Improving on-device speaker verification using federated learning with privacy

Filip Granqvist, Matt Seigel, Rogier van Dalen, Áine Cahill, Stephen Shum, Matthias Paulik

Apple

{fgranqvist,mseigel,rogier_vandalen,aine_cahill,stephen_shum,mpaulik}@apple.com

Abstract

Information on speaker characteristics can be useful as side information in improving speaker recognition accuracy. However, such information is often private. This paper investigates how privacy-preserving learning can improve a speaker verification system, by enabling the use of privacy-sensitive speaker data to train an auxiliary classification model that predicts vocal characteristics of speakers. In particular, this paper explores the utility achieved by approaches which combine different federated learning and differential privacy mechanisms. These approaches make it possible to train a central model while protecting user privacy, with users' data remaining on their devices. Furthermore, they make learning on a large population of speakers possible, ensuring good coverage of speaker characteristics when training a model. The auxiliary model described here uses features extracted from phrases which trigger a speaker verification system. From these features, the model predicts speaker characteristic labels considered useful as side information. The knowledge of the auxiliary model is distilled into a speaker verification system using multi-task learning, with the side information labels predicted by this auxiliary model being the additional task. This approach results in a 6 % relative improvement in equal error rate over a baseline system.

Index Terms: Speaker Verification, Multi-task Learning, Federated Learning, Differential Privacy

1. Introduction

Speaker verification is the problem of determining whether the person speaking is a specific individual or someone else. It is a vital feature for devices that use a "wake-up phrase" to provide access to information, as actions should only be triggered when this phrase is uttered by the device owner and not an impostor. Speaker verification systems usually consist of two components: a speaker embedding network; and a discriminative method for comparing pairs of embeddings to determine whether or not those embeddings originate from the same speaker [1, 2, 3, 4]. Additional side information can be useful for speaker verification [5, 6, 7]. This side information could be obtained through manual labelling. The setting that this paper considers instead is one where side information is available on many users' devices, but it is privacy-sensitive and should therefore not be uploaded to a central server. At a high level, this paper tests three hypotheses:

1. It is possible to train a classifier on the audio of trigger phrases to predict personal attributes of the speaker considered to be useful as side information.

2. Such a classifier can be improved with federated learning while preserving users' privacy.

3. The predictions of this classifier can be used to improve the performance of speaker verification.

Previous work has shown that neural networks can learn to predict speaker-dependent labels, such as gender [8, 9, 10], and emotion [11, 12, 8], from utterances. The desired outcome from testing the first hypothesis is a classifier that can predict similar speaker-dependent labels from the same input as the baseline speaker verification system used in this paper.

The second hypothesis is that it is possible to train a useful classifier on distributed user data while preserving user privacy. This is achieved through the combination of federated learning with differential privacy, which has been proposed and put into practice successfully in a large body of prior work [13, 14, 15, 16]. In federated learning, a batch of clients compute statistics on their local data using the latest version of a central model. The resulting statistics are combined on a server to improve the central model. This process is repeated with a different subset of users. Federated averaging [14] is commonly used for federated learning. In this algorithm, models are trained locally on devices and the changes in model parameter values are averaged on a central server and used to update the central model. However, local model updates, which are derived from the data, might leak sensitive information. To prevent this, differential privacy (DP) [17] is used in this paper. Prior work has provided few examples of high-utility applications on real-world models and datasets, and none on classifying speakers. This paper presents an analysis of different privacy regimes on training accuracy and convergence in this domain.

The third hypothesis states that the encoded knowledge of an auxiliary model trained on side information can be used to improve a speaker verification system. Manually labelled side information has been shown effective for improving speaker verification systems [5, 6, 7]. The baseline system [18] employs the common approach of using a speaker embedding network and scoring pairs of embeddings using cosine similarity. This paper shows that it can be improved by enriching the speaker embedding network with knowledge distilled from the auxiliary model.

The structure of this paper is as follows. Section 2 provides a high-level overview of how federated learning can be made private using differential privacy. Section 3 introduces the baseline speaker verification system used in production. Section 4 introduces the classifier trained on user data in a privacy-preserving manner to predict side information from trigger phrases, named the "vocal classification model". Section 5 explains the approach for including the vocal classification model in the speaker verification training setup to ultimately improve performance. Section 6 presents experimental results.

2. Federated learning with privacy

Data that can be used to improve machine-learned models often belongs to individuals or users and is therefore distributed over their devices. Federated learning is an approach that makes

learning in this scenario possible. In federated learning, a central model is trained over a distributed dataset, where a large number of nodes (e.g. user devices) hold variable-sized subsets of the data. A model update or gradient [19, 20, 14] is computed at the node on the local data, and communicated to a central server. A large number of these updates or gradients are combined at the central server during each iteration of training. A global update to the central model is computed as the average of local updates. This is called "federated averaging" [14].

Many organisations and policy makers are committed to upholding user privacy. This makes federated learning an important approach to consider when dealing with data that is private, as it goes some way to protecting privacy. However, even though raw user data is not communicated with the server, it has been shown that model updates can leak information about the raw data [21, 22]. As mentioned in Section 1, there is a large body of work investigating and putting approaches into practice which combine federated learning with some privacy protection. One common way to mitigate these threats to privacy is to apply differential privacy (DP) [17, 13, 15, 16]. Differential privacy makes it possible to add noise to the model updates to give a guaranteed upper bound on the amount of information that can be leaked. DP can be used to protect an individual's update by applying noise at the distributed node. DP can also be applied centrally to protect the privacy of individuals' updates after aggregation [23].

In this paper, a number of privacy regimes are explored in simulation. One of these regimes is to use a weaker form of local DP [24], combined with central DP. This form of local DP is applied to individual updates which are sent for aggregation on a secure server. This algorithm is an optimal method for providing updates (i.e. high-dimensional vectors) with the highest possible signal-to-noise ratio (SNR). The algorithm is tuned to achieve an SNR that permits high accuracy, while still allowing strong DP guarantees in deployment scenarios where it is applicable, e.g., because of shuffling [25] and subsampling [26]. In addition, while doing federated averaging, the server adds enough additional noise to ensure strong central DP guarantees (as per the moment's accountant [13].

3. Speaker verification

The baseline speaker verification system improved in this work is the system described in [18], which builds on an underlying voice trigger model that recognizes the trigger phrase. The input of the speaker verification system is a fixed-length supervector with features generated from forward propagation of the trigger phrase audio through the voice trigger model. The voice trigger system uses 26 Mel Frequency Cepstral Coefficients to parameterize a hidden Markov model (HMM) which models the trigger phrase. The means of the 31 HMM states (other than those modelling silence) are concatenated to form the supervector, resulting in $26 \times 20 = 520$ dimensions.

The supervector consists of features about the particular trigger phrase and the baseline speaker verification system contains a neural network that transforms the supervector into "speaker space", only focusing on retaining characteristics about the speaker itself. The neural network is a fully connected neural network with five layers. The first four layers consist of 256 dimensions, with batch normalization [27] and sigmoid activations. The fifth layer is designed to be the embedding layer, and is therefore only a linear transform of dimension 100 with batch normalization. In the training phase, a sixth layer of size K with softmax activation is added as a head to the architec-

ture, where K is the number of speakers in the training dataset. Training is defined as a speaker identification task, where the labels are one-hot representations of the K unique speakers and training is performed by minimizing the cross-entropy with the output softmax distribution. The trained embedding is used for measuring the similarity between two utterances by measuring the distance in "speaker space".

The speaker verification system stores multiple speaker vectors generated from a set of enrolment utterances. At test time, the acoustic instance of a trigger phrase is transformed into a fixed-length supervector from the HMM states of the voice trigger model, transformed again into a speaker vector by the speaker embedding network, and compared to the enrolment speaker vectors using cosine similarity. The average cosine similarity between the test speaker vector and the enrolment vectors can be interpreted as a speaker verification score. Following the notation of [28], the score is defined as

$$SV_{score} = \frac{1}{N} \sum_{i=1}^{N} \frac{f_{nn}(u_a)^{\top} f_{nn}(u_i^{spk})}{\|f_{nn}(u_a)\| \|f_{nn}(u_i^{spk})\|} \tag{1}$$

where u_i^{spk} is the ith out of N supervectors from the enrolment phase, u_a is the test supervector, and f_{nn} is the function expressed by the speaker embedding network. The speaker verification score SV_{score} is then compared to the operating threshold λ to reject or accept the request following the trigger phrase.

The work presented in this paper mainly targets the Japanese language (ja_JP), where the training dataset originates from a speaker population of size $K = 18700$. The speaker embedding network is trained with a batch size of 256, a weight decay of $5 \cdot 10^{-4}$, initial learning rate of 10^{-4} and momentum of factor 0.9. The performance of this setup is presented as *Baseline* in Section 6. Improvements on the baseline model presented in the paper are also trained with the same hyperparameters unless otherwise stated.

4. Vocal classification

The first hypothesis proposes that classification of side information can be learned from the voice trigger phrase only, if the side information is correlated with vocal characteristics. To test the hypothesis, a fully connected deep neural network was defined which uses the same input features as the baseline speaker verification system, i.e. the 520-dimensional supervector, and predicts the side information of the speaker. This model is referred to as the "vocal classification model". The target labels are sensitive information and are stored privately on devices, hence larger-scale experiments of training the vocal classification model were only possible using the framework explained in Section 2.

Experiments were carried out with limited central data to evaluate the effect of applying differential privacy on federated training of the vocal classification model. In addition to accuracy, the signal-to-noise ratio is measured throughout experiments to quantify the signal quality of model updates where DP has been applied. Here, SNR is computed as the ratio of the L2 norm of the un-noised model update to that of the added DP noise, i.e. $\text{SNR} = \frac{\|\text{unnoised update}\|_2}{\|\text{DP noise}\|_2}$. Figure 1 shows the accuracy of the following experiments in simulation on an evaluation set:

No DP No DP mechanism is applied. The resulting accuracy of 95.6% acts as an upper bound for the remaining experiments because introducing any DP mechanism to improve privacy guarantees is expected to have a negative

impact on accuracy. Note that even in the "No DP" scenario, there is some privacy protection, as anonymity is assumed. The accuracy of this model proves the first hypothesis, namely that a predictor of vocal characteristics can be learned from only the voice trigger phrase.

Local DP The strongest form of differential privacy, local DP, is applied using the Gaussian mechanism [29]. The privacy parameters used were $\epsilon = 2$ and $\delta = 10^{-5}$. This results in a significant negative impact on the performance of the model, yielding an accuracy of 82.1%, with a low SNR of 0.19 observed for the first central model update. The result is still much better than random, which shows that useful knowledge can be learned even with such strict privacy guarantees and low SNR.

Central DP The Gaussian moments accountant is applied on the aggregate model update to provide central privacy guarantees. The privacy parameters used were $\epsilon = 2$ and $\delta = 10^{-5}$. For the moments accountant, the population size is assumed to be 100M, the cohort size is 300 and the maximum number of central iterations is 60. The accuracy of the trained vocal classification model is 94.6%, which is close to the "No DP" case. An SNR of 1.31 is observed for the first central model update, which is significantly higher than "Local DP". However, this does not have any local privacy guarantees.

Central DP with weaker local DP Falling between the "local DP" and "Central DP" experiments, a weaker form of local DP [24] is used in combination with the Gaussian moments accountant for the central DP mechanism. This combination results in an accuracy of 94.1%, which is a good privacy-utility trade-off. The privacy parameters used in the weaker form of local DP translate to $\epsilon = 25.7$, which is in the high-epsilon regime. However, with the assumed privacy amplification through shuffling and sampling for anonymity, and the application of central DP with $\epsilon = 2$ and $\delta = 10^{-5}$, this may be considered a reasonable operating point. The SNR observed here is 1.07, which is significantly higher than "local DP" and expected considering the high ϵ value.
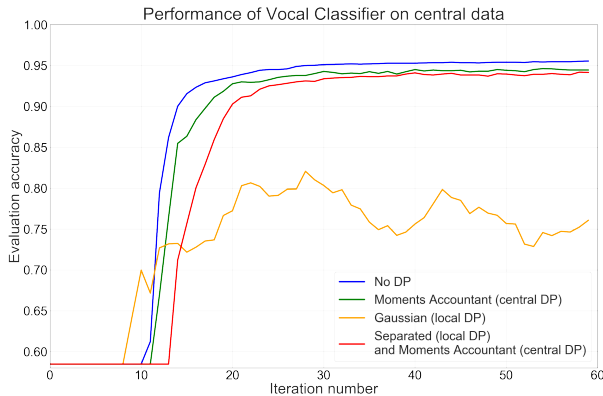

Figure 1: *Accuracy of the vocal classification model on an evaluation dataset, trained with different DP mechanisms.*

Hyperparameter tuning was performed in simulations with limited central data. The best performing model was used as an initialization when training with real devices. Switching to training the vocal classification model distributed on-device yielded multiple benefits. Firstly, additional categories

of side information not previously available in the limited central dataset were used. Secondly, magnitudes more data is available distributed on devices. The cohort of users for each central model update was increased to 5000, resulting in higher SNR for the same local DP parameters and smaller noise variance from the moments accountant used centrally. This larger effective corpus means that there is more speaker coverage. Thirdly, while labels of the central dataset may be erroneous due to errors in manual human annotation, on-device training uses ground truth labels, thereby increasing accuracy, while protecting privacy.

5. Multi-task learning of the speaker verification system

The third hypothesis proposed in this paper is that the encoded knowledge of a vocal classification model can act as complementary information for training a more accurate speaker verification system. Multiple approaches for utilizing the knowledge of the vocal classification model were experimented with: static rules for rejecting a request based on the model output, using the output of the model as input to intermediate layers when training the speaker embedding network, and multi-task learning with pseudo-labels. The latter approach was the most successful and is the focus of the rest of the paper.

Specifically, the network was trained to predict the speaker, like the baseline system, and, additionally, the side information. The loss was the sum of the original loss and a new term. Minimizing this loss distilled the knowledge [30] of the vocal classification model, encoded in the pseudo-labels it generates. Just as with the baseline system, the final classification layer was removed during inference to expose the embedding layer. Not only did distilling the knowledge of the vocal classification model outperform the two other approaches mentioned, but the final architecture of the speaker verification system remains unchanged. Since the vocal classification model is a model trained with differential privacy, any knowledge that is distilled by the speaker embedding network is also protected by the post-processing theorem of differential privacy [17].

The vocal classification model is generally confident in its predictions, mostly generating an output probability of 0.995 for the highest predicted class, even where it is incorrect. To better distill the knowledge for cases like this, the concept of temperature is used [30]. A temperature higher than 1 softens the output distribution, making the probabilities that were previously minuscule more representative.

Given a mini-batch of data $\mathbf{X} = \{\mathbf{x}^{(t)}\}$ and corresponding labels $\mathbf{Y} = \{\mathbf{y}^{(t)}\}$, the objective function to minimize for the multi-task setup is

$$\mathcal{L}_{mtl}(\mathbf{X}, \mathbf{Y}) = \sum_t \left(\mathcal{L}_{spk}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) + \gamma \mathcal{L}_{vc}(\mathbf{x}^{(t)}) \right), \quad (2)$$

where $\mathcal{L}_{spk}$ is the cross-entropy loss function for speaker identification as in the baseline setup, γ is a weight for the vocal classification loss relative to speaker identification, and the vocal classification loss $\mathcal{L}_{vc}$ is defined as

$$\mathcal{L}_{vc}(\mathbf{x}) = \frac{T^2}{N} \sum_{i=0}^N \left(\frac{e^{\mathcal{V}_i(\mathbf{x})/T}}{\sum_j e^{\mathcal{V}_j(\mathbf{x})/T}} - \frac{e^{z_i/T}}{\sum_j e^{z_j/T}} \right)^2. \quad (3)$$

Here, z_i is the ith logit of the predicted side information from the vocal classification layer in the speaker embedding training setup, $\mathcal{V}_i(\mathbf{x})$ is the ith logit from forward propagating the

Table 1: *Performance of the two tasks in the multi-task learning setup on an evaluation set.*

Model	Speaker accuracy (%)	Side information accuracy (%)
Baseline	82.56%	-
VC offline	83.37%	98.22%
VC FL	83.59%	98.15%

Table 2: *Performance of speaker verification on a test set.*

Model	EER (%)
Baseline	10.10
VC offline	9.95
VC FL	9.50

supervector $\mathbf{x}$ through the vocal classification model and T is the temperature. The above can be interpreted as the mean squared error between the "softened" softmax distributions. The mean squared error is multiplied by T^2 because otherwise $\frac{\partial \mathcal{L}_{vc}}{\partial z_i}$ scales with a factor of $\frac{1}{T^2}$ for large T (see section 2.1 of [30]).

6. Results

In this section, results of three models are presented, all with the same final network architecture. The difference is only in how the speaker embedding is trained. The first model, *Baseline*, is the baseline production system described in Section 3. The second model, *VC offline*, follows the setup described in Section 5 where the knowledge to distil is from a vocal classification model trained on the limited offline data (blue line in Figure 1). The third model, *VC FL*, also follows the setup from Section 5, but with a vocal classification model trained with federated learning. The mechanism in [24] was used for local privacy guarantees and the Gaussian mechanism with moments accountant was used for central privacy guarantees.

Multi-task learning of the speaker embedding was conducted on 1.7 million utterances of the trigger phrase from 18700 speakers, all preprocessed by the voice trigger system to extract 520-dimensional supervectors. The softmax layer classifying side information has six outputs and the softmax layer classifying speakers has one output per speaker. The temperature T was set to 10 for all experiments, and the weight γ of the side information classification loss was roughly tuned to balance the losses of the two tasks. Evaluation accuracy was measured on another set of 93600 utterances from the same population of speakers as the training dataset.

The accuracies of speaker identification and side information classification on the evaluation dataset are shown in Table 1. The accuracy on the speaker identification task in the multi-task setting increases relative to the baseline. It is possible that the side information has both a regularizing effect on speaker identification as well as helping propagate signals through the network. The accuracy on the classification of the side information is expected to be close to 100% because the labels are generated from the vocal classification model, and this larger DNN should be able to capture the encoded knowledge of the vocal classification model.

As mentioned in Section 3, a speaker profile is defined by a set of supervectors from enrolment utterances. When evaluating the speaker verification system with the newly trained speaker embedding network, each of 234 speaker profiles available were compared to supervectors of test utterances using Equation 1. A subset of the test utterances were unique utterances from the 234 speakers which had profiles, and the rest originated from imposter speakers that did not match any profile. A total of 65545 pairs of speaker profiles and test utterances were com-

pared and a test utterance was accepted or rejected by applying a threshold θ on the speaker vector score SV_{score}. Performance at the equal error rate (EER) is shown in Table 2 for the three experiments. The multi-task setup with knowledge distillation of a vocal classification model trained on limited central data yields a 1.5% absolute improvement over the baseline. The same setup with a vocal classification model trained on-device with privacy-preserving federated learning yields a 6% relative improvement in EER. These results prove both our second and third hypotheses, namely that privacy-preserving federated learning can be used to improve the vocal classification-based system (due to the gain over the "VC offline" experiment), and that the vocal classification model can be used through multi-task learning to improve the speaker verification system (due to the gain over the "Baseline" system).

7. Conclusions

This paper demonstrates how a centrally trained speaker verification system can be improved by distilling the knowledge of an auxiliary model that was trained with side information on a much broader population using federated learning, while protecting user privacy. Firstly, the auxiliary model, which classifies side information, was trained using data distributed over millions of real devices. Additional experiments simulated different combinations of federated learning and differential privacy when training this model, to highlight the utility/privacy trade-off expected when using such approaches. The accuracy, time to convergence and signal-to-noise ratio clearly show the relative ordering of these approaches in terms of utility. Secondly, the encoded knowledge of the auxiliary model was distilled into the speaker embedding network of an existing speaker verification baseline system using multi-task learning. Finally, a relative improvement of 6% in equal error rate for speaker verification was achieved using this technique while maintaining the same network architecture as the baseline. This result shows that the speaker characteristic knowledge distilled into the speaker verification network resulted in speaker embeddings which are more discriminative.

8. Acknowledgements

9. References

[1] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4052–4056.

[2] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification." in *Interspeech*, 2017, pp. 999–1003.

[3] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-attentive speaker embeddings for text-independent speaker verification." in *Interspeech*, 2018, pp. 3573–3577.

[4] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[5] L. Ferrer, M. Graciarena, A. Zymnis, and E. Shriberg, "System combination using auxiliary information for speaker verification," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4853–4856.

[6] O. Plchot, S. Matsoukas, P. Matějka, N. Dehak, J. Ma, S. Cumani, O. Glembek, H. Hermansky, S. Mallidi, N. Mesgarani *et al.*, "Developing a speaker identification system for the DARPA RATS project," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 6768–6772.

[7] F. Kelly and J. H. Hansen, "Evaluation and calibration of short-term aging effects in speaker verification," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] Z.-Q. Wang and I. Tashev, "Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 5150–5154.

[9] H. Meinedo and I. Trancoso, "Age and gender classification using fusion of acoustic and prosodic features," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[10] S. H. Kabil, H. Muckenhirn, and M. Magimai-Doss, "On learning to identify genders from raw speech signal using CNNs," in *Interspeech*, 2018, pp. 287–291.

[11] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Fifteenth annual conference of the international speech communication association*, 2014.

[12] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition," in *Interspeech*, 2016, pp. 3603–3607.

[13] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS'16*, 2016. [Online]. Available: http://dx.doi.org/10.1145/2976749.2978318

[14] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," 2016.

[15] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning differentially private language models without losing accuracy," *ICLR*, 2018.

[16] T. Ryffel, A. Trask, M. Dahl, B. Wagner, J. Mancuso, D. Rueckert, and J. Passerat-Palmbach, "A generic framework for privacy preserving deep learning," *CoRR*, 2018.

[17] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy." *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2014.

[18] Siri Team, "Personalized Hey Siri," *Apple Machine Learning Journal*, vol. 1, no. 9, 2018. [Online]. Available: https://machinelearning.apple.com/2018/04/16/personalized-hey-siri.html

[19] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Privacy aware learning," 2012.

[20] J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," 2016.

[21] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 10 2015, pp. 1322–1333.

[22] L. Melis, C. Song, E. D. Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," 2018.

[23] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 3–18.

[24] A. Bhowmick, J. C. Duchi, J. Freudiger, G. Kapoor, and R. Rogers, "Protection against reconstruction and its applications in private federated learning." *CoRR*, vol. abs/1812.00984, 2018.

[25] Ú. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, K. Talwar, and A. Thakurta, "Amplification by shuffling: From local to central differential privacy via anonymity," in *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2019, pp. 2468–2479.

[26] B. Balle, G. Barthe, and M. Gaboardi, "Privacy amplification by subsampling: Tight analyses via couplings and divergences," in *Advances in Neural Information Processing Systems*, 2018, pp. 6277–6287.

[27] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[28] E. Marchi, S. Shum, K. Hwang, S. Kajarekar, S. Sigtia, H. Richards, R. Haynes, Y. Kim, and J. Bridle, "Generalised discriminative transform via curriculum learning for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5324–5328.

[29] B. Balle and Y.-X. Wang, "Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising," *arXiv preprint arXiv:1805.06530*, 2018.

[30] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.