Membership Inference Attacks Against Self-supervised Speech Models

Wei-Cheng Tseng, Wei-Tsung Kao, Hung-yi Lee

Graduate Institute of Communication Engineering, National Taiwan University, Taiwan
{r09942094, r09942067, hungyilee}@ntu.edu.tw

Abstract

Recently, adapting the idea of self-supervised learning (SSL) on continuous speech has started gaining attention. SSL models pre-trained on a huge amount of unlabeled audio can generate general-purpose representations that benefit a wide variety of speech processing tasks. Despite their ubiquitous deployment, however, the potential privacy risks of these models have not been well investigated. In this paper, we present the first privacy analysis on several SSL speech models using Membership Inference Attacks (MIA) under black-box access. The experiment results show that these pre-trained models are vulnerable to MIA and prone to membership information leakage with high adversarial advantage scores in both utterance-level and speaker-level. Furthermore, we also conduct several ablation studies to understand the factors that contribute to the success of MIA.

1 Introduction

As the applications of deep learning become more and more widespread, it is inevitable for people to pay extra attention to the privacy issues of deep learning models. (Shokri et al. 2017) first investigates the information leakage from the viewpoint of data and propose the shadow model strategy for Membership Inference Attacks (MIA), which aims to inference whether a given data was used to train the model. In the speech processing community, the privacy issues of some important applications such as ASR also have been explored (Shah et al. 2021).

Beyond the supervised learning models mentioned above, in recent years, self-supervised learning (SSL) pre-trained models have become an important component of natural language processing (NLP) and speech processing. SSL speech models can be pre-trained on large-scale unlabeled speech datasets with different manners such as discriminative loss (Hsu et al. 2021; Baevski et al. 2020; Oord, Li, and Vinyals 2018; Riviere et al. 2020), generative loss (Chung et al. 2019; Liu, Li, and Lee 2020; Liu et al. 2020) or multi-task (Pascual et al. 2019; Ravanelli et al. 2020). The SSL models can extract high-level, informative, and compact feature vectors from the raw audio inputs. The extracted features improve downstream tasks like speech recognition, speaker verification, speech translation, spoken language understanding, and so on (wen Yang et al. 2021). Only requiring unlabeled audios is a desirable property of SSL since large-scale unlabeled audios can be collected easily compared to labeled data. Furthermore, if the audios are collected from an online service, using only unlabeled data can protect the speaker's privacy because their speech will not be listened to by any annotators.

Nevertheless, the extreme size of the unlabeled corpus also makes it hard for the developers to ensure that there is not any private information in the corpus. It is still possible that a malicious person can attack the SSL models to retrieve some sensitive information in the pre-training data. In the NLP community, researchers have successfully eavesdropped on sensitive information such as phone numbers from SSL NLP models (Carlini et al. 2020). But for SSL speech models, to our best knowledge, there is still a lack of systematic analyses or successful attack strategies. It is still an open question that whether the SSL speech models would have similar behavior or not. As the self-supervised pre-trained model becomes more and more ubiquitous, the practitioners would consider deploying the pre-trained models to their products and services by using Machine Learning as a Service (MLaaS) engines. Consequently, it is imperative for us to either eliminate these concerns or verify the existence of the privacy leakage.

In this paper, we perform the first MIA against several SSL speech models under black-box access. The results show that SSL speech models are vulnerable to such attack at both speaker and utterance-level. Besides, we also conduct an ablation study to understand how the size of the model, the pre-trained dataset, and a simple data perturbation affect the attack performance.

2 Methodology

2.1 Threat Modeling

Here we consider an adversary who has black-box input-output access to the target SSL speech model $\mathcal{M}$ which has been pre-trained on dataset D_{target} that possibly contains sensitive information. The adversary can only inference $\mathcal{M}$ with some utterances and compute their output representations while having no knowledge about both the structure and the pre-training algorithm of $\mathcal{M}$. In addition, the adver-

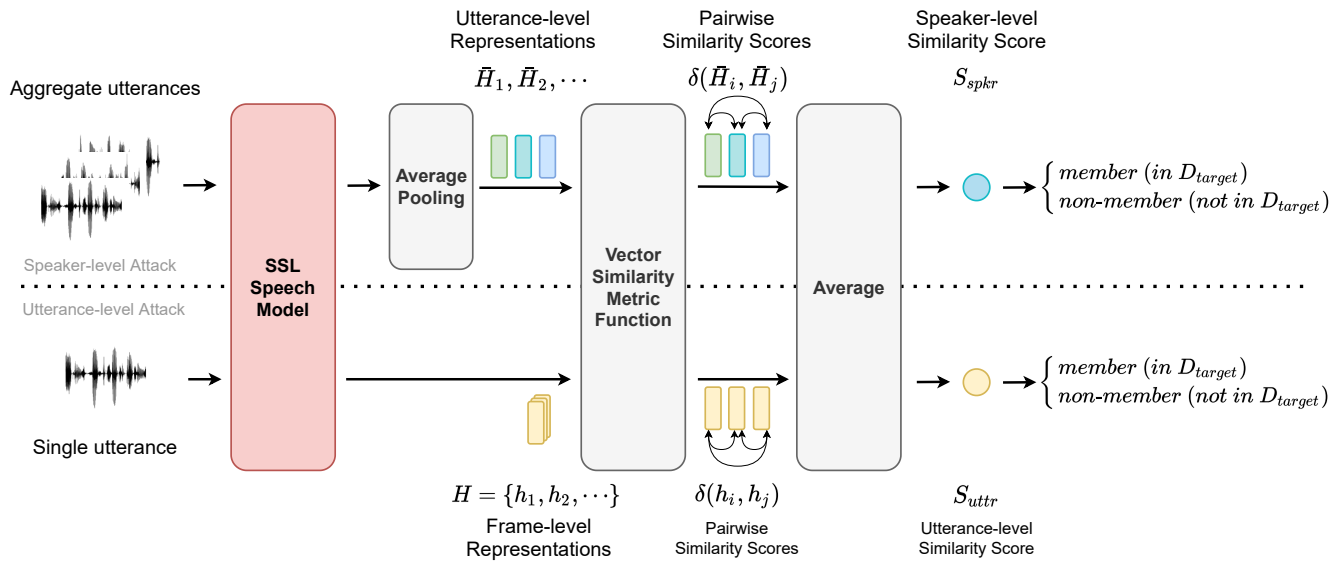

Figure 1: The proposed membership inference attack against self-supervised pre-trained speech models.

sary may utilize an auxiliary dataset D_{aux} to perform MIA. D_{aux} comprises aggregate utterances from several speakers that are not included in D_{target}.

Since SSL speech models are pre-trained on large-scale utterances from various speakers, the adversary could launch MIA in two levels, according to the speaker identities or the utterance information. We introduce the utterance-level attacks in section 2.2 and the speaker-level attack in section 2.3. For the attack strategy, we perform thresholding attacks based on similarity scores (Song and Raghunathan 2020). This is according to the observation that SSL speech models usually minimize or maximize certain similarity between models' representations and some acoustic features in each utterance. For example, Hubert (Hsu et al. 2021) maximizes the similarity between its representations and the centroids of the MFCC clusters; Wav2vec2 (Baevski et al. 2020) maximizes the similarity between its output representations and the quantized input representations of the transformer model. So we believe that the similarity score statistics of the data in D_{target} and D_{aux} could be very different and distinguishable. The detailed methods are introduced in the next subsections. Figure 1 illustrates our attacks.

2.2 Utterance-level Attack

In utterance-level attack, the adversary inputs a specific utterance to $\mathcal{M}$ and try to decide whether it belongs to the pre-training dataset or not. The attack falls into two stages, namely *basic attack* and *improved attack*. *Basic attack* does not require any additional parameters and attacker models. While in *improved attack*, we need to learn a neural network.

Basic attack: in basic attack, we start with an utterance with T sample points $\boldsymbol{x} = [x_1, x_2, \cdots, x_T]$. The adversary first encodes the utterance into a sequence of representations $\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_m]$, where m is the length of features of the utterance. Here $\boldsymbol{h}_i \in \mathbb{R}^d$ is referred to as frame-level representations, and d is their dimensionality.

Then the adversary calculates the **utterance-level similarity score** by averaging the similarity score between each frame-level representations:

$$S_{uttr} = \frac{2}{m(m-1)} \sum_{i=1}^{m} \sum_{j=i+1}^{m} \delta(\boldsymbol{h}_i, \boldsymbol{h}_j)$$

where $\delta : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ is a vector similarity metric function (e.g. cosine similarity).

Finally, the adversary uses S_{uttr} to decide membership: if S_{uttr} is above some threshold then $\boldsymbol{x}$ is a member of the training data; otherwise, it is not a member.

Improved attack: using a pre-defined similarity metric may not achieve the best membership inference performance. Here we adapt *pseudo-labeling* on a subset of D_{aux} to learn a customized similarity metric function. Given utterances $\{\boldsymbol{x}_i\}_{i=1}^N \in D_{aux}$, the adversary first compute S_{uttr} for each x_i. Then the adversary collects a dataset comprising of the frame-level representations of utterances of both k highest and lowest S_{uttr}, respectively labeled as *member* and *non-member*. This dataset is used to train a neural network $f : \mathbb{R}^{m \times d} \mapsto \mathbb{R}$ that computes the customized utterance-level similarity score from the frame-level representations. We use binary cross-entropy loss to train the network.

2.3 Speaker-level Attack

In speaker-level attack, the adversary inputs aggregate utterances from a certain speaker to $\mathcal{M}$ and aims to determine whether this speaker involves in D_{target} or not. Likewise, we divide the attack into two stages.

Basic attack: in basic attack, we start with the aggregation $\mathcal{X}_C$ of n utterances $\{\boldsymbol{x}_i^C\}_{i=1}^n$ from certain speaker C. The adversary first computes the utterance-level representations $\{\bar{\boldsymbol{H}}_i\}_{i=1}^n$ by taking the average of the frame-level representations of each utterance x_i^C, as we believe that

utterance-level representation may contain more speaker information. Then the adversary calculates the **speaker-level similarity score** by averaging the similarity score between the utterance-level representations:

$$S_{spkr} = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \delta(\bar{\boldsymbol{H}}_i, \bar{\boldsymbol{H}}_j)$$

If S_{spkr} is above some threshold, the adversary considers that C involves in the pre-training data, and vice versa.

Improved attack: for better performance on the speaker-level attack, we also utilize D_{aux} to learn a customized similarity metric function. Given a set of aggregate utterances $\{\mathcal{X}_i\}_{i=1}^{m}$ from m speakers in D_{aux}, the adversary first compute S_{spkr} for each $\mathcal{X}_i$. Then the adversary collects a dataset comprising of the utterance-level representations from the speakers with k highest and lowest S_{spkr}. The utterance-level representations from the speakers with top-k similarity scores are all labeled as *member*, and the ones from the speakers with k lowest similarity scores are all labeled as *non-member*. This dataset is used to train a neural network that acts as the customized δ. We also use binary cross-entropy loss to train the network.

3 Experimental Setting

In the experiments, we use four SSL speech models from S3PRL toolkits (Liu and Shu-wen 2020): HuBERT, wav2vec 2.0, CPC, and TERA[1]. These models were pre-trained on large-scale unlabeled data such as LibriSpeech (Panayotov et al. 2015) and Libri-Light (Kahn et al. 2020). Five subsets of the LibriSpeech corpus are involved in the experiments: *train-clean-100*, *dev-clean*, *dev-other*, *test-clean* and *test-other*, where *train-clean-100* serves as member for $\mathcal{M}$. The rest constitute D_{aux} and serve as non-members. We also conduct experiments in which VCTK-Corpus (Yamagishi, Veaux, and MacDonald 2019) servers as D_{aux} to amplify the difference between the members and non-members.

For utterance-level attack, since we find out that most utterances have relatively low similarity scores, we use *1 - cosine similarity* for δ. And for the learned similarity metric function f, it contains an attention pooling layer (Safari, India, and Hernando 2020) followed by two linear layers. We optimize f for 20 epochs with learning rate set to 10^{-5} and $k = 500$.

As for the speaker-level attack, unlike utterance-level, we find out that most speakers have similarity scores close to 1, so we use *cosine similarity* for δ. And for the learned similarity metric function f, it contains an attention pooling layer followed by a linear layer and a dot product layer. We optimize f for 20 epochs with learning rate set to 10^{-5} and $k = 1$.

[1]The variants we use here are HuBERT-base, wav2vec 2.0-base, modified CPC and TERA-960hr. Please refer to their original paper for a more detailed description.

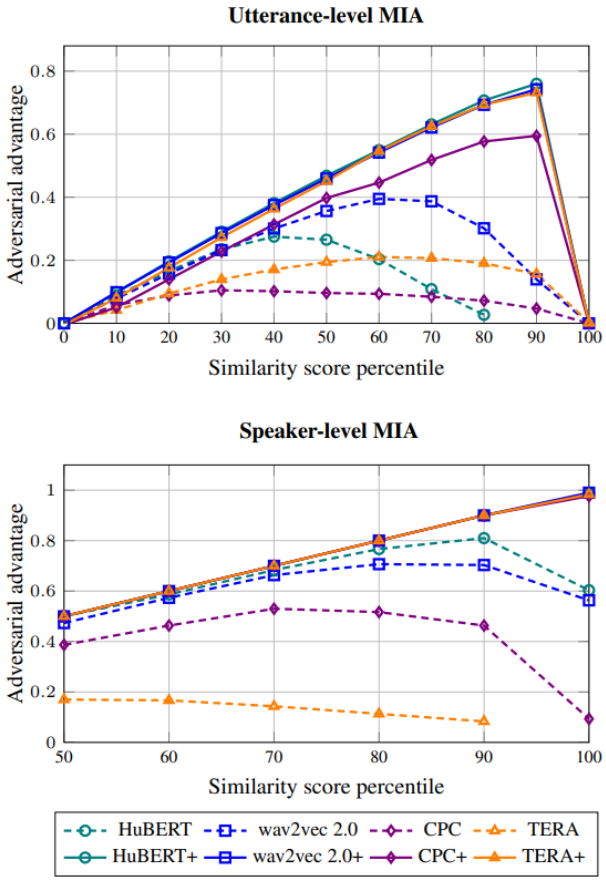

Figure 2: The performance of speaker-level and utterance-level MIA. The x-axis represents which percentile range of the similarity scores of the non-members is selected as the threshold. The dashed lines are the results of basic attack, and the solid lines along with the plus (+) denote the results of improved attack.

4 Results

4.1 Quantitative Results

We evaluate membership inference attacks with thresholds equal to each decile range as well as the highest and lowest value of the similarity scores of non-members. The performance is reported with adversarial advantage (Yeom et al. 2018), defined as the difference between true and false positive rate. A higher advantage score implies a better attack performance. Random guesses offer an advantage of 0.

Figure 2 shows the performances of the membership inference attacks against different SSL speech models in both utterance-level and speaker-level. At the utterance-level attack, we can observe that all pre-trained models are vulnerable to such attack to reveal membership information except CPC. Wav2vec 2.0 even has an advantage score of around 0.4 when we apply the basic attack. Additionally, using the learned similarity metric function in the improved attack significantly boosts the attack performance, with the highest advantage score for the pre-trained models being 0.7598 (Hu-

	HuBERT	wav2vec 2.0	TERA	CPC
D_{aux} comes from LibriSpeech				
utterance	0.7598	0.7423	0.7317	0.5948
speaker	0.9867	0.9900	0.9833	0.9767
D_{aux} comes from VCTK				
utterance	0.7692	0.7757	0.6276	0.7716
speaker	1	1	1	1

Table 1: The attack performance when the unseen data comes from VCTK corpus.

BERT), 0.7423 (wav2vec 2.0), 0.5948 (CPC), and 0.7317 (TERA), respectively.

Furthermore, at the speaker-level attack, the membership information leakage of most pre-trained models is more severe than the one in utterance-level except TERA. For the basic attack, the worst one, HuBERT, even has an advantage score of around 0.8. Moreover, we surprisingly find out that with only $k = 1$, the learned similarity metric function can further improve the attack to obtain a near-optimal advantage score, where the highest score is 0.9867 (HuBERT), 0.9900 (wav2vec 2.0), 0.9767 (CPC), 0.9833 (TERA). The results remain similar if D_{aux} comes from VCTK-Corpus, as shown in table 1. So the success of the proposed attacks does not result from any special selections of D_{aux}. These results indicate that SSL speech models are weak in preventing membership information leakage regardless of the self-supervised objectives. Especially, the state-of-the-art models, HuBERT and wav2vec 2.0, have the worst privacy-preserving ability, which validates the intuition – the utility-privacy trade-off still exists when people apply SSL.

4.2 Ablation Study

We then inspect how the number of parameters of the pre-trained models and the size of the pre-training dataset affects the attack performance by applying the attack against several variants of these models. For the number of parameters, we compare (1) HuBERT-{Large, X-Large}, which are pre-trained on LibriLight-60Khr (LL-60K), and (2) wav2vec 2.0-{Base, Large}, which are pre-trained on LibriSpeech-960hr (LS-960). As for the size of the pre-training dataset, we compare wav2vec 2.0 models that are pre-trained on LS-960 or LL-60K, and TERA models that are pre-trained on LibriSpeech-100hr (LS-100) or LS-960. All of these models are published by their original author to ensure the same pre-training policy. For simplicity, here we only report the highest advantage score against each model under improved attack.

Table 2 lists the attack performance when the model size differs. We can observe that using a larger model leads to a lower speaker-level attack performance, preventing itself from speaker-level membership information leakage. But when it comes to utterance-level attack, there's no such guarantee that which model size is better. Table 3 shows the attack performance when the pre-training dataset size changes. In utterance-level and speaker-level MIA, pre-

Model	Model size ablation		
	Base	Large	X-Large
Utterance-level MIA			
HuBERT	–	0.7002	**0.7069**
wav2vec 2.0	**0.7423**	0.7080	–
Speaker-level MIA			
HuBERT	–	**0.9433**	0.86
wav2vec 2.0	**0.9900**	0.9167	–

Table 2: The attack performance of HuBERT and wav2vec 2.0 when the model size varies.

Model	Dataset size ablation		
	LS-100	LS-960	LL-60K
Utterance-level MIA			
wav2vec 2.0	–	0.7080	**0.7134**
TERA	0.5772	**0.7317**	–
Speaker-level MIA			
wav2vec 2.0	–	0.9167	**0.9333**
TERA	0.9000	**0.9833**	–

Table 3: The attack performance of TERA and wav2vec 2.0-large when the dataset size varies.

training the model on smaller dataset results in a lower attack performance.

These results are surprising as most of the previous works demonstrated that using a larger model or utilizing more training data may reduce the hassle of privacy risks (Carlini et al. 2020; Melis et al. 2019; Shokri, Strobel, and Zick 2021). This strongly motivates the need for an in-depth study on the influence of these factors and developing privacy-preserving techniques for pre-training SSL speech models in the future.

5 Preliminary study of Defense

Several defenses have been proposed to mitigate the privacy risks of machine learning models. For example, (Song, Chaudhuri, and Sarwate 2013) incorporated noisy-SGD in the training to ensure the model for differential privacy. (Salem et al. 2019) consider the privacy leakage as a by-product of overfitting and adopt ensemble learning to reduce the harm of it. (He and Zhang 2021) alleviate the privacy leakage of contrastive models using adversarial training.

Beyond these methods, here we conduct a preliminary study that tries to defend the proposed attack on SSL speech models. We discuss the effect of a simple data perturbation, *waveform reversing*. Intuitively, a normal waveform and a reversed one sound very different for humans. One can possibly pre-train a SSL model on only the reversed waveforms to decrease the attack performance of attackers using the normal waveforms without sacrificing utility. We start with pre-training a TERA model on reversed LS-100

	TERA (LS-100)	TERA-reverse
utterance	0.5772	0.5577
speaker	0.9	0.9900

Table 4: The attack performance of TERA models pre-trained on normal LS-100 and the reversed one (TERA-reverse).

(called TERA-reverse). On several tasks such as phoneme classification and speaker verification, the performance of this model is close to the one pre-trained on normal LS-100. We then perform improved attack against it, as shown in table 4. We find that waveform reversing partly helps to alleviate utterance-level privacy leakage. The reason it fails to prevent speaker-level attack may be that the adversary utilizes the average of frame-level representations in speaker-level attack, which diminishes the effect of waveform reversing.

We further extend our experiment to other SSL models. Due to resource limitations, however, we could not fully pre-train these models with the reversed LS-960 or LL-60K from scratch. So we consider an approximated scenario and try to get some insights. In the scenario, the SSL models are pre-trained on normal waveforms while the attackers perform MIA with the reversed waveforms. We speculate that its effect is close to pre-training on reversed waveforms and attacking the models with normal waveforms. The results are shown in table 5. Similarly, waveform reversing slightly decreases the performance of the improved utterance-level attack but still fails to prevent improved speaker-level attack. Overall, the privacy issues of SSL speech models are critical. Only waveform reversing may not be enough to resolve them. Especially, more advanced techniques are required if we would like to keep the speaker information secure. We leave this issue and the combination of other defense methods and SSL speech models for the future work.

6 Conclusions

This paper performs the first membership inference attack against several self-supervised pre-trained speech models under black-box access. The results show that these models are vulnerable to both speaker and utterance-level attacks, with the highest adversarial advantage over 0.99 and 0.75, respectively. We also conduct an ablation study indicating that with smaller datasets, one can slightly reduce the risk of privacy leakage, which is different to the observation of previous works. The success of the proposed attacks suggests that the representations of SSL models encode the membership information of the pre-training data, which can cause severe privacy issues. We also conduct a preliminary study of defense but find that a simple data augmentation is not enough to prevent the proposed attack. This strongly gives rise to the need for caution and motivates a high demand for developing privacy-preserving pre-training techniques in the future.

	HuBERT	wav2vec 2.0	TERA	CPC
Attack with normal waveforms				
utterance	0.7598	0.7423	0.7317	0.5948
speaker	0.9867	0.9900	0.9833	0.9767
Attack with reversed waveforms				
utterance	0.7577	0.7440	**0.3317**	**0**
speaker	0.9900	0.9900	0.9900	0.9900

Table 5: The attack performance when the attacker applies MIA with the reversed waveforms.

References

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems*, 33.

Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U.; et al. 2020. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*.

Chung, Y.-A.; Hsu, W.-N.; Tang, H.; and Glass, J. 2019. An Unsupervised Autoregressive Model for Speech Representation Learning. *Proc. Interspeech 2019*, 146–150.

He, X.; and Zhang, Y. 2021. Quantifying and Mitigating Privacy Risks of Contrastive Learning. *arXiv preprint arXiv:2102.04140*.

Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *arXiv preprint arXiv:2106.07447*.

Kahn, J.; Rivière, M.; Zheng, W.; Kharitonov, E.; Xu, Q.; Mazaré, P.-E.; Karadayi, J.; Liptchinsky, V.; Collobert, R.; Fuegen, C.; et al. 2020. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7669–7673. IEEE.

Liu, A. T.; Li, S.-W.; and Lee, H.-y. 2020. Tera: Self-supervised learning of transformer encoder representation for speech. *arXiv preprint arXiv:2007.06028*.

Liu, A. T.; and Shu-wen, Y. 2020. S3PRL: The Self-Supervised Speech Pre-training and Representation Learning Toolkit.

Liu, A. T.; Yang, S.-w.; Chi, P.-H.; Hsu, P.-c.; and Lee, H.-y. 2020. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6419–6423. IEEE.

Melis, L.; Song, C.; De Cristofaro, E.; and Shmatikov, V. 2019. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, 691–706. IEEE.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210.

Pascual, S.; Ravanelli, M.; Serra, J.; Bonafonte, A.; and Bengio, Y. 2019. Learning problem-agnostic speech representations from multiple self-supervised tasks. *arXiv preprint arXiv:1904.03416*.

Ravanelli, M.; Zhong, J.; Pascual, S.; Swietojanski, P.; Monteiro, J.; Trmal, J.; and Bengio, Y. 2020. Multi-task self-supervised learning for robust speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6989–6993. IEEE.

Riviere, M.; Joulin, A.; Mazaré, P.-E.; and Dupoux, E. 2020. Unsupervised pretraining transfers well across languages. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7414–7418. IEEE.

Safari, P.; India, M.; and Hernando, J. 2020. Self-Attention Encoding and Pooling for Speaker Recognition. *Proc. Interspeech 2020*, 941–945.

Salem, A.; Zhang, Y.; Humbert, M.; Berrang, P.; Fritz, M.; and Backes, M. 2019. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. In *Network and Distributed Systems Security (NDSS) Symposium 2019*.

Shah, M. A.; Szurley, J.; Mueller, M.; Mouchtaris, A.; and Droppo, J. 2021. Evaluating the Vulnerability of End-to-End Automatic Speech Recognition Models To Membership Inference Attacks. *Proc. Interspeech 2021*, 891–895.

Shokri, R.; Strobel, M.; and Zick, Y. 2021. On the privacy risks of model explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 231–241.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, 3–18. IEEE.

Song, C.; and Raghunathan, A. 2020. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 377–390.

Song, S.; Chaudhuri, K.; and Sarwate, A. D. 2013. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, 245–248. IEEE.

wen Yang, S.; Chi, P.-H.; Chuang, Y.-S.; Lai, C.-I. J.; Lakhotia, K.; Lin, Y. Y.; Liu, A. T.; Shi, J.; Chang, X.; Lin, G.-T.; Huang, T.-H.; Tseng, W.-C.; tik Lee, K.; Liu, D.-R.; Huang, Z.; Dong, S.; Li, S.-W.; Watanabe, S.; Mohamed, A.; and yi Lee, H. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, 1194–1198.

Yamagishi, J.; Veaux, C.; and MacDonald, K. 2019. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92).

Yeom, S.; Giacomelli, I.; Fredrikson, M.; and Jha, S. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, 268–282. IEEE.