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

Wei-Cheng Tseng

Wei-Tsung Kao

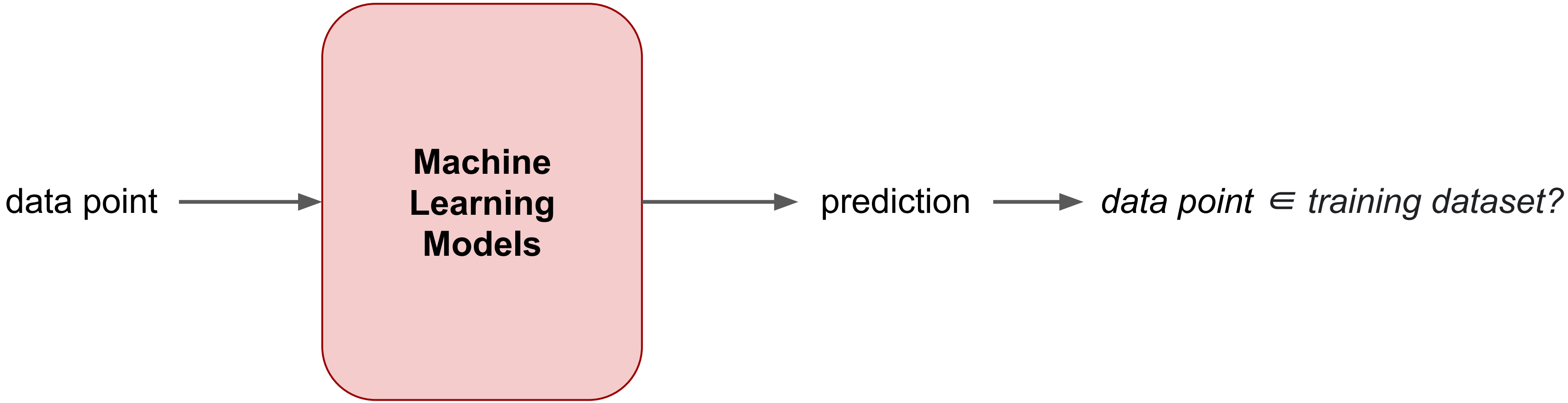
Hung-yi Lee

Motivation

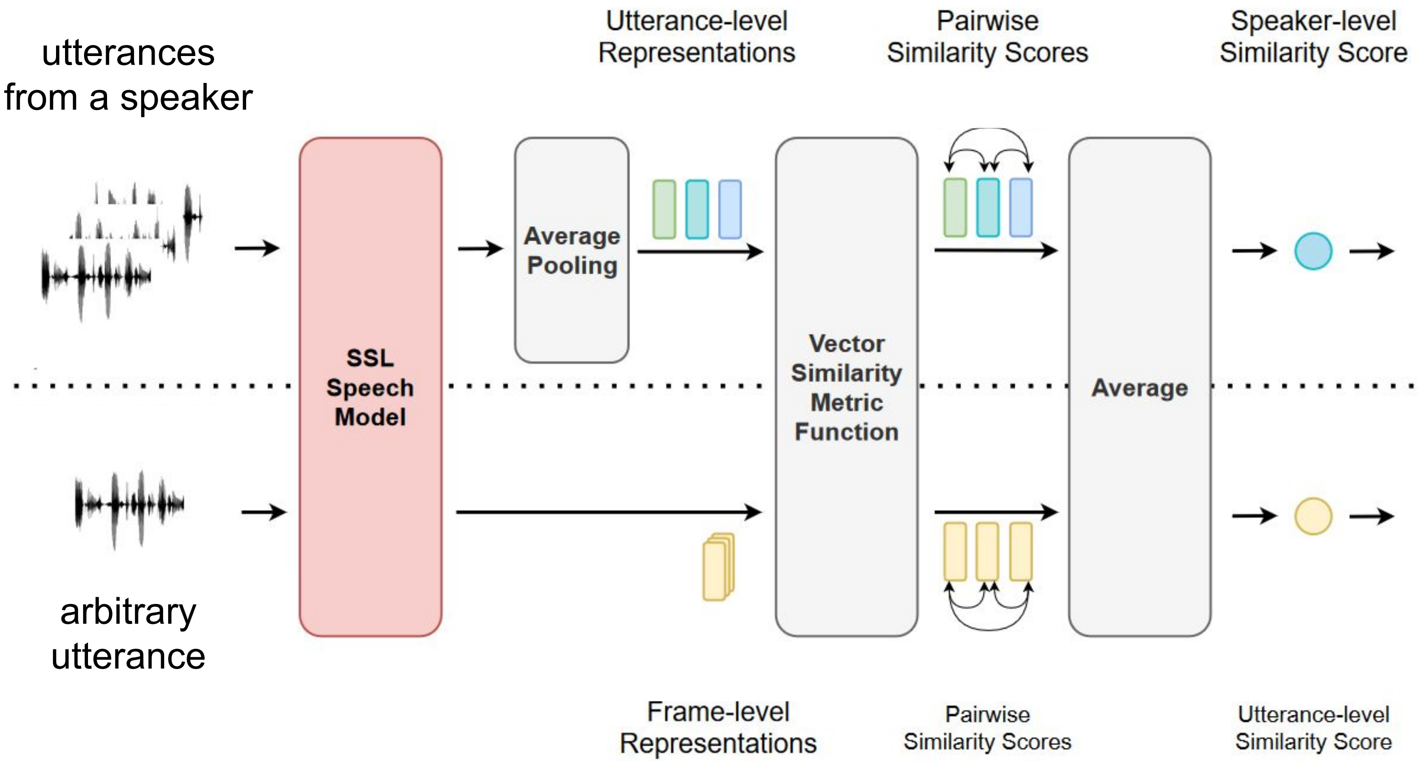
- Self-supervised learning (SSL) models now becomes an important component of speech processing.
- To deploy SSL models to products, it is inevitable for us to make sure whether there are potential privacy issues in SSL models or not.
- In this paper, we study and propose a basic method to perform **membership inference attack** to SSL speech models.

Membership inference attack (MIA)

- Given a model and an exact datapoint, the adversary want to know whether this datapoint was used to train the model or not
- Serves as canary of more severe privacy issues



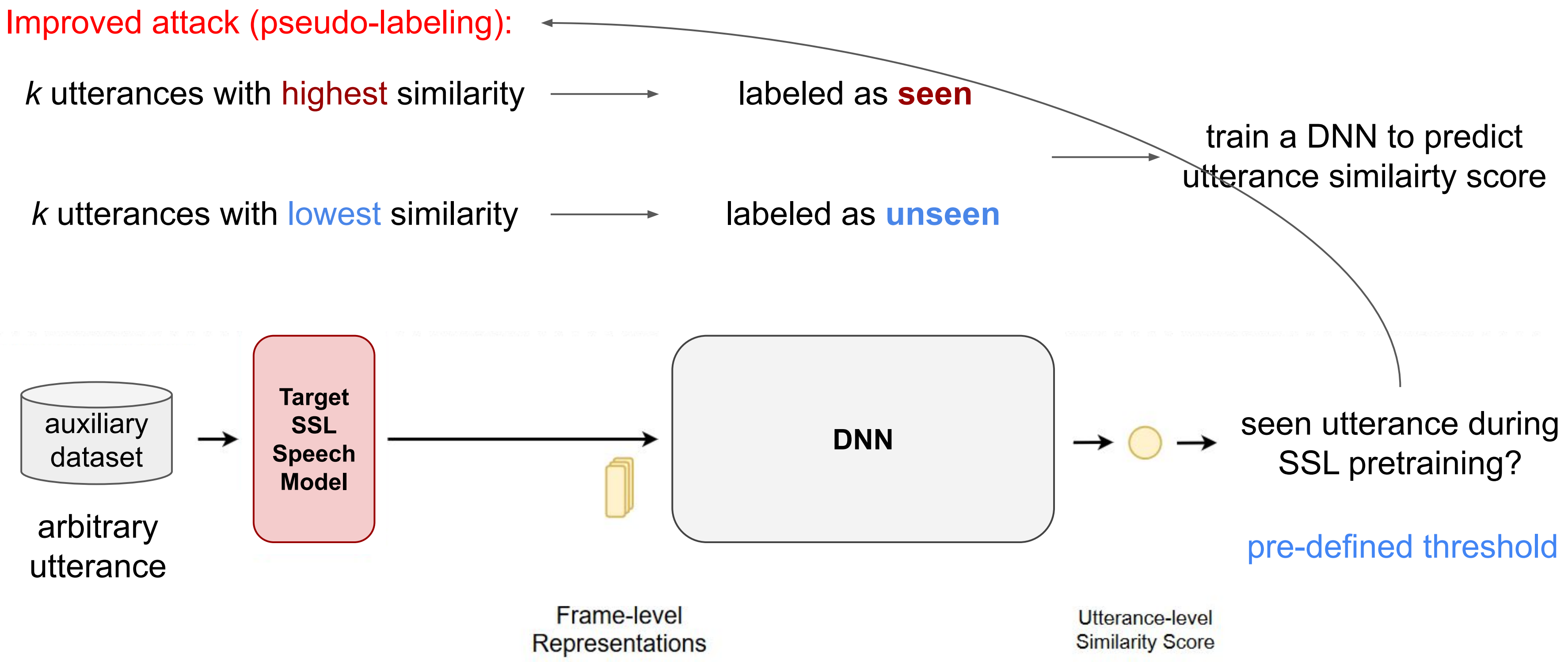
Proposed black-box MIA framework


utterances
from a speaker
Utterance-level Representations
Pairwise Similarity Scores
Speaker-level Similarity Score
SSL Speech Model
Average Pooling
Vector Similarity Metric Function
Average
arbitrary utterance
Frame-level Representations
Pairwise Similarity Scores
Utterance-level Similarity Score

Speaker-level attack: seen speaker during SSL pretraining?

Utterance-level attack: seen utterance during SSL pretraining?

Utterance-level MIA: seen utterance during pretraining?



Improved attack (pseudo-labeling):

k utterances with highest similarity → labeled as **seen**

k utterances with lowest similarity → labeled as **unseen**

→ train a DNN to predict utterance similairty score

auxiliary dataset

arbitrary utterance

Target SSL Speech Model

Frame-level Representations

DNN

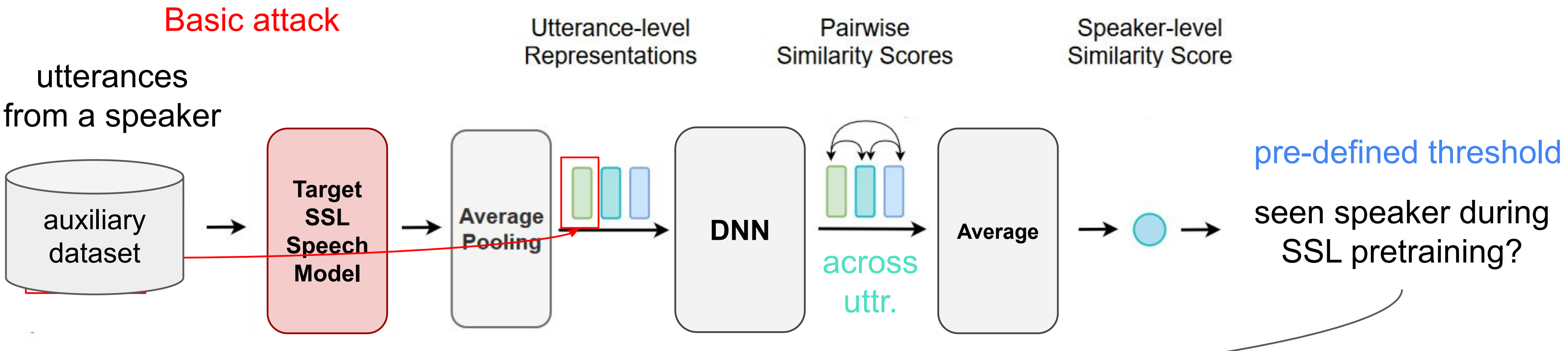
Utterance-level Similarity Score

seen utterance during SSL pretraining?

pre-defined threshold

Basic attack

Speaker-level MIA: seen speaker during pretraining?


Basic attack
utterances from a speaker
auxiliary dataset
Target SSL Speech Model
Average Pooling
Utterance-level Representations
DNN
Pairwise Similarity Scores
across uttr.
Speaker-level Similarity Score
Average
pre-defined threshold
seen speaker during SSL pretraining?

Improved attack (pseudo-labeling):

k speakers with highest similarity → labeled as **seen**

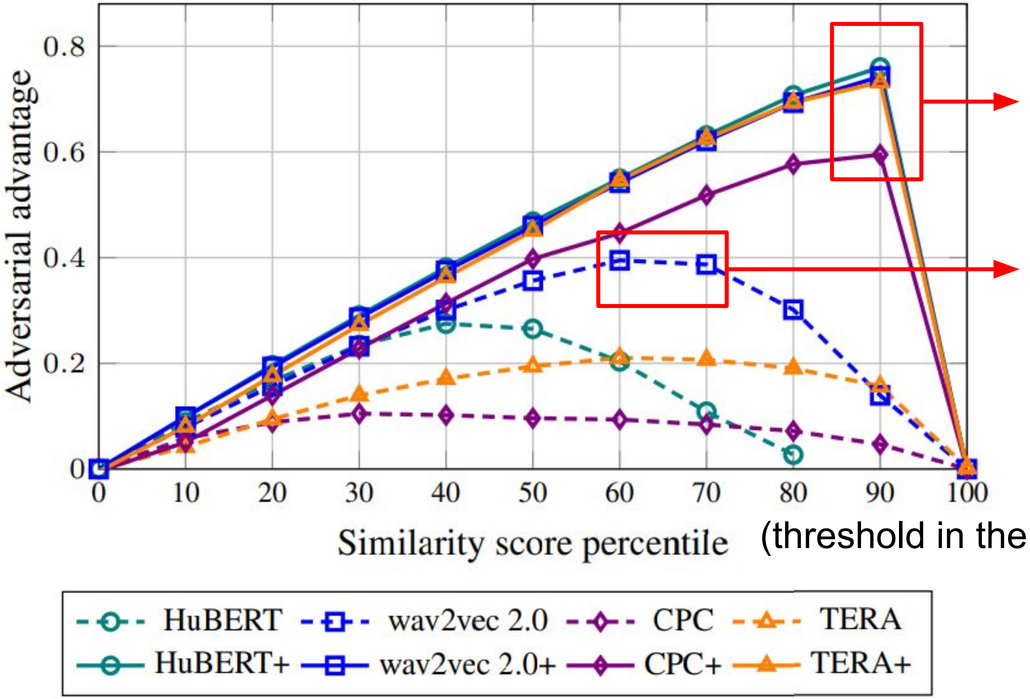
k speakers with lowest similarity → labeled as **unseen**

→ train a DNN to predict pairwise similarity score

Experiment setup

1. SSL speech models:
   - HuBERT, wav2vec 2.0, CPC and TERA
2. Dataset: LibriSpeech
   - seen: train-clean-100
   - unseen: dev-clean, dev-other, test-clean, test-other
3. Predefined vector similarity metric function:
   - utterance-level attack: 1 - (cosine similarity)
   - speaker-level attack: cosine similarity
4. *k* parameter used to train the DNN:
   - utterance-level attack: 500 (utterances)
   - speaker-level attack: 1 (speaker)
5. Evaluation: Adversarial advantage: True Positive Rate - False Negative Rate

Utterance-level MIA result

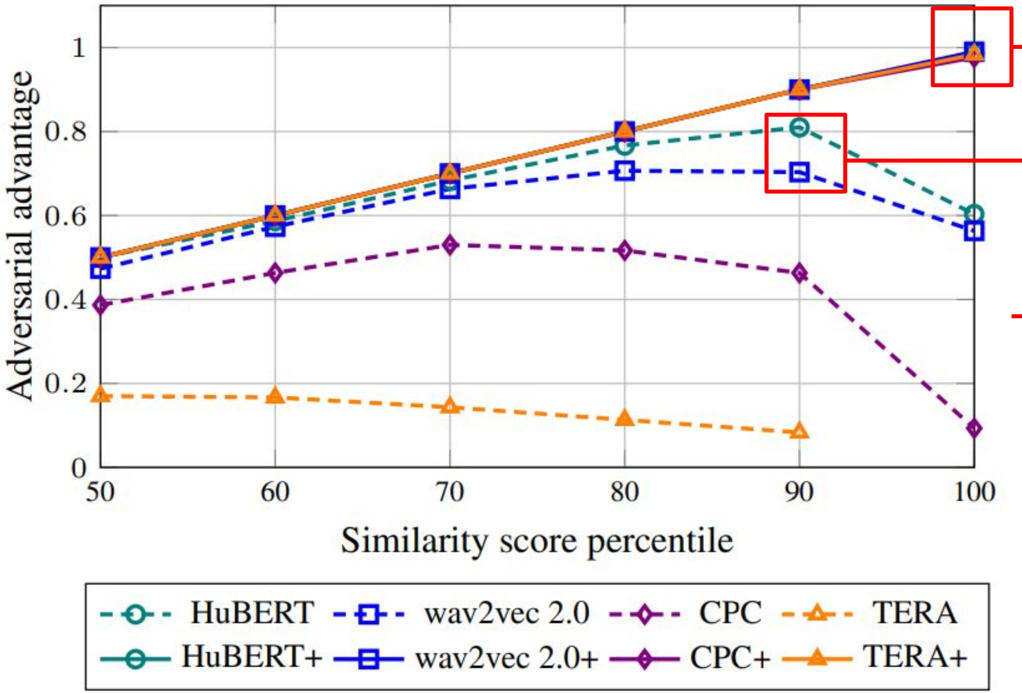


DNN significantly boosts the attack.

wav2vec2 is already vulnerable when applying basic attack

(threshold in the binary cls.)

Speaker-level MIA result



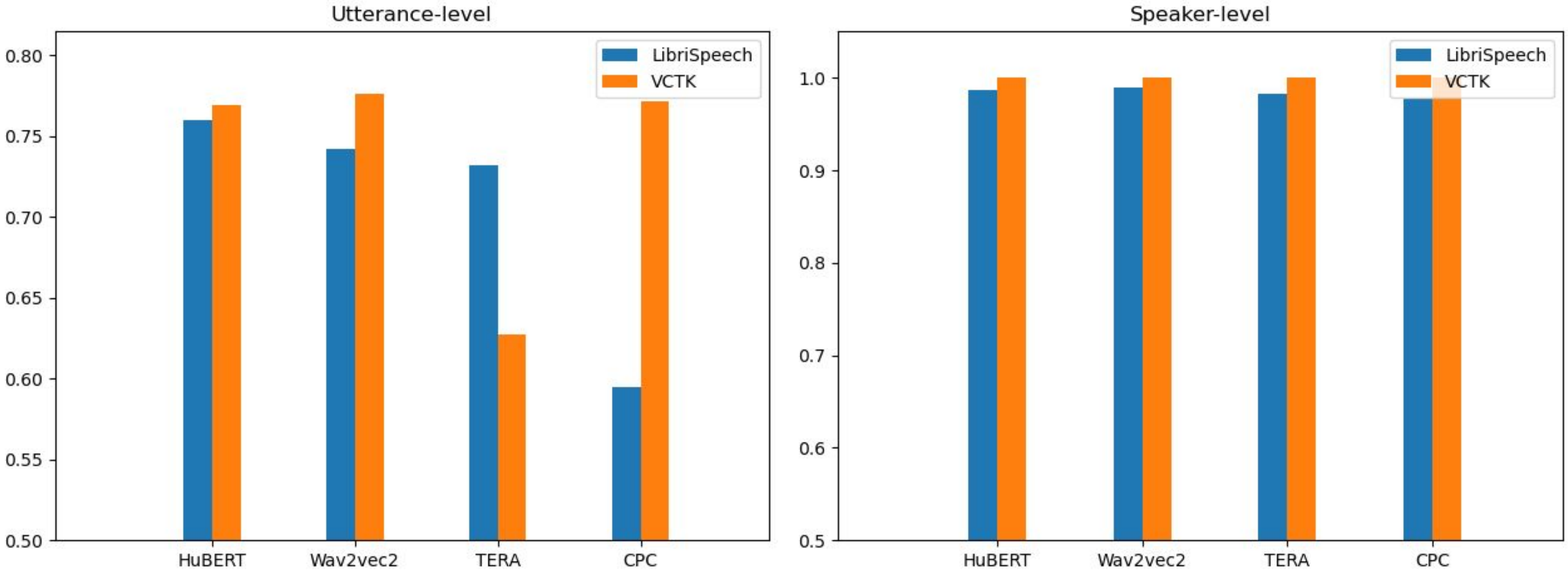
near-optimal performance for improved attack

Better models suffer from the attack more when applying basic attack

Speaker-level attack is more severe.

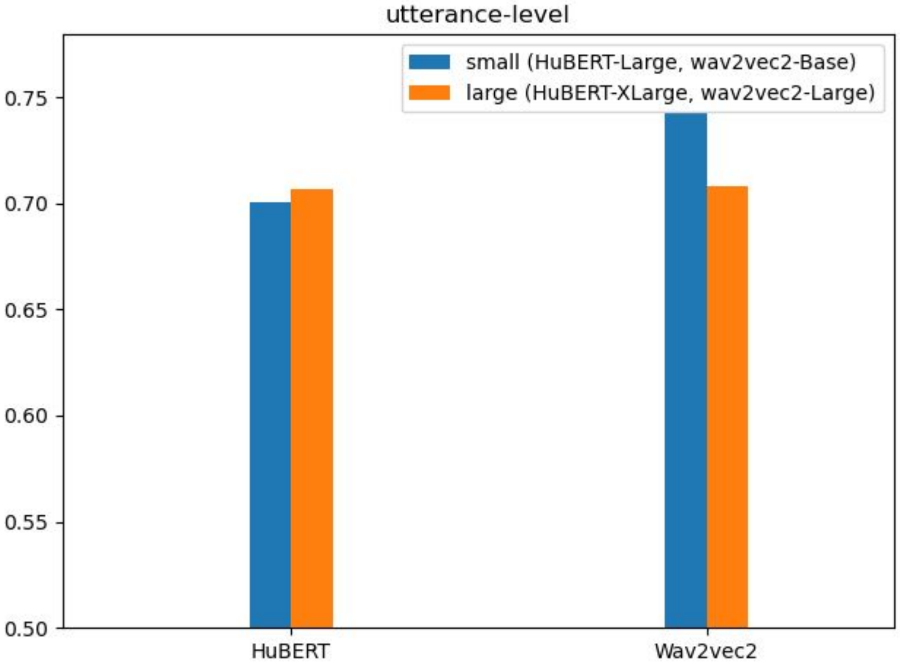
Choice of the auxiliary dataset

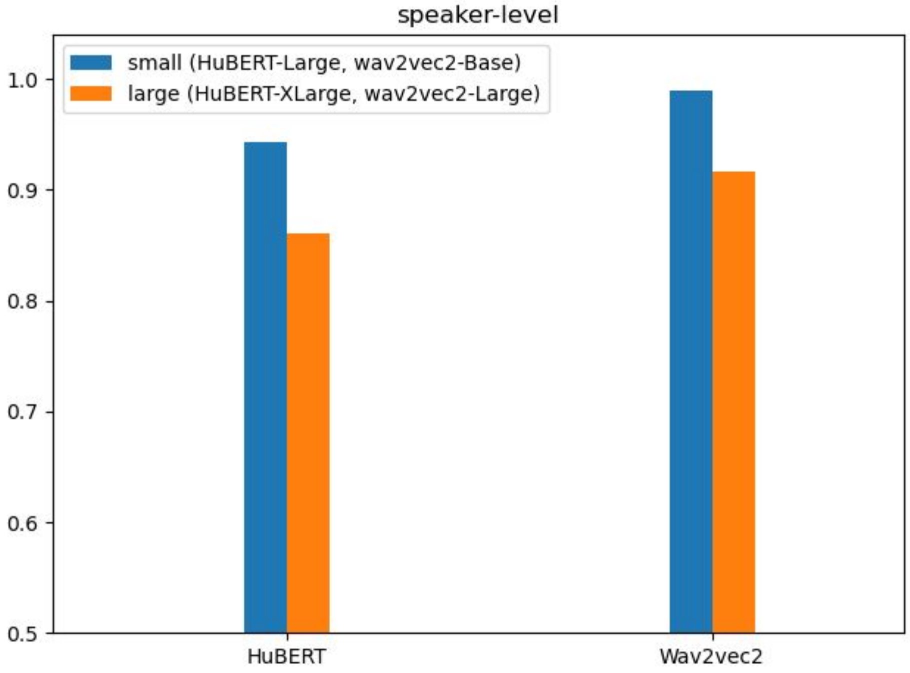
The proposed attack is robust to the choices of the auxiliary dataset.



Effect of the model size

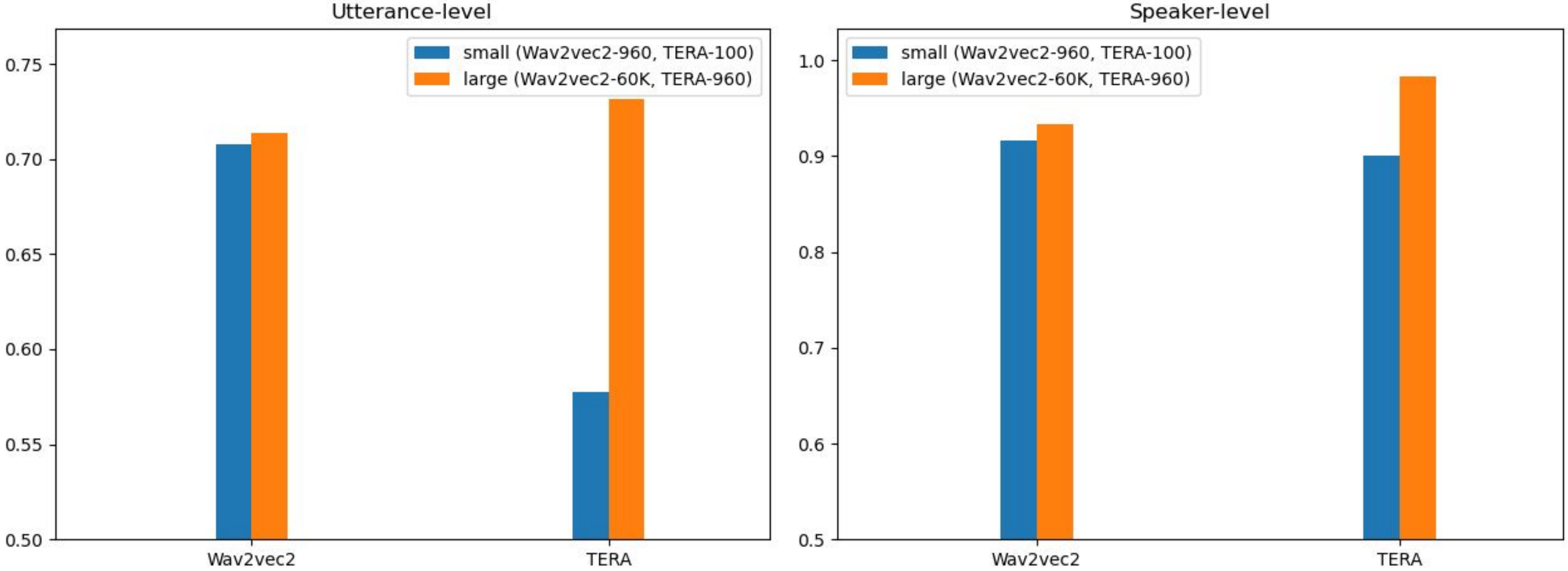
Speaker-level: Larger model -> lower attack performance





Effect of the pre-training corpus size

Smaller pre-training corpus -> lower attack performance



Conclusion

1. We propose the first membership inference attack against SSL speech models under black-box access.
2. SSL models are vulnerable to speaker-level and utterance-level attacks to reveal sensitive membership information.
3. The proposed attack is robust to the choice of the auxiliary datasets.
4. We also conduct ablation study on how the model size and the corpus size affect the attack performance.

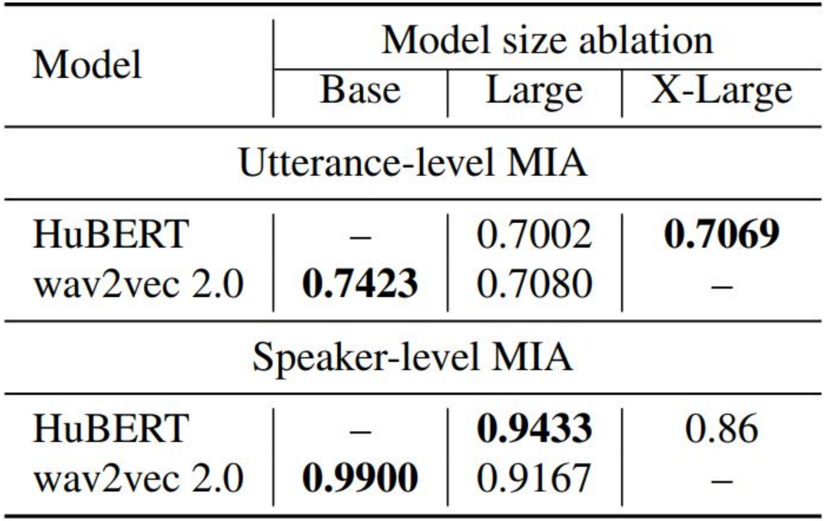
Ablation study

1. Size of the model:
   a. HuBERT-{Large, XLarge} pretrained on LibriLight
   b. Wav2vec2-{Base, Large} pretrained on LibriSpeech-960hr
2. Size of pre-training datasets:
   a. Wav2vec2-Large pretrained on {LibriSpeech-960hr, LibriLight}
   b. Tera pretrained on {LibriSpeech-100hr, LibriSpeech-960hr}
3. Choices of unseen data:
   a. LibriSpeech dev and test set
   b. VCTK dataset

Effect of the model size

Utterance-level: No consistent results

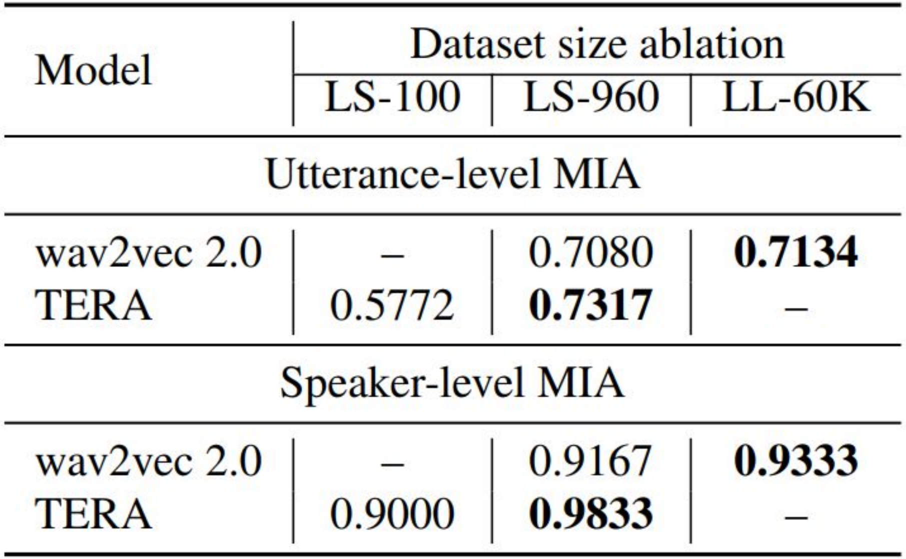
Speaker-level: Larger model -> lower attack performance



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	–

Effect of the pre-training corpus size

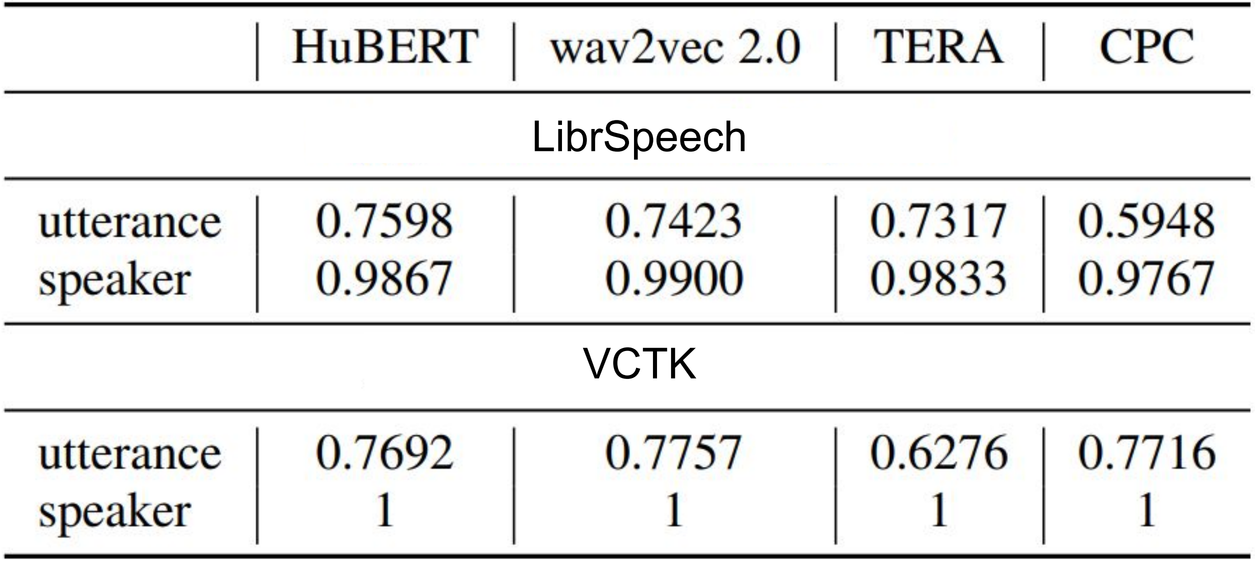
Smaller pre-training corpus -> lower attack performance



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**	–

Effect of unseen dataset

The proposed attack is robust to the choices of unseen dataset.



	HuBERT	wav2vec 2.0	TERA	CPC
LibrSpeech				
utterance	0.7598	0.7423	0.7317	0.5948
speaker	0.9867	0.9900	0.9833	0.9767
VCTK				
utterance	0.7692	0.7757	0.6276	0.7716
speaker	1	1	1	1