PERSONAL ATTRIBUTE LEAKAGE IN FEDERATED SPEECH MODELS

Hamdan Al-Ali[1], Ali Reza Ghavamipour[2], Tommaso Caselli[3]
Fatih Turkmen[3], Zeerak Talat[4], Hanan Aldarmaki[1]

[1]Mohamed bin Zayed University of Artificial Intelligence, UAE
[2]Maastricht University, Netherlands [3]University of Groningen, Netherlands
[4]University of Edinburgh, UK

ABSTRACT

Federated learning is a common method for privacy-preserving training of machine learning models. In this paper, we analyze the vulnerability of ASR models to attribute inference attacks in the federated setting. We test a non-parametric white-box attack method under a passive threat model on three ASR models: Wav2Vec2, HuBERT, and Whisper. The attack operates solely on weight differentials without access to raw speech from target speakers. We demonstrate attack feasibility on sensitive demographic and clinical attributes: gender, age, accent, emotion, and dysarthria. Our findings indicate that attributes that are underrepresented or absent in the pre-training data are more vulnerable to such inference attacks. In particular, information about accents can be reliably inferred from all models. Our findings expose previously undocumented vulnerabilities in federated ASR models and offer insights towards improved security.

Index Terms— Federated Learning, ASR, Privacy

1. INTRODUCTION

Speech technologies are increasingly integrated into daily life, powering applications from virtual assistants to healthcare. These systems rely on machine learning models trained on raw speech to extract acoustic features, enabling robust recognition, synthesis, and understanding across diverse environments and speaker populations. Self-supervised acoustic models——such as Wav2Vec2 [1] and HuBERT [2]–have become central to modern speech processing tasks, including automatic speech recognition (ASR), due to their ability to learn rich representations from unlabeled audio data. Large pre-trained semi-supervised models, like Whisper [3], are also widely used. These models often need fine-tuning on target speaker data for optimal performance, which raises privacy concerns given the sensitive nature of speech. A single voice sample can inadvertently disclose a wide range of personal attributes—including demographic characteristics (e.g., gender [4], age [5]), and accent [6]), clinical conditions (e.g., dysarthria [7]), and emotional states. Federated learning (FL) offers privacy-aware training for such data-sensitive domains by keeping raw audio on user devices and sharing only model updates [8]. This approach is well-suited to speech, which is naturally generated on personal devices, and addresses both privacy and scalability concerns. However, FL is not immune to threats [9]. Even without raw inputs, model updates can leak sensitive information [10, 11]. In the speech domain, prior work has demonstrated speaker re-identification and membership inference against federated ASR models using small samples of target user speech [12, 13]. Beyond these membership attacks, the extent of attribute leakage from model weights remains underexplored. In this work, we investigate the risk of inferring personal attributes from weight updates in FL settings, without access to any target user data. Inferring personal attributes from weights alone could enable profiling, surveillance, or discrimination, and may violate legal protections such as GDRP, HIPAA, and anti-discrimination laws in the US and EU. Additionally, the ADA protects people with disabilities from discrimination [14]. As noted by [15], even revealing attributes like accent or emotional state—without identity—can constitute a serious privacy breach. Our experiments on Wav2Vec2, HuBERT, and Whisper across five attributes (gender, age, accent, emotion, and dysarthria) reveal substantial leakage of some attribute but not others, and our analysis provides evidence that leakage is correlated with attribute coverage and representation in the base model.

2. METHODOLOGY

2.1. Threat Model

We consider a passive adversary at the server side of an FL system. The attacker does not interfere with training but performs a white-box attack, leveraging access to the global model W_g (distributed before personalization) and the locally trained model W_s for a target speaker s, obtained after fine-tuning W_g on local speech. The attacker has no access to raw audio, transcripts, or metadata. We assume each model is fine-tuned on a single utterance per speaker, with personalization occurring locally. Given W_g and updated weights W_s, the attacker's goal is to infer *private and protected at-*

tributes using only these weights. To support the attack, the adversary leverages public datasets to simulate fine-tuning on speech with known attributes. These simulations (shadow models [10]) provide labeled training data for auxiliary classifiers mapping weight updates to attributes. This reflects a realistic threat model in which the attacker follows the FL protocol but extracts sensitive information from updates.

2.2. Attribute Inference Attack Model

To test attribute inference feasibility, we simulate a passive attacker building a labeled dataset for the target attribute (e.g., gender) using public datasets. We collect labeled samples $\{(x_i, y_i)\}_{i=1}^{n}$, with $y_i \in \{\text{male}, \text{female}\}$, and fine-tune W_g on each sample x_i, yielding n shadow models $W_i, i = 1, \ldots, n$. For each W_i, we extract summary statistics—mean, standard deviation, minimum, and maximum—from each parameter tensor $p \in W_i$, concatenating them into a fixed-length feature vector $z_i \in \mathbb{R}^d$:

$$z_i = \text{concat}\left([\mu_p, \sigma_p, \min(p), \max(p)] \text{ for each } p \in W_i\right)$$

Here, d is the total dimensionality, determined by the number of parameter tensors.[1] We then compute class centroids (e.g., $\bar{z}_m$ for male, $\bar{z}_f$ for female) by averaging the feature vectors of shadow models within each class:

$$\bar{z}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} z_i \quad \text{where } z_i \in \text{class } c$$

To infer the attribute of a new, unseen model from a target speaker, W_s, the attacker extracts its feature vector z_s in the same way and compares it to the class centroids using normalized Euclidean distance. The predicted class corresponds to the closest centroid:

$$\hat{c} = \arg\min_c \left(\frac{\|z_s - \bar{z}_c\|_2}{\|z_s\|_2 \cdot \|\bar{z}_c\|_2} \right) \tag{1}$$

This approach scales naturally to multi-class settings.

3. EXPERIMENTS

The proposed attack is evaluated on three personal attributes: gender, age, and accent (correlated with country of origin)—as well as two speech-related conditions: emotional state and the presence of a speech disorder (dysarthria).

3.1. Datasets

We use publicly available speech datasets containing the target attributes. For *gender*, *age*, and *accent*, we use the **Speech Accent Archive** (SAA) [16], which provides controlled recordings of speakers reading the same scripted English sentence, ideal for isolating speaker traits. For *speech disorders*, we use the **TORGO** dataset [17], with recordings from 15 speakers—8 with dysarthria (due to cerebral palsy or amyotrophic lateral sclerosis) and 7 without. For *emotions*, we use the **RAVDESS** dataset [18], containing acted emotional speech from 24 professional speakers (12 female, 12 male) reading matched statements in a North American accent.

3.2. Experimental Conditions

We evaluate five attributes in total, each designed as a binary detection task. For **gender detection**, we use 200 native-English speakers from SAA, comprising 100 male and 100 female speakers, with a 75/25 train-test split, balanced by gender. The **age detection** task is also framed as a binary classification between two well-separated age ranges, 18–24 and 35–44 years, to avoid ambiguous boundary cases since age is continuous. We used 70 male native-English speakers (35 per class) from SAA and report results using five-fold cross-validation. For binary **Accent detection**, we used 200 speakers from SAA, equally divided between native and non-native English speakers, with a 75/25 split. **Emotion detection** consists of three binary tasks from RAVDESS—*Calm* versus *Angry*, *Happy* versus *Sad*, and *Calm* versus *Fearful*—each based on one fixed statement from 24 speakers (48 utterances) and evaluated with 6-fold cross-validation. For speech disorder detection, we select 10 utterances per speaker from TORGO, prioritizing longer ones. This is the only task with varying linguistic content, and it is evaluated using leave-one-speaker-out cross-validation.[2]

3.3. ASR Models

We evaluate three widely used self-supervised speech models: **Wav2Vec2**-Base,[3] which has 95 million parameters and was pre-trained and fine-tuned for ASR on 960 hours of LibriSpeech [1]; **HuBERT**-Large,[4] which has 300 million parameters and was also trained and fine-tuned on 960 hours of LibriSpeech [2]; and **Whisper**-Small,[5] which has 244 million parameters and was trained on 680,000 hours of multilingual and multitask labeled data [3]. LibriSpeech [19] is a corpus of read English speech by native adult speakers, both male and female, derived from audiobooks.

4. RESULTS

We evaluated the attribute inference attack (Section 2.2) for each binary condition (Section 3.2) using Wav2Vec2, HuBERT, and Whisper as base ASR models. Table 1 reports the

[1] For example, Whisper$_{small}$ (244M parameters) is reduced to $d = 1916$.

[2] Cross-validation reduces random effects but cannot eliminate systematic linguistic differences.

[3] huggingface.co/facebook/wav2vec2-base-960h

[4] huggingface.co/facebook/hubert-large-ls960-ft

[5] huggingface.co/openai/whisper-small

Task	Wav2Vec2	HuBERT	Whisper
Gender: Male / Female	64%	63%	46%
Age: 18-24 / 35-44	100%	97%	94%
Accent: Native / Accented	100%	80%	93%
Dysarthria: True / False	59%	76%	81%
Emotion: Calm / Angry	52%	67%	83%
Emotion: Happy / Sad	50%	54%	75%
Emotion: Calm / Fearful	46%	48%	73%

Table 1. Attack success rate in terms of accuracy (%)

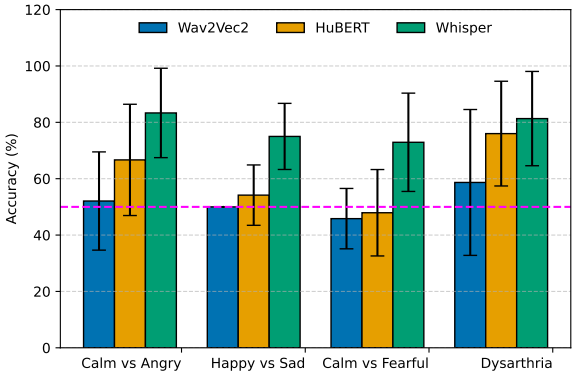


Fig. 1. Mean accuracy for binary classification of Emotion & Dysarthia with 95% confidence intervals.

mean classification accuracy (attack success rate) with cross-validation applied where relevant. Gender was the most difficult attribute to predict, with accuracies ranging from 46% to 64%. Accent and age exhibited the strongest leakage, reaching 80–100% accuracy across models; Wav2Vec2 achieved 100% for both. Whisper consistently leaked attributes at >70% accuracy, except for gender. Among emotions, anger yielded the highest detection accuracy, while other categories showed more variability. Variability in cross-validated results was assessed using standard error and 95% confidence intervals (t-distribution), shown in Figure 1 (chance level: dashed magenta line). Age detection results were statistically significant for all models (CIs $\geq$ 95%), so they are omitted from the figure. Whisper produced the most reliable results for all emotion categories and dysarthria despite small sample sizes, whereas Wav2Vec2 and HuBERT often performed near chance. Overall, these findings confirm that model updates can reveal sensitive attributes, with leakage strength depending on both the attribute and the model architecture.

5. PER-LAYER ANALYSIS

We analyzed where attribute-specific information is encoded in Wav2Vec2 by applying the attribute inference attack (Section 2.2) to parameter tensors from individual transformer layers. Accuracy was evaluated for age, emotion (calm vs. angry), and dysarthria, with results in Figure 2 (mean accuracy

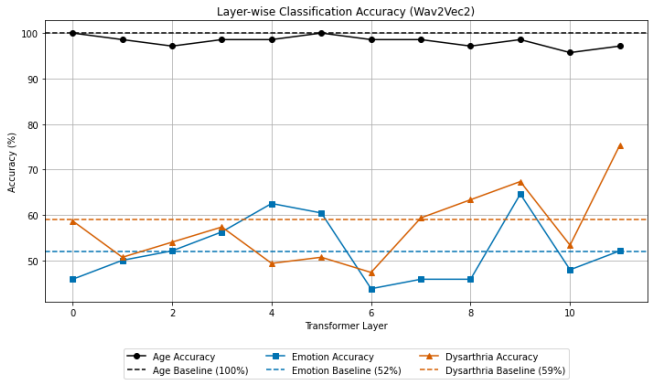


Fig. 2. Layer-wise classification accuracy using Wav2Vec2 encoder layers for three tasks. The dashed lines show the result obtained using the full set of parameters.

across cross-validation folds). Age detection remained consistently high across all layers, indicating that effective attacks may be implemented using only a subset of parameters. In contrast, emotion and dysarthria showed greater variability, with certain mid-to-late layers outperforming full-model inference. These findings suggest that attribute-specific information is distributed unevenly across layers, and that targeting specific layers can sometimes enhance inference accuracy.

6. FINE-GRAINED ACCENT CLASSIFICATION

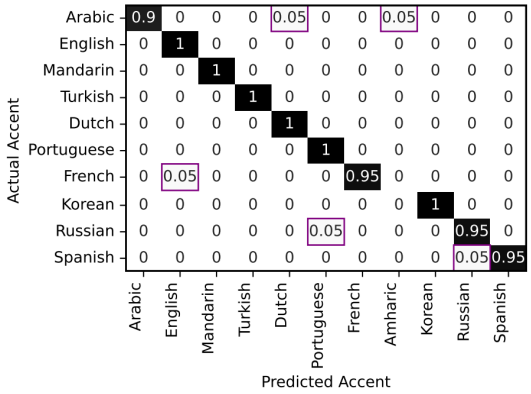
We further analyze accent prediction, as it yielded some of the most consistent binary classification results. The SAA corpus provides controlled recordings from speakers with diverse native accents, enabling multi-class classification. We use Wav2Vec2, which achieved perfect binary accuracy, and select the ten most represented accents—English, Spanish, Arabic, Dutch, Mandarin, French, Korean, Portuguese, Russian, and Turkish—plus Amharic (training only). For each accent, 15 samples are used for training and 20 unseen speakers for testing; Amharic has no test set due to limited data. As shown in Figure 3a, the attack achieves $\geq$ 90% accuracy for all tested accents[6].

6.1. Effect of Targeted Fine-Tuning

One observation to note from the results so far is the high disparity in detection accuracy between different attributes. Gender, for instance, was predicted at near chance levels using FL model weights, in spite of it being one of the most distinguishable features in raw speech [4]. On the other hand, accent was predicted almost perfectly, even in multi-class setting. A likely factor behind the high leakage is the mismatch between ASR pre-training data (e.g., LibriSpeech) and target-speaker accents, which can induce larger and more distinct

[6]Similar results were observed for HuBERT and Whisper.

(a) Before (incorrect predictions in magenta).



(b) After (correct predictions in green).

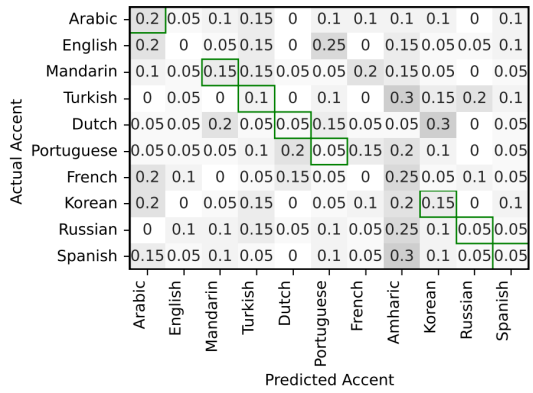


Fig. 3. Wav2Vec2 accent classification results before and after fine-tuning. Results are shown as proportions of the total for each class, so the diagonal corresponds to precision. Note: Amharic is only used in training the attribute classifier, with no test samples.

gradients due to higher surprisal. Another possible cause is catastrophic forgetting [20], where attributes included in early training but unseen in later updates still produce high surprisals. To test this, we fine-tune the global model on previously unused speakers of the ten most frequent accents in SAA (up to 20 sample per accent, total 137 samples). After fine-tuning, the attack's success drops sharply, with no accent exceeding 0.2 accuracy (Figure 3b). This confirms that limited accent diversity in training amplifies leakage, and that training on a wide range of accents not only serves inclusion [21] but also mitigates accent-based attribute inference. This may also explain the negative results for gender, as gender variations are well represented in the underlying models.

6.2. Generalization to Unseen Accents

Fine-tuning on accented speech can defend against our attack for known accents, but what about unseen accents? We test this by selecting five accents from SAA absent in prior experiments: Japanese, Italian, German, Polish, and Macedonian. For each, we use six additional samples for training shadow models and 14 unseen speakers for testing. As shown in Table 2, the attack achieves high precision and recall across all accents. These results indicate that accents not adequately represented in ASR pre-training or fine-tuning data remain highly vulnerable, as their acoustic characteristics yield more distinguishable parameter updates. Partial coverage may therefore be insufficient for privacy preservation, particularly during initial client enrollment when novel attributes are more likely to appear.

Accent	Precision	Recall	F1-score
Japanese	1.00	1.00	1.00
Italian	0.81	0.93	0.87
German	0.85	0.79	0.82
Polish	1.00	1.00	1.00
Macedonian	1.00	1.00	1.00

Table 2. Wav2Vec2 Accent Classification - Generalization to Unseen Accents

7. CONCLUSION

This study presents an investigation into attribute leakage in federated Automatic Speech Recognition (ASR). We show attack success rates for age, accent, dysarthria, and emotion detection that exceed chance levels, particularly using the Whisper model as the base ASR system. Age and accent, in particular, are highly vulnerable to such attribute attacks. Inferring such attributes–particularly in combination–can enable profiling, discrimination, and surveillance. For example, a speech disorder label combined with IP/geolocation data can uniquely identify a user. While it is understood that FL improves privacy by keeping audio on-device, our work highlights that this protection is uneven—certain traits (e.g., accent and age) can be inferred successfully more than others. We also show experimentally that this leakage may be explained by the lack of coverage of these traits in the training data of the underlying models. When models were exposed to even a small, diverse sample of accented speech prior to client updates, the effectiveness of the leakage attack significantly declined for the covered accents. This suggests that pre-training on diverse demographies and conditions prior to federated learning could result in better robustness against privacy attacks in FL settings. Finally, our controlled experiments demonstrate privacy leakage using minimal attribute training data, but results on gender detection also show that complex and interdependent attributes may be hard to detect in real-world scenarios, especially if these attributes are well represented in the model's initial training data prior to FL setup. This work reveals previously undocumented vulnerabilities in federated ASR models and demonstrates a potential approach to improve their security.

8. REFERENCES

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[2] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman 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.

[3] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28492–28518.

[4] Yeptain Leung, Jennifer Oates, and Siew Pang Chan, "Voice, articulation, and prosody contribute to listener perceptions of speaker gender: A systematic review and meta-analysis," *Journal of Speech, Language, and Hearing Research*, vol. 61, no. 2, pp. 266–297, 2018.

[5] Anthony Mulac and Howard Giles, "'your're only as old as you sound': Perceived vocal age and social meanings," *Health Communication*, vol. 8, no. 3, pp. 199–215, 1996.

[6] Cynthia G Clopper and David B Pisoni, "Effects of talker variability on perceptual learning of dialects," *Language and speech*, vol. 47, no. 3, pp. 207–238, 2004.

[7] Dong-Her Shih, Ching-Hsien Liao, Ting-Wei Wu, Xiao-Yin Xu, and Ming-Hung Shih, "Dysarthria speech detection using convolutional neural networks with gated recurrent unit," in *Healthcare*, 2022, vol. 10, p. 1956.

[8] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[9] Natalia Tomashenko, Salima Mdhaffar, Marc Tommasi, Yannick Estève, and Jean-François Bonastre, "Privacy attacks for automatic speech recognition acoustic models in a federated learning framework," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6972–6976.

[10] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 3–18.

[11] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *2019 IEEE symposium on security and privacy (SP)*, 2019, pp. 691–706.

[12] Guangke Chen, Yedi Zhang, and Fu Song, "SLMIA-SR: speaker-level membership inference attacks against speaker recognition systems," in *31st Annual Network and Distributed System Security Symposium, NDSS 2024, San Diego, California, USA, February 26 - March 1, 2024*. 2024, The Internet Society.

[13] Wei-Cheng Tseng and Wei-Tsung Kao and Hung-yi Lee, "Membership Inference Attacks Against Self-supervised Speech Models," in *Interspeech 2022*, 2022, pp. 5040–5044.

[14] Shehu Mohammed and Neha Malhotra, "Ethical and regulatory challenges in machine learning-based healthcare systems: A review of implementation barriers and future directions," *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, vol. 5, no. 1, pp. 100215, 2025.

[15] Emma Ritter, "Your voice gave you away: The privacy risks of voice-inferred information," *Duke LJ*, vol. 71, pp. 735, 2021.

[16] Steven H Weinberger and Stephen A Kunath, "The speech accent archive: towards a typology of english accents.," *Language & Computers*, vol. 73, no. 1, 2011.

[17] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language resources and evaluation*, vol. 46, pp. 523–541, 2012.

[18] Steven R Livingstone and Frank A Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, pp. e0196391, 2018.

[19] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev 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.

[20] Robert M French, "Catastrophic forgetting in connectionist networks," *Trends in cognitive sciences*, vol. 3, no. 4, pp. 128–135, 1999.

[21] Lucas Maison and Yannick Esteve, "Improving accented speech recognition with multi-domain training," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.