PRIVACY ATTACKS FOR AUTOMATIC SPEECH RECOGNITION ACOUSTIC MODELS IN A FEDERATED LEARNING FRAMEWORK

Natalia Tomashenko[1], Salima Mdhaffar[1], Marc Tommasi[2], Yannick Estève[1], Jean-François Bonastre[1]

[1] LIA – University of Avignon – France [2] Inria – University of Lille – France

1 Introduction

Context

- Federated learning: collaborative training of machine learning models while keeping the raw training data decentralized.
- Automatic speech recognition (ASR) acoustic models (AM).
- Indirect privacy leakage: adversary can access the model parameters and aims to infer information about the speaker identity.

Research question

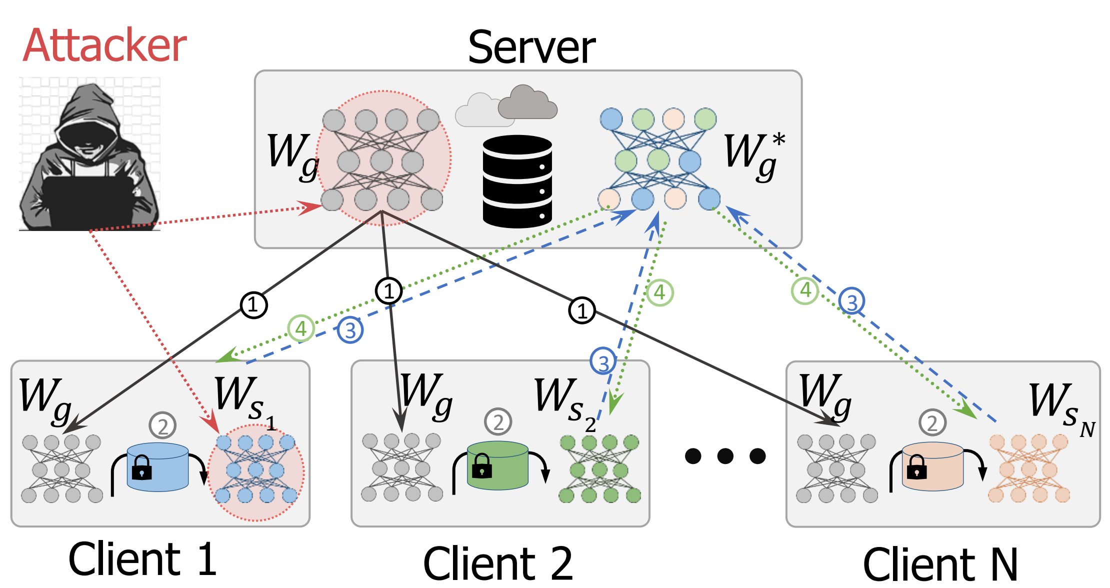
How to effectively and easily analyze (speaker) information in neural network AMs?

Proposed approach

- Use an external indicator dataset to analyze the footprint of AMs on this data.

2 Federated learning and privacy preservation scenario

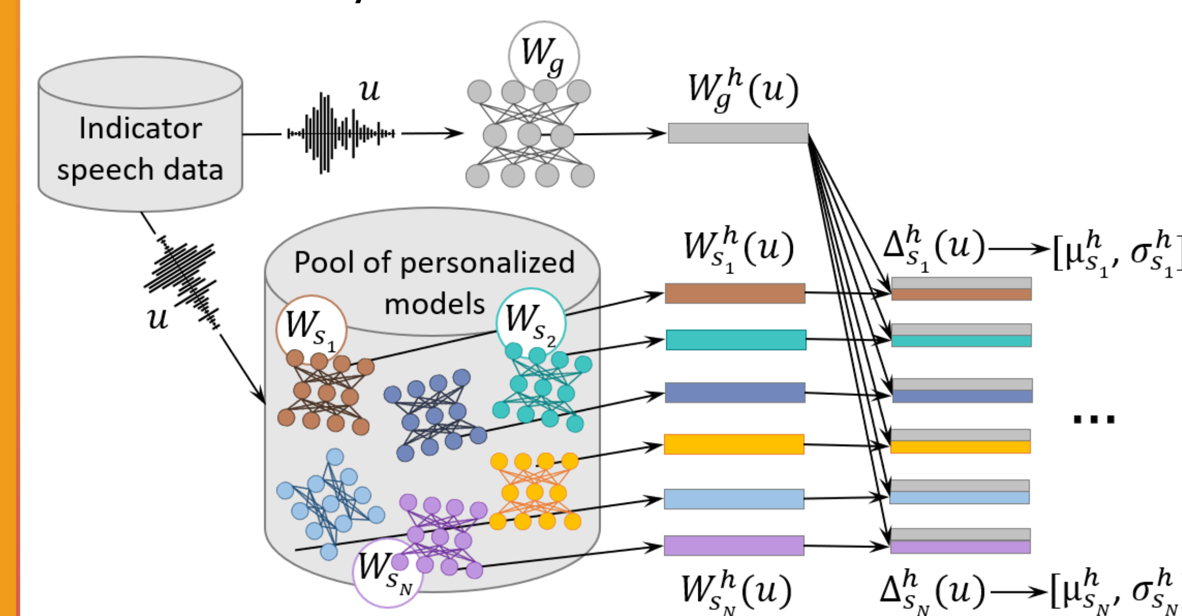
- Users (clients): share their personalized model updates with the server; no speech data is transmitted.
- Attacker has access: global model W_g & personalized model W_s of the target speaker s enrolled in the FL system & other personalized models of speakers: $W_{s_1},\dots,W_{s_N}$.
- Attacker's objective: automatic speaker verification (ASV) by using the enrollment model W_s and test trials in the form of models $W_{s_1},\dots,W_{s_N}$.



3 Attack models

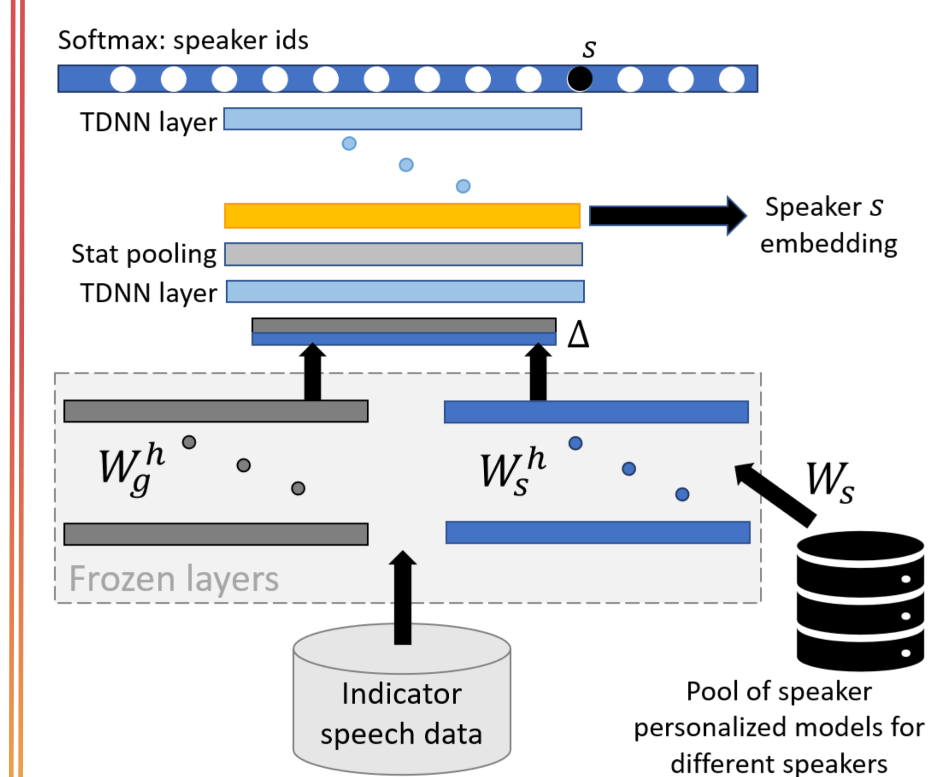
Approach: capture information about the identity of speaker s from the corresponding speaker-adapted model W_s and the global model W_g by comparing the outputs of these two neural AMs taken from hidden layers h on some external speech dataset → analyze the footprint of the NN model on the indicator data.

A1: comparative statistical analysis of the NN outputs from hidden layer h



$$\rho(W^h_{s_i}, W^h_{s_k}) = \alpha_\mu \frac{\|\mu^h_{s_i} - \mu^h_{s_k}\|_2}{\|\mu^h_{s_i}\|_2\|\mu^h_{s_k}\|_2} + \alpha_\sigma \frac{\|\sigma^h_{s_i} - \sigma^h_{s_k}\|_2}{\|\sigma^h_{s_i}\|_2\|\sigma^h_{s_k}\|_2}$$

A2: neural-network (NN) based approach



4 Experimental results

Data and models

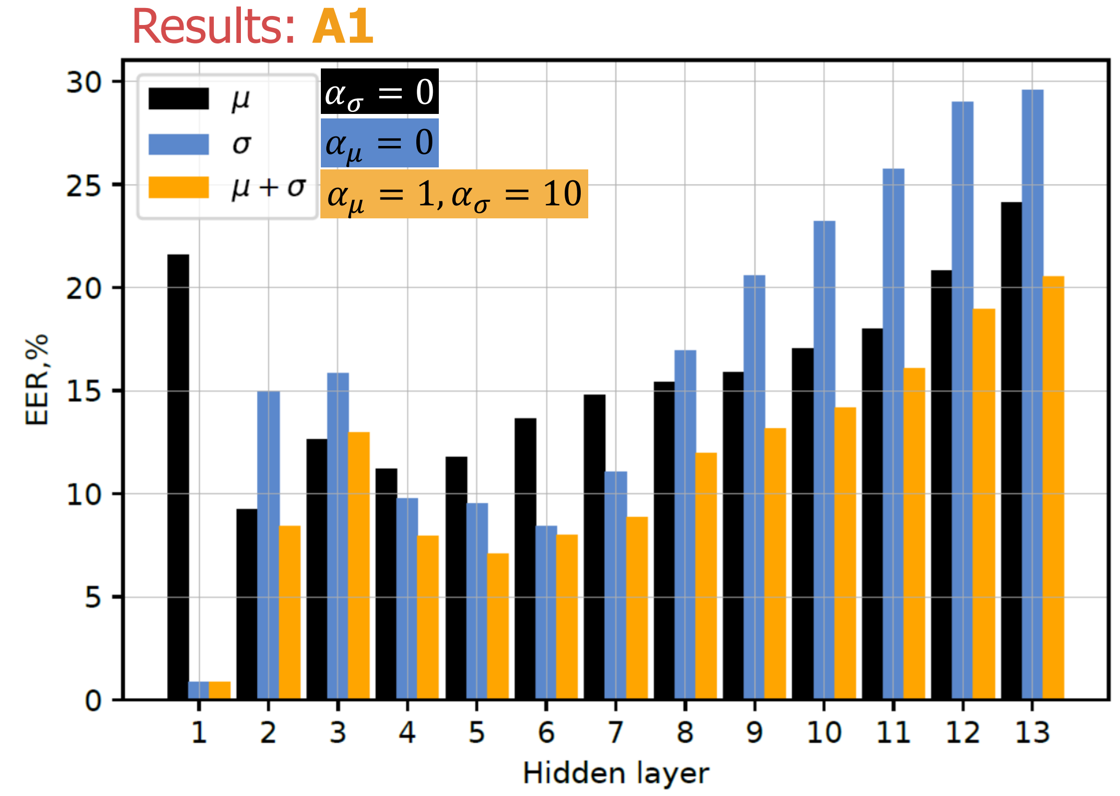
	Train-G Global model	Part-1 Train	Part-2 Test	Indicator
Duration, h	200	86	73	0.5
# speakers	880	736	634	32
# models	-	1300	1079	-

TED-LIUM 3 corpus. Adaptation data for personalized models W_{s_i}: 4 minutes per model

Results: A1 & A2 EER,%

Attack model	Hidden layer #1	Hidden layer #5
A1	**0.86**	7.11
A2	12.31	**1.94**

Results: **A1**



5 Conclusions

- ASR acoustic models are vulnerable to privacy attacks which aim to infer speaker identity from the updated (personalized) models.
- We propose an efficient method to analyze information in neural network AMs based on a neural network footprint on the indicator dataset.
- On the TED-LIUM 3 corpus both attack models are shown to be very effective:
  - EER=1% for the simple attack model **A1**.
  - EER=2% for the NN attack model **A2**.
- The first layer of personalized AMs contains a large amount of speaker information that is mainly contained in the standard deviation values computed on the indicator dataset.
- **Future work:** developing an efficient ASV system based on this property of the adapted NN models