The First VoicePrivacy Attacker Challenge

Natalia Tomashenko[1*], Xiaoxiao Miao[2], Emmanuel Vincent[1*], Junichi Yamagishi[3]

[1]*Université de Lorraine, CNRS, Inria, Loria, F-54000* Nancy, France;

[2]*Singapore Institute of Technology*, Singapore; [3]*National Institute of Informatics*, Tokyo, Japan

natalia.tomashenko@inria.fr, xiaoxiao.miao@singaporetech.edu.sg, emmanuel.vincent@inria.fr, jyamagis@nii.ac.jp

Abstract—**The First VoicePrivacy Attacker Challenge is an ICASSP 2025 SP Grand Challenge which focuses on evaluating attacker systems against a set of voice anonymization systems submitted to the VoicePrivacy 2024 Challenge. Training, development, and evaluation datasets were provided along with a baseline attacker. Participants developed their attacker systems in the form of automatic speaker verification systems and submitted their scores on the development and evaluation data. The best attacker systems reduced the equal error rate (EER) by 25–44% relative w.r.t. the baseline.**

Index Terms—**Voice privacy, voice anonymization, attacker system, automatic speaker verification**

I. CONTEXT

Speech conveys a lot of personal data, e.g., age and gender, health, geographical or ethnic origin, and socio-economic status. Formed in 2020, the VoicePrivacy initiative [1] promotes privacy enhancing solutions for speech technology via a series of benchmarking challenges. Privacy preservation is formulated as a game between *users* who process their utterances (referred to as *trial* utterances) with a privacy enhancing system prior to sharing with others, and *attackers* who access these processed utterances and wish to infer information about the users. The level of privacy offered by a given solution is measured as the lowest error rate among all attackers.

The first three VoicePrivacy Challenge editions [1], [2] focused on improving voice anonymization systems. In particular, the systems submitted to the VoicePrivacy 2024 Challenge had to: (a) output a speech waveform; (b) conceal speaker identity at the *utterance level*; (c) not distort linguistic and emotional content. The processed utterances sound as if they were uttered by another *pseudo-speaker*, which is selected independently for every utterance and can be an artificial voice not matching any real speaker. A *semi-informed attack model* was assumed, whereby attackers have access to the voice anonymization system and seek to re-identify the original speaker behind each anonymized trial utterance. Specifically, an ECAPA-TDNN automatic speaker verification (ASV) system was trained by the participants on data anonymized using their anonymization system. While this attack model is undeniably the most realistic to date, the provided attacker system is not its strongest possible implementation as it does not exploit spoken content similarities, specific pseudo-speaker selection strategies, or stronger ASV architectures, among others. To ensure a fair and reliable privacy assessment, it is essential to find the strongest possible attacker against every anonymization system. Hence, the current challenge edition takes

the attacker's perspective and focuses on the development of attacker systems against voice anonymization systems [3].

II. TASK

Participants were required to develop one or more attacker systems against one or more voice anonymization systems selected among three VoicePrivacy 2024 Challenge baselines [2] and four systems developed by the VoicePrivacy 2024 Challenge participants. For each speaker of interest, the attacker is assumed to have access to one or more utterances spoken by that speaker, which are referred to as *enrollment* utterances. The attacker system shall output an ASV score for every given pair of trial utterance and enrollment speaker, where higher (resp., lower) scores correspond to same-speaker (resp., different-speaker) pairs.

To develop and evaluate their attacker system against a given voice anonymization system, in line with the assumed semi-informed attack model, participants had access to: (1) anonymized trial utterances; (2) original and anonymized enrollment utterances; (3) original and anonymized training data (as well as other publicly available training resources specified in Section III) for the ASV system; (4) a description of the voice anonymization system; (5) the software implementation of that system when available.

III. DATA

The datasets are presented in Table I.

TABLE I
NUMBER OF SPEAKERS AND UTTERANCES IN THE ATTACKER TRAINING, DEVELOPMENT, AND EVALUATION SETS.

Subset		Female	Male	Total	#Utter.
Train	LibriSpeech: train-clean-360	439	482	921	104,014
Dev	LibriSpeech dev-clean	Enrollment	15	14	29	343
		Trial	20	20	40	1,978
Eval	LibriSpeech test-clean	Enrollment	16	13	29	438
		Trial	20	20	40	1,496

Training resources. The training set is the *train-clean-360* subset of *LibriSpeech*. In addition, participants were allowed to propose other training resources such as speech corpora and pretrained models before the deadline. Based on these suggestions, the final list of training resources was published in the evaluation plan [3].

Development and evaluation data. The development and evaluation sets comprise *LibriSpeech dev-clean* and *test-clean*.

Voice anonymization systems. The voice anonymization systems to be attacked include three baseline systems (**B3**, **B4**, and **B5**) [2], [3] and four selected systems developed by the VoicePrivacy 2024 Challenge participants (**T8-5**, **T10-2**, **T12-5**, and **T25-1**):

- **B3** – based on phonetic transcription, pitch and energy modification, and artificial pseudo-speaker embedding generation.
- **B4** – based on neural audio codec language modeling.
- **B5** – based on vector quantized bottleneck (VQ-BN) features extracted from an ASR model and on original pitch.

- **T8-5** [4] – random selection of one of two methods for each utterance (with probability p for the second method): (1) a cascaded ASR-TTS system with *Whisper* for ASR and *VITS* for TTS and (2) a k-nearest neighbor (kNN) voice conversion (VC) system operating on *WavLM* features.
- **T10-2** [5] – neural audio codec, with specific disentanglement of linguistic content, speaker identity and emotional state.
- **T12-5** [6] – based on **B5**, with additional pitch smoothing.
- **T25-1** [7] – disentanglement of content (VQ-BN as in **B5**) and style (global style token (GST) features and emotion transfer from target speaker utterances.

The code of **B3**, **B4**, and **B5** is available and could be used to develop attacker systems by, e.g., generating different or additional training data to train those systems.

IV. EVALUATION METRIC

We use the equal error rate (EER) metric to evaluate the attacker's performance. This metric has been used in all VoicePrivacy Challenge editions. The lower this metric, the stronger the attacker. The number of same-speaker and different-speaker trials in the development and evaluation datasets is given in Table II. The attackers were ranked separately for each voice anonymization system.

TABLE II
NUMBER OF SPEAKER VERIFICATION TRIALS.

Subset		Trials	Female	Male	Total
Dev	LibriSpeech dev-clean	Same-speaker	704	644	1,348
		Different-speaker	14,566	12,796	27,362
Eval	LibriSpeech test-clean	Same-speaker	548	449	997
		Different-speaker	11,196	9,457	20,653

V. BASELINE ATTACKER SYSTEM

As a baseline, we consider the attacker system used in the VoicePrivacy 2024 Challenge [2] (see Fig. 1). The ASV system (denoted $ASV_{\text{eval}}^{\text{anon}}$) is an ECAPA-TDNN with 512 channels in the convolution frame layers, implemented by adapting the *SpeechBrain VoxCeleb* recipe to *LibriSpeech*, and it is trained on anonymized training data. For a given trial utterance and enrollment speaker, the attacker computes the average speaker embedding of all anonymized enrollment utterances from that speaker and compares it to the speaker embedding of the anonymized trial utterance.
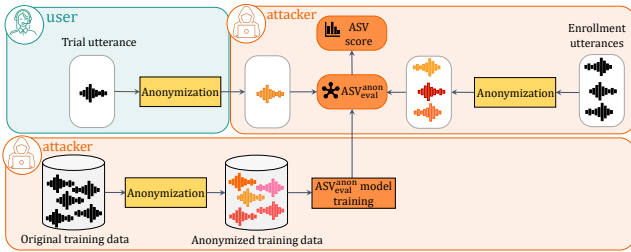

Fig. 1. Baseline attacker: training $ASV_{\text{eval}}^{\text{anon}}$ on anonymized training data and using it to compare anonymized trial and enrollment data.

VI. CHALLENGE RESULTS AND CONCLUSIONS

The challenge attracted 41 registered teams from academia and industry in 11 countries. Among them, 11 teams successfully submitted their results (55 submissions for 7 anonymization systems), which are summarized in Fig. 2. Many attackers significantly outperform the baseline attacker. The best ones reduce EER by 7–18% absolute (25–44% relative) for different anonymization systems[1].

---

[1]The attacker **A.42-2*** uses only textual data for training. It does not comply with the challenge rules due to the use of an undeclared BERT model.

The best attacker was developed by team **A.5** [8] for anonymization system **T8.5** and by team **A.20** [9] for every other anonymization system. **A.20** adapts a pretrained *ResNet34* ASV model from *WeSpeaker* using the *LoRA* technique on the provided anonymized data. **A.5** for **T8-5** proposes to use a so-called *ECAPA-PLDA-Mix* model, which combines an *ECAPA-TDNN* feature extractor trained on mixed datasets with a *PLDA*-based scoring module trained on anonymized data, and *SpecAugment* data augmentation. Other successful attackers' strategies include, among others, using a proposed *SpecWav* attack based on the *wav2vec2.0* feature extractor and spectrogram resizing (**A.41**) [10]; fine-tuning the *TitaNet-Large* model on anonymized data (**A.22-2**) [11]; using alternative distance metrics and voice *kNN-VC*-based voice normalization (**A.1**) [12]. The findings of the Attacker Challenge reveal that the privacy protection offered by the best anonymization systems from the VoicePrivacy 2024 Challenge was overestimated. They also show that, while attackers have made great progress in reducing the baseline EERs, the best anonymization methods still provide moderate protection against speaker re-identification (EER > 25%). A paper with a more detailed analysis of the results will be published in the future.
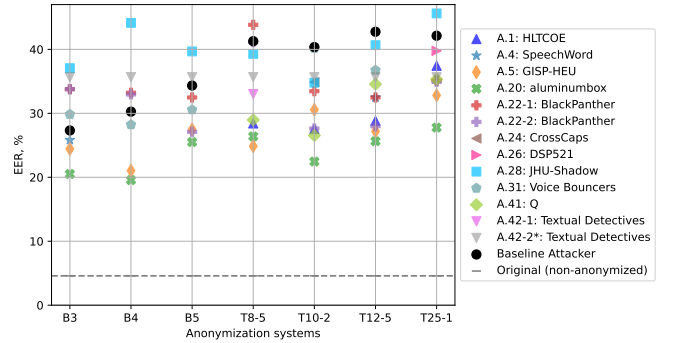

Fig. 2. Challenge results on the evaluation dataset.

REFERENCES

[1] Natalia Tomashenko et al., "The VoicePrivacy 2020 Challenge: Results and findings," *Computer Speech and Language*, vol. 74, 2022.

[2] Natalia Tomashenko et al., "The VoicePrivacy 2024 challenge evaluation plan," *arXiv preprint arXiv:2404.02677*, 2024.

[3] Natalia Tomashenko et al., "The First VoicePrivacy Attacker Challenge evaluation plan," 2024.

[4] Henry Li Xinyuan et al., "HLTCOE JHU submission to the Voice Privacy challenge 2024," in *SPSC 2024*, 2024.

[5] Jixun Yao et al., "NPU-NTU System for Voice Privacy 2024 Challenge," *arXiv preprint arXiv:2409.04173*, 2024.

[6] Nikita Kuzmin et al., "NTU-NPU System for Voice Privacy 2024 Challenge," *SPSC 2024*, 2024.

[7] Wenju Gu et al., "USTC-PolyU system for the VoicePrivacy 2024 Challenge," *SPSC 2024*, 2024.

[8] Yanzhe Zhang et al., "GISP-HEU's submission for VoicePrivacy Attacker Challenge at ICASSP 2025," 2024.

[9] Xiang Lyu et al., "Fast adaptation of pretrained speaker verification system for source speaker tracking," 2024.

[10] Yuqi Li et al., "SpecWav-Attack: Leveraging spectrogram resizing and wav2vec 2.0 for attacking anonymized speech," 2024.

[11] Candy Olivia Mawalim et al., "Fine-tuning TitaNet-Large model for speaker anonymization attacker systems," 2024.

[12] Henry Li Xinyuan et al., "HLTCOE submission to the VoicePrivacy Attacker challenge," 2024.