Interspeech 2024
1-5 September 2024, Kos, Greece

Quantifying Unintended Memorization in BEST-RQ ASR Encoders

Virat Shejwalkar, Om Thakkar, Arun Narayanan

Google

{vshejwalkar, omthkkr, arunnt}@google.com

Abstract

Self-supervised ASR encoders are increasingly being adopted in real-world applications as they enable downstream ASR tasks with impressive performances. This raises concerns around privacy of the data used to train such encoders, especially since neural networks are known to unintentionally memorize rare/unique samples from their training data. To this end, we perform the first systematic auditing of unintended memorization in ASR encoders. Specifically, we focus on a state-of-the-art Conformer-based ASR encoder pre-trained using the BEST-RQ technique, which forms the foundation of many real-world ASR applications. We propose a novel auditing method that can successfully demonstrate such memorization in ASR encoders, even for samples occurring just once in their training data. Finally, we show the promise of pre-training with per-sample gradient clipping towards mitigating such memorization in ASR encoders without significantly impacting downstream model quality.

1. Introduction

Recent advancements in self-supervised learning have shown impressive improvements in Automatic Speech Recognition (ASR). Furthermore, self-supervised learning enables model training when there is scarcity of supervised data. Due to these benefits, pre-training encoders using self-supervised learning and then fine-tuning them for various downstream tasks has become commonplace [1, 2]. The increasing adoption of encoders in various application domains raises concerns around the privacy of their training data, especially since machine learning models are notorious for their tendency to leak parts of their training data. A long line of research demonstrates how this can lead to private information extraction attacks [3, 4, 5] in practical settings. Motivated by such attacks, prior works have systematically demonstrated unintentional memorization in models from various domains, including language models [6, 7], vision models [5, 8], and even in end-to-end ASR models [9]. However, so far there has been no work on quantifying, or *auditing*, unintentional memorization in ASR encoders.

In this work, we focus on auditing memorization in state-of-the-art Conformer-based [10] ASR encoders pre-trained using the BEST-RQ algorithm [11], as they form the foundation of many real-world ASR applications [1]. We achieve this by using the foundational Secret Sharer framework [6]. At a high level, the Secret Sharer inserts carefully-crafted atypical samples, called *canaries*, in the training data of the target model, i.e., the model under audit. These canaries are intended to be difficult to learn from regular training data, as the canary distribution is designed to be different from distribution of the regular training data. Hence, if the trained model performs much better on the inserted canaries compared to un-inserted canaries from the same canary distribution, we conclude that the model exhibits memorization.

Our contributions: To the best of our knowledge, ours is the first work to systematically audit the unintentional memorization in state-of-the-art ASR encoders. A key aspect of the Secret Sharer is how it measures the performance of the target model on canaries. We propose a novel performance metric to improve such auditing in ASR encoders. Note that BEST-RQ trains ASR encoders to predict masked portions of an input training sample, where the masks are applied randomly to the sample's frames. The general practice for auditing is to use the same performance metric that the target model was optimized for during training, which in our case would be the random masks based loss/accuracy for BEST-RQ. However, we show that doing so can result in highly underestimating the propensity of an ASR encoder to unintentionally memorize samples from its training data. In this work, we propose a novel *word-level mask* based performance metric to improve Secret Sharer auditing. In other words, instead of masking random frames, we propose masking all frames corresponding to a word in a sample. We demonstrate via extensive experiments how using word-level masks can significantly improve auditing results.

Apart from the novel performance metric, we show that the design of canaries is also a key to successful auditing. To this end, we show that the canaries proposed in recent work on end-to-end ASR models auditing [9] may not be useful towards auditing memorization in state-of-the-art ASR encoders. We observe that this is due to the fundamentally different nature of the self-supervised BEST-RQ training. To this end, we propose three novel types of out-of-distribution (OOD) canaries.

Empirically, we first show that our OOD canaries can detect memorization in ASR encoders even with traditional BEST-RQ loss as the performance metric. Then, we demonstrate that our novel word-level masks based performance metric is successful at uncovering significantly more memorization in ASR encoders. Though we focus on auditing only BEST-RQ trained ASR encoders in this work, our techniques can also be useful towards auditing memorization in other BERT-style ASR encoders [12, 13, 14, 15] due to similarities in the pre-training techniques; we leave such exploration for future work.

Lastly, we investigate strategies to mitigate unintentional memorization in ASR encoders. Differentially private ML [16] can mitigate such memorization via strong theoretical guarantees, but often significantly degrades model utility for large models. Hence, to mitigate memorization without significantly compromising model quality, we evaluate *per-sample gradient clipping* during BEST-RQ training. We show that per-sample gradient clipping significantly reduces memorization in ASR encoders with minimal impact on their utility.

10.21437/Interspeech.2024-574

2. Background and Related Work

BEST-RQ [11]: **BE**RT-based **S**peech pre-**T**raining with **R**andom-projection **Q**uantizer (BEST-RQ) is a state-of-the-art self-supervised training algorithm to train ASR encoders. At a high level, BEST-RQ starts with constructing a quantizer that projects samples using a randomly initialized matrix, and assign a label to the projection using a randomly initialized codebook. During training, samples are randomly masked before feeding them to the encoder. The encoder is trained to predict labels for the masked region based on the unmasked regions around it. Both the quantizer and codebook are frozen throughout the training, which removes any limitation on the architectural design of the encoder. Due to its simplicity and high efficacy, BEST-RQ has been used to pre-train state-of-the-art ASR encoders [1].

Secret Sharer [6]: The Secret Sharer framework is a foundational framework that allows auditing (quantifying) unintended memorization of training data in various ML models, including language models [6, 7], even when they are fused with acoustic models for ASR [17], as well as in end-to-end ASR models [9]. To audit a model, it first creates specially-designed out-of-distribution samples called *canaries* and inserts some of the canaries, called *seen canaries*, in the data used to train the model. To measure memorization, Secret Sharer computes the model's performance (using some metric, e.g., loss, accuracy, perplexity, word error rate, etc.) on the seen canaries, and on the *unseen canaries* that were drawn from the same distribution but never seen during training. If the performance on seen canaries is much better than on the unseen canaries, the Secret Sharer concludes that the model has memorized the seen canaries. To quantify the memorization of a canary, it uses the *rank* of the model's performance on the canary among its performance on unseen canaries. This is formalized in the following definition of *exposure*, the metric used to measure such memorization.

Definition 1 (Exposure [6]) *Given a canary c, a model $\mathcal{M}$, and examples in a holdout set r_i, the exposure of c is*

$$\textbf{\textit{exposure}}_{\mathcal{M}}(c, \{r_i\}) = \log_2 |\{r_i\}| - \log_2 \textbf{\textit{rank}}_{\mathcal{M}}(c, \{r_i\}),$$

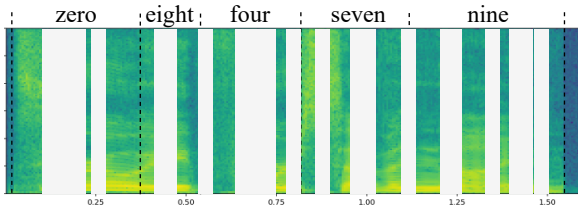
where $|\{r_i\}|$ is the size of the holdout set, and $\textbf{\textit{rank}}_{\mathcal{M}}(c, \{r_i\})$ is the rank of canary c among r_i in terms of a metric of interest, such as perplexity, character error rate, or loss.

Related Work: Numerous works have demonstrated leakage of sensitive information from ML models about their training data [3, 18, 5]. In the speech domain, there has been work showing that model updates during ASR training can leak potentially sensitive artifacts like labels [19] and speaker identity [20] of the utterances used in computing the updates. To analyze vulnerabilities in trained ASR models, Amid et al. [21] proposed a *noise-masking attack* to extract targeted training data from an ASR model. Consequently, auditing memorization in ASR models is becoming increasingly important. There is only one work [9] on efficiently auditing end-to-end ASR models using the Secret Sharer framework. Unfortunately, despite of their wide-spread use, no work has attempted to quantify memorization in ASR encoders; we aim to bridge this major gap through this work.

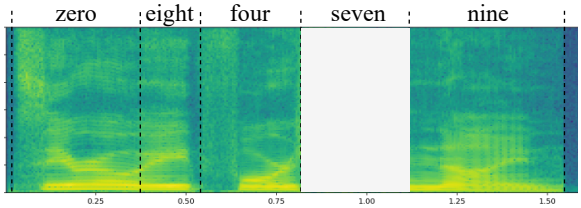
3. Auditing Memorization in ASR Encoders

To audit a model using the Secret Sharer, prior works [6, 22, 17] generally compute exposures using the same performance metric that the model was optimized for during training. For ASR encoders pre-trained using BEST-RQ, this metric is the label prediction accuracy/loss on fixed-length masks applied randomly to input audio samples. However, as we will see from the experiments in Section 4, how well does the exposure estimate the encoder's memorization of canaries heavily depends on the metric used to compute it. We propose a novel performance metric which can detect significantly more memorization than the naive metric based on BEST-RQ accuracy/loss.

Recall that BEST-RQ performs the following steps to compute loss on a given sample: 1) computes target labels for all frames using random-projection quantizer, 2) selects each frame of the sample with certain probability, 3) applies masks to a constant number of frames following the selected frame, 4) obtains the encoder's predicted labels for the masked portions of the sample, and 5) computes cross-entropy loss over the target and predicted labels. Our key observation is that the randomly applied masks of fixed-length in BEST-RQ may not be sufficient to cleanly distinguish memorization from *learning*. For instance, if a BEST-RQ mask lies within the boundaries of a word, e.g., the word "seven" in Figure 1-(a), in a rare sample, the model could provide the correct labels for the masked frames either due to unintentional memorization and/or due to *intra-word context*, i.e., model predicts correct labels using the context available from the unmasked portions of the word, e.g., "se–n" for the word "seven" in Figure 1-(a). In many situations, this can lead to underestimating the amount of unintended memorization in the encoder.



(a) Random mask of BESTRQ training



(b) Our proposed word-level mask for improved auditing

Figure 1: *(a) Random masks used in BEST-RQ training can provide strong intra-word context when predicting the masked portions. In the example above, "ve" of "seven" is masked, but access to "se–n" can make it easier for an ASR encoder to predict labels for the masked region with high confidence, regardless of the context around the word itself. (b) Our proposed word-level masks for improved auditing. Here, the goal is to provide no intra-word context to the encoder, to better capture its memorization given the context around the mask.*

Thus, to compute exposures better, we propose to use *word-level masks* instead of the random masks as in the traditional BEST-RQ training. Specifically, given a canary, we mask the frames corresponding to one word in its transcript, and compute cross-entropy loss over the masked frames as in BEST-RQ training. We compute losses for all the words and use their average as the measure of performance on the canary.

Figure 1 provides an illustration for our approach. From Figure 1-(a), we can see that the random masks used in BEST-RQ could provide strong intra-word context to the ASR encoder about the word(s) that they partially mask. Word-level masks (Figure 1-(b)) alleviate this issue as they mask a word in its entirety. This limits any intra-word context to the encoder. Hence, given sufficient *entropy* in the canary construction, the encoder can provide correct labels only by making use of some context around the masked word. Using our experiments in the next section, we show that word-level masks can be significantly better at computing canary exposure, more accurately estimating the propensity of such ASR encoders to unintentionally memorize rare/atypical samples in their training data.

4. Empirical analysis

In this section, we first describe our experimental setup. We discuss shortcomings of recently-proposed canaries [9] for auditing ASR encoders, and propose novel out-of-distribution (OOD) canaries. Next, we demonstrate the efficacy of our OOD canaries and word-level masks towards auditing memorization in ASR encoders. Finally, we analyze the effectiveness of *per-sample gradient clipping* towards mitigating such memorization in ASR encoders without significantly affecting their utility.

4.1. Experimental setup

For our model, we choose the 300M variant [9] of the state-of-the art ASR model architecture, Conformer XL [23]. We pre-train the encoder using BEST-RQ on the LibriLight dataset [24] for 1 million (1M) steps. We use the same hyperparameters from the original BEST-RQ work [11].

Details of our auditing setup: We use the Secret Sharer framework for auditing, and insert canaries with different number of repetitions $\in \{1, 2, 4, 8, 16\}$ in the LibriLight dataset used for pre-training the encoders. To account for randomness in memorization across different samples, we create 30 unique canaries for each repetition frequency. All our canary transcripts contain 10 words. We use a WaveNet Text-to-Speech (TTS) engine [25] to generate samples for each canary transcript.

4.1.1. The need for novel canaries to audit ASR encoders

Prior work [9] has proposed fast-paced canaries with transcripts containing potentially-repeating random English words for auditing end-to-end supervised ASR models. However, such *fast-paced canaries may not be useful towards auditing ASR encoders*. This is because, empirically, we observe that the exposures of such fast-paced canaries remain close to 1, which is the expected lower bound on exposures using the Secret Sharer. The reason for this is as follows. Recall that the BEST-RQ training is *unsupervised*, i.e., for a training sample, it uses a code-book to compute target labels for the sample and uses them to compute the cross-entropy loss. We observe that due to the fast-paced nature of canaries, the code-book maps all of their frames to a very small set of labels. We see that the trained encoder simply learns to arbitrarily predict from this set of labels when it sees fast-paced samples, resulting in similar losses on seen/unseen canaries. Due to this, fast-paced canaries can lend confidence to the belief that the encoder does not memorize rare training data, which motivates the need for designing novel canaries for better auditing.

4.1.2. Designing novel out-of-distribution (OOD) canaries

Given the shortcomings of auditing ASR encoders using fast-paced canaries, we propose three designs of increasingly OOD canaries that we use for auditing:
(C1) *Random English words canaries with English voices:* We compute the top 10,000 words from LibriSpeech test set, and for each canary, pick 10 random words with replacement to generate a transcript. We use normal-paced English voices to generate canaries. Note that, C1 canaries are similar to the fast-paced canaries in Section 4.1.1, but C1 canaries use voices at $1\times$ speed while fast-paced canaries use voices at $4\times$ speed.
(C2) *Non-repeating English digits canaries with English voices:* Transcripts of these canaries contain only digits sampled *without* replacement.
(C3) *Random Non-English (Afrikaans) words canaries with non-English voices:* The idea here is to use transcripts with words which model is never expected to encounter in its training data. We choose to generate transcripts in Afrikaans language; we use top 10,000 words vocab from Afrikaans text from CC-100 corpora [26], and pick 10 random words with replacement to generate a canary transcript. We generate these canaries using normal-paced non-English voices.

4.2. Auditing using BEST-RQ training objective

Now, we provide results for auditing using the random-masks based BEST-RQ loss as the performance metric to compute exposure. Figure 2-(a) shows the exposures for the three types of canaries when each canary occurs just once in the training data. We make several important observations from this plot: 1) *We see that exposure increases for all canary types as training progresses.* This indicates higher susceptibility for memorization with increasing training steps. 2) As evident from the high exposure values, we see *ASR encoders exhibit high unintended memorization for well-crafted canary types, even when they occur only once* in the training data of 2.8 million utterances. 3) *Random Afrikaans words based canaries (C3) are the most effective at uncovering unintentional memorization in ASR encoders*. At $1M^{th}$ training step, we see that the encoder may have almost completely memorized them as the exposure is close to its upper bound. We also notice that the exposure is the least for random English words canaries (C1). We conjecture that this may be because the Afrikaans canaries are more OOD compared to English canaries with respect to LibriLight data that contains only English samples.

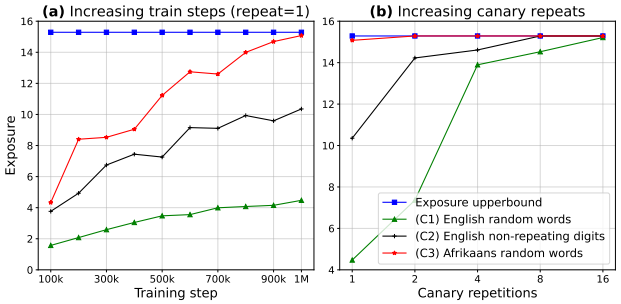


Figure 2: *(a) Exposure of **C1**-**C3** canaries (Section 4.1.2) at different training steps of BEST-RQ. (b) Exposures with varying number of canary repetitions in training data at $1M^{th}$ step. Exposure computation uses random-masks based BEST-RQ loss.*

Next, Figure 2-(b) shows exposures of the three OOD

canary types with increasing number of repetitions in training data. Along the observations in prior works [17, 9], we see that ***with increasing canary repetitions in training data, the model's memorization increases***. ASR encoders completely memorize Afrikaans canaries with just two repetitions and English non-repeating digits with 8 repetitions. These results demonstrate that ***ASR encoders pre-trained using BEST-RQ can unintentionally memorize their training data***. Our experiments with various OOD canaries highlight that ***designing effective canaries can be a key to detecting unintended memorization in ML models***.

4.3. Auditing using loss based on word-level masks

In this section, we demonstrate efficacy of using our novel word-level masks based loss in auditing ASR encoders. For clarity, we use the two most effective canaries (C2, C3) from the experiments in Section 4.2. Figure 3 compares exposures of canaries computed using random masks and word-level masks. Figure 3-(a) shows exposures of the canaries as training progresses; here canaries occur only once in the training data. We can consistently see a significant improvement in exposures when auditing using word-level masks. This implies that ***word-level masks based loss can estimate the unintentional memorization in ASR encoders much better than the traditional random masks based BEST-RQ loss***. Using word-level masks and Afrikaans canaries, the encoder memorization becomes evident as soon as 100k$^{\text{th}}$ training step, and it reaches its upper bound as soon as the 200k$^{\text{th}}$ step.

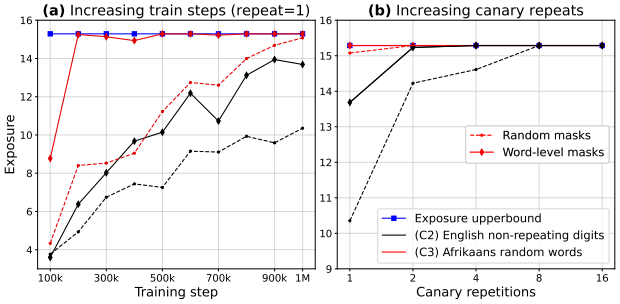


Figure 3: *Comparison of random and word-level masks: Word-level masks are much more effective at auditing BEST-RQ trained ASR encoders. (a) Exposures of singly-occurring canaries as training progresses. (b) Exposures at 1Mth training step for varying number of repetitions in training data.*

Figure 3-(b) shows exposures of C2 and C3 canaries with increasing repetitions in training data; here, we show exposures at 1Mth step. We note that the encoder completely memorizes Afrikaans canaries even if they occur just once in training data. Even for English non-repeating digits canaries, auditing using word-level masks detects their complete memorization for as low as two repetitions in training data, compared to 8 using the random-masks based auditing. These results show that ***our word-level masking based auditing approach can uncover the unintentional memorization of even the rarest of data, and early on during training***.

4.4. Towards mitigating memorization in ASR encoders

Differentially private (DP) ML, e.g., DP-SGD training [16, 27], can provide strong theoretical guarantees towards mitigating unintended memorization [28]. However, training using DP-SGD often significantly affects the utility of large models [29, 16, 30]. Hence, literature has proposed other solutions that can empirically mitigate memorization without significantly degrading model utility [31, 32, 17, 9]. In particular, *per-sample gradient clipping*, i.e., bounding the L2-norm of the gradients of each sample in a mini-batch before averaging them to update an ML model during training, has been shown to be an effective solution towards reducing such memorization in ASR models [17, 9]. This is because per-sample gradient clipping prevents any training sample from having an out-sized impact on the training process, and hence, on the final trained model.

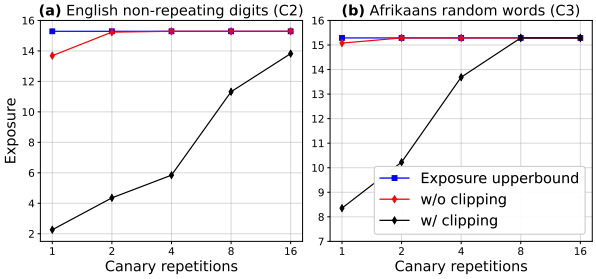


Figure 4: *Per-sample gradient clipping reduces memorization in BEST-RQ trained ASR encoders to some extent; exposures values are for the 1Mth training step.*

Results: Figure 4 shows exposures of the English non-repeating digits (C2) and Afrikaans random words canaries (C3) when we use per-sample clipping during BEST-RQ training; we use word-level masks to compute exposures. We note that, ***per-sample gradient clipping can significantly reduce unintentional memorization in ASR encoders***. In line with prior works [17, 9], we observe that clipping is effective at reducing memorization for low canary repetitions, though its efficacy reduces with increasing repetitions. One way to measure quality of pre-trained ASR encoders is to evaluate performance on a downstream task. Thus we evaluate encoder performance using LibriSpeech task. Specifically, we pre-train two ASR encoders using BEST-RQ with and without clipping, attach a decoder to each of the encoders to obtain two end-to-end ASR models, fine-tune the two complete models on LibriSpeech dataset [33] for 40k steps, and compute their word-error rates (WERs) on LibriSpeech test-other data. WERs of the models fine-tuned from encoders trained with and without clipping are 4.4 and 4.3, respectively. In other words, we observe that ***per-sample gradient clipping has minimal impact on the utility of the ASR encoders***.

5. Conclusions

We perform the first systematic auditing of unintended memorization in state-of-the-art ASR encoders pre-trained using self-supervised BEST-RQ training. We use the Secret Sharer framework for auditing and show that its effectiveness heavily relies on performance metric and canaries it uses for auditing. We design a novel performance metric based on word-level masks. We demonstrate the shortcomings of canaries from prior work in auditing ASR encoders and design novel out-of-distribution canaries. We demonstrate that our proposed performance metric and canaries are highly effective at auditing memorization in ASR encoders. Finally, in line with prior works, we observe that per-sample gradient clipping during BEST-RQ training can significantly reduce such memorization in ASR encoders.

6. References

[1] Y. Zhang, W. Han, J. Qin, Y. Wang, A. Bapna, Z. Chen, N. Chen, B. Li, V. Axelrod, G. Wang *et al.*, "Google usm: Scaling automatic speech recognition beyond 100 languages," *arXiv preprint arXiv:2303.01037*, 2023.

[2] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elsahar, H. Gong, K. Heffernan, J. Hoffman *et al.*, "Seamlessm4t-massively multilingual & multimodal machine translation," *arXiv preprint arXiv:2308.11596*, 2023.

[3] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Security and Privacy (SP), 2017 IEEE Symposium on*, 2017.

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

[5] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramer, B. Balle, D. Ippolito, and E. Wallace, "Extracting training data from diffusion models," *arXiv preprint arXiv:2301.13188*, 2023.

[6] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, "The secret sharer: Evaluating and testing unintended memorization in neural networks," in *28th USENIX Security Symposium (USENIX Security 19)*. Santa Clara, CA: USENIX Association, Aug. 2019, pp. 267–284. [Online]. Available: https://www.usenix.org/conference/usenixsecurity19/presentation/carlini

[7] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang, "Quantifying memorization across neural language models," *arXiv preprint arXiv:2202.07646*, 2022.

[8] V. Feldman, "Does learning require memorization? a short tale about a long tail," in *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, 2020, pp. 954–959.

[9] L. Wang, O. Thakkar, and R. Mathews, "Unintended memorization in large asr models, and how to mitigate it," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024.

[10] 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.

[11] C.-C. Chiu, J. Qin, Y. Zhang, J. Yu, and Y. Wu, "Self-supervised learning with random-projection quantizer for speech recognition," in *International Conference on Machine Learning*. PMLR, 2022, pp. 3915–3924.

[12] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[13] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[14] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[15] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 244–250.

[16] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016.

[17] W. R. Huang, S. Chien, O. Thakkar, and R. Mathews, "Detecting unintended memorization in language-model-fused asr," *Interspeech*, 2022.

[18] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. B. Brown, D. Song, U. Erlingsson *et al.*, "Extracting training data from large language models." in *USENIX Security Symposium*, vol. 6, 2021.

[19] T. Dang, O. Thakkar, S. Ramaswamy, R. Mathews, P. Chin, and F. Beaufays, "Revealing and protecting labels in distributed training," *Advances in Neural Information Processing Systems*, vol. 34, pp. 1727–1738, 2021.

[20] ——, "A method to reveal speaker identity in distributed asr training, and how to counter it," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4338–4342.

[21] E. Amid, O. Thakkar, A. Narayanan, R. Mathews, and F. Beaufays, "Extracting targeted training data from asr models, and how to mitigate it," *Interspeech*, 2022.

[22] S. Ramaswamy, O. Thakkar, R. Mathews, G. Andrew, H. B. McMahan, and F. Beaufays, "Training production language models without memorizing user data," *arXiv preprint arXiv:2009.10031*, 2020.

[23] Y. Zhang, D. S. Park, W. Han, J. Qin, A. Gulati, J. Shor, A. Jansen, Y. Xu, Y. Huang, S. Wang *et al.*, "Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[24] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen *et al.*, "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)*, 2020.

[25] A. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," in *International conference on machine learning*. PMLR, 2018, pp. 3918–3926.

[26] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 8440–8451. [Online]. Available: https://aclanthology.org/2020.acl-main.747

[27] V. Shejwalkar, A. Ganesh, R. Mathews, O. Thakkar, and A. Thakurta, "Recycling scraps: Improving private learning by leveraging intermediate checkpoints," *arXiv preprint arXiv:2210.01864*, 2022.

[28] M. Jagielski, "A note on interpreting canary exposure," 2023.

[29] R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization: Efficient algorithms and tight error bounds," in *2014 IEEE 55th annual symposium on foundations of computer science*. IEEE, 2014, pp. 464–473.

[30] M. H. Brendan, D. Ramage, K. Talwar, and L. Zhang, "Learning differentially private recurrent language models," *International Conference on Learning and Representation*, 2018.

[31] V. Shejwalkar and A. Houmansadr, "Membership privacy for machine learning models through knowledge transfer," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[32] X. Tang, S. Mahloujifar, L. Song, V. Shejwalkar, M. Nasr, A. Houmansadr, and P. Mittal, "Mitigating membership inference attacks by self-distillation through a novel ensemble architecture," in *31th {USENIX} Security Symposium ({USENIX} Security 22)*, 2022.

[33] 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.