to the ASR model. This is understandable considering that our targets are sound English sentences: if a model tends to transcribe that target with mistakes, the language model can bridge that gap. To put it differently, the more prone an ASR model is to output sentences in a given distribution, the more vulnerable it is to attacks with targets sampled from that distribution. Language models are therefore more of a liability than a defense against attacks, and most likely so would be many tricks applied to an ASR model in order to improve its general performance.



(a) With dropout

(b) Without dropout

Figure 3. Evolution over attack steps of the loss on one adversarial input for three models: the Wav2Vec2 Large proxy and two targets, respectively with and without SSL pretraining. We run attacks (a) with dropout in the proxy model, and (b) without dropout in the proxy model.

C.3. Effect of model regularization on transferability

As mentioned in Section 2.2 we use regularization tricks like dropout in all proxy models when optimizing the adversarial perturbation. In Figure 3b we plot the loss on proxy and private models without that regularization, for comparison with Figure 3a. We observe that the loss degrades significantly on private models without regularization.

On the other hand, the loss on the proxy converges much faster in Figure 3b: removing model regularization makes for better, faster white-box attacks, at the cost of all transferability. To the extent of our knowledge, past work like (Carlini & Wagner, 2018) have not used regularization for generation, explaining why they report better white-box attacks than we do in terms of WER and SNR. However, as we have established above, applying regularization against standard ASR models does *not* lead to transferable adversarial examples: for that SSL pretraining is also required.

D. Defending against adversarial examples

Although we have shown that adversarial attacks can represent an important threat for private, SSL-based ASR models, it is possible to defend against them. Randomized smoothing (Cohen et al., 2019) is a popular adversarial defense that has been applied to ASR in the past (Olivier & Raj, 2021) and comes with some robustness guarantees. It consists in applying to the inputs, before feeding them to the model, amounts of random gaussian noise that are significantly larger than potential adversarial perturbations in L_2 norm. For reference we try applying it on some of our models.

We follow (Olivier & Raj, 2021) and enhance randomized smoothing with a-priori SNR estimation and ROVER voting (with 8 outputs) to boost performance. We use gaussian deviation $\sigma = 0.02$. For evaluation, we simply check the effect of our adversarial examples generated in section 3.1 on the smoothed model. A rigorous evaluation would require us to design adaptive attacks (Athalye et al., 2018; Tramer et al., 2020); since this paper does not focus on claiming robustness to attacks, we restrict ourselves to a simpler setting.

We report our results in Table 7 for the Wav2Vec2-Base, Wav2Vec2-Large and Data2Vec-Large models, pretrained and fine-tuned on 960h of LibriSpeech training data. We observe that randomized smoothing is sufficient to block the targeted attack completely (0% success rate) and recover most of the original transcription (the untargeted success rate drops to 14-34% depending on the model). However, due to the addition of gaussian noise on all inputs the defense takes a toll on the performance on clean data: the WER jumps by 4-10%. The standard deviation σ controls this tradeoff between robustness and performance; we chose the value of σ that minimizes the untargeted success rate.

Unsurprisingly, randomized smoothing is a promising protection against transferred attacks, but it does leave room for improvement. These results illustrate the need for additional research on adversarial defenses.

Model\Proxy	**W2V- B LS960 960h**		**W2V-B LS960 100h**		**W2V-B LS960 10h**		**W2V-B LS960 1h**		**D2V-B LS960 960h**	
	CER	*WER*	*CER*	*WER*	*CER*	*WER*	*CER*	*WER*	*CER*	*WER*
W2V-B LS960 960h	96.37	84.72	80.61	49.01	64.11	30.69	53.41	18.46	20.61	0
W2V-B LS960 100h	81.42	54.46	99.24	97.74	81.18	47.6	64.18	25.04	25.23	0
W2V-B LS960 10h	42.64	42.64	87.9	60.25	99.117	97.6	72.91	30.13	23.47	0
W2V-B LS960 1h	69.3	33.52	78.84	43.71	81.12	45.12	99.498	98.66	20.91	0
D2V-B LS960 960h	37.9	0	17.68	0	10.88	0	7.94	0	98.44	94.13
W2V-L LS960 960h	44.61	11	20.36	0	13.46	0	8.32	0	16.8	0
D2V-L LS960 960h	28.72	0	8.68	0	5.36	0	4.94	0	25.03	0
W2V-L LV60k 960h	29.24	0	11.12	0	5.68	0	3.19	0	13.73	0
HB-L LV60k 960h	23.83	0	7.29	0	4.83	0	3.91	0	14.92	0
HB-XL LV60k 960h	26.55	0	6.71	0	5.21	0	4.37	0	17.53	0
W2V-L CV CV+1h	27.38	0	12.59	0	11.01	0	9.61	0	19.24	0
W2V-B None 960h	7.84	0	4.45	0	4.05	0	3.83	0	5.51	0
W2V-L None 960h	8.12	0	4.63	0	4.55	0	3.44	0	5.44	0

	W2V- L LS960 960h		**D2V-L LS960 960h**		**W2V-L LV60k 960h**		**HB-L LV60k 960h**		HB-XL **LV60k 960h**	
	CER	*WER*	*CER*	*WER*	*CER*	*WER*	*CER*	*WER*	*CER*	*WER*
W2V-B LS960 960h	47.08	8.49	24.9	0	44.7	9.48	55.55	19.17	47.46	8.98
W2V-B LS960 100h	46.01	5.73	26.77	0	48.57	9.76	58.41	18.03	48.42	8.13
W2V-B LS960 10h	43.14	0	25.1	0	42.67	0	53.12	5.59	44.36	0
W2V-B LS960 1h	41.21	8.63	25.48	0.57	36.68	4.74	45.32	6.65	42.95	10.18
D2V-B LS960 960h	34.49	0	24.2	0	47.15	0.92	58.75	14.29	46.71	0.14
W2V-L LS960 960h	67.07	30.69	20.89	0	37.34	1.84	56.87	19.02	42.21	5.87
D2V-L LS960 960h	28.27	0	94.53	80.69	47.75	15.21	68.97	38.61	51.02	18.6
W2V-L LV60k 960h	25.19	0	16.05	0	97.13	88.61	71.78	39.18	46.61	11.88
HB-L LV60k 960h	27.19	0	30.08	0	49.27	17.47	97	87.98	56.83	28.71
HB-XL LV60k 960h	33.31	0	30.5	0	51.68	14.99	83.92	55.3	87.66	62.38
W2V-L CV CV+1h	27.8	0	26.85	0	56.72	11.67	46.94	0	39.95	0
W2V-B None 960h	11.19	0	9.6	0	7.16	0	6.72	0	11.07	0
W2V-L None 960h	11.15	0	9.19	0	7.52	0	7.45	0	11.23	0

	W2V- L CV CV+1h		**W2V-B None 960h**		**W2V-L None 960h**	
	CER	*WER*	*CER*	*WER*	*CER*	*WER*
W2V-B LS960 960h	10.81	0	2.62	0	2.53	0
W2V-B LS960 100h	11.01	0	2.82	0	2.58	0
W2V-B LS960 10h	11.19	0	2.65	0	2.66	0
W2V-B LS960 1h	11.81	0	3.03	0	3.04	0
D2V-B LS960 960h	8.01	0	2.32	0	2.38	0
W2V-L LS960 960h	8.2	0	2.39	0	2.54	0
D2V-L LS960 960h	8.76	0	2.44	0	2.39	0
W2V-L LV60k 960h	9.08	0	2.59	0	2.47	0
HB-L LV60k 960h	8.65	0	2.5	0	2.55	0
HB-XL LV60k 960h	8.41	0	2.49	0	2.36	0
W2V-L CV CV+1h	97.46	88.68	3.25	0	3.18	0
W2V-B None 960h	5.77	0	99.57	99.01	19.05	0
W2V-L None 960h	5.53	0	22.93	0	99.93	99.58

Table 4. Targeted Character-level and Word-level success rate for adversarial attacks when varying the proxy and the target model. All proxy-model pairs are evaluated within a pool of 13 models varying in training scheme, training data and size. The format is [Model]-[Size] [Unlabeled data] [Labeled data]. Model is equal to W2V (Wav2Vec2), D2V (Data2Vec) or HB (HuBERT). Size is equal to B (Base), L (Large) or XL (XLarge).

Model	Unlabeled data	Labeled data	Attack SNR	Attack success rate (word level)	
				targeted	untargeted
Wav2Vec2-Large	LV60k	LS960	30dB	88.0%	100%
			22dB	98.4%	100%
HuBERT-Large	LV60k	LS960	30dB	87.2%	100%
			22dB	98.5%	100%
Data2Vec-Base	LS960	LS960	30dB	63.4%	100%
			22dB	92%	100%
Wav2Vec2-Base	LS960	LS960	30dB	55.7%	100%
			22dB	62.9%	100%
Wav2Vec2-Base	LS960	LS100	30dB	53.9%	100%
			22dB	59.5%	100%
Wav2Vec2-Large	LS960	LS960	30dB	50.7%	100%
			22dB	49.4%	100%
Data2Vec-Large	LS960	LS960	30dB	66%	100%
			22dB	86.4%	100%
HuBERT-XLarge	LV60k	LS960	30dB	80.9%	100%
			22dB	95.5%	100%
UniSpeech-Sat-Base	LS960	LS100	30dB	50.4%	100%
			22dB	62.4%	100%
WavLM-Base	LV60k+VoxP+GS	LS100	30dB	21.7%	100%
			22dB	22.9%	100%
Wav2Vec2-Large	CV	CV+LS1	30dB	19.7%	100%
			22dB	36.1%	100%
M-CTC-Large	None	CV	30dB	7.5%	76.4%
			22dB	3.5%	83.4%
Speech2Text	None	LS960	30dB	7.3%	63.3%
			22dB	2.3%	74.6%
SB CRDNN	None	LS960	30dB	5.9%	86.39%
			22dB	1.5%	76.8%
SB Transformer	None	LS960	30dB	6.49%	90.56%
			22dB	1.2%	76.1%

Table 5. Results of the transferred adversarial attack on different ASR models, with multiple Signal-Noise Ratios. The first three models correspond to the proxies used to generate the adversarial examples. On all other models, the inputs have been transferred directly. We report for each model how much unlabeled data was used for SSL pretraining and for ASR finetuning. We also report its Word-Error-Rate on the LibriSpeech test-clean set, and the targeted and untargeted word-level attack success rate (see section 3.3)

Model	Unlabeled data	Labeled data	Clean WER		Attack success rate (word level)	
			w/o LM	with LM	w/o LM	with LM
Wav2Vec2-Large	LV60k	LS960	2.2%	2.0%	80.2%	88.0%
HuBERT-Large	LV60k	LS960	2.1%	1.9%	77.3%	87.2%
Data2Vec-Base	LS960	LS960	3.2%	2.5%	51.7%	63.4%
Wav2Vec2-Base	LS960	LS960	3.4%	2.6%	43.6%	55.7%
Wav2Vec2-Base	LS960	LS100	6.2%	3.4%	41.8%	53.9%
Wav2Vec2-Large	LS960	LS960	2.8%	2.3%	41.4%	50.7%
Data2Vec-Large	LS960	LS960	2.2%	1.9%	56.9%	66%
HuBERT-XLarge	LV60k	LS960	2.0%	1.8%	63.9%	80.9%
UniSpeech-Sat-Base	LS960	LS100	6.4%	3.5%	39.5%	50.4%

Table 6. Results of the transferred adversarial attack on different ASR models, with and without language models. We report for each model how much unlabeled data was used for SSL pretraining and for ASR finetuning. We also report its Word-Error-Rate on the LibriSpeech test-clean set, and the targeted word-level attack success rate (see section 3.3)

Model	Smoothing	Clean WER	Attack success rate	
			targeted	untargeted
Wav2vec2-Base	No	3.4%	55.7%	100%
Wav2vec2-Base	Yes	13.5%	0%	33.9%
Wav2vec2-Large	No	2.2%	50.7%	100%
Wav2vec2-Large	Yes	7.3%	0%	19.5%
Data2Vec-Large	No	2.2%	66%	100%
Data2Vec-Large	Yes	6.7%	0%	14.1%

Table 7. Results of the transferred adversarial attack (generated in section 3.1) on the Wav2Vec2-Base, Wav2Vec2-Large and Data2Vec-Large models. Each model was pretrained and fine-tuned on 960h of LibriSpeech training data. We report results on both the undefended version of each model and one defended with randomized smoothing at $\sigma = 0.02$. We report the WER of each model on the LibriSpeech test-clean set, and the word-level success rate of the attack (see Section 3.3).