Privacy Attacks on Voice Anonymization Systems: Overview and Key Findings from the First VoicePrivacy Attacker Challenge

Natalia Tomashenko, Xiaoxiao Miao, Emmanuel Vincent, Junichi Yamagishi

▶ **To cite this version:**

Natalia Tomashenko, Xiaoxiao Miao, Emmanuel Vincent, Junichi Yamagishi. Privacy Attacks on Voice Anonymization Systems: Overview and Key Findings from the First VoicePrivacy Attacker Challenge. Computer Speech and Language, In press, pp.102036. ⟨10.1016/j.csl.2026.102036⟩. ⟨hal-05543730v2⟩

HAL Id: hal-05543730

https://hal.science/hal-05543730v2

Submitted on 23 Jul 2026

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons CC BY-NC-ND 4.0 - Attribution - Non-commercial use - No Derivative Works - International License

Privacy Attacks on Voice Anonymization Systems: Overview and Key Findings from the First VoicePrivacy Attacker Challenge

Natalia Tomashenko[a,*], Xiaoxiao Miao[b], Emmanuel Vincent[a] and Junichi Yamagishi[c]

[a]Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France
[b]Duke Kunshan University, Kunshan, 215316, China
[c]National Institute of Informatics, Chiyoda-ku, Tokyo, 101-8340, Japan

ARTICLE INFO

Keywords:
Privacy
Anonymization
Attack model
Speaker verification
Privacy metrics
Attacker challenge

Abstract

The First VoicePrivacy Attacker Challenge is an ICASSP 2025 SP Grand Challenge that 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. 11 teams submitted 55 attackers targeting 7 anonymization systems. The best attacker systems reduced the equal error rate (EER) by 25–44% relative to the baseline. Beyond global EER, analyses based on linkability and speaker-level risk show substantial variability in privacy protection across speakers and changes in anonymization system rankings under stronger attacks.

1. Challenge context and motivation

Speech carries sensitive personal information encompassing demographic characteristics (age, gender), health indicators, geographical and ethnic origin, and socioeconomic status. To address growing privacy concerns in speech technology, the VoicePrivacy initiative, formed in 2020, has developed a series of benchmarking challenges that systematically advance privacy-enhancing solutions (Tomashenko et al., 2022, 2020). 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 attempt to infer sensitive personal information from the processed utterances. The level of privacy offered by a given solution is measured as the lowest speaker recognition error rate among all attackers.

The first three VoicePrivacy Challenge (VPC) editions (Tomashenko et al., 2022, 2024a, 2026a; Panariello et al., 2024b) focused on improving voice anonymization systems. In particular, the systems submitted to the VoicePrivacy 2024 Challenge (VPC 2024) 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.

The challenge adopted a *semi-informed* attack model that reflects realistic threat scenarios in deployed systems. Under this threat model, attackers have access to the voice anonymization system and seek to re-identify the original speaker behind each anonymized trial utterance. Specifically, an Emphasized Channel Attention, Propagation and Aggregation Time-Delay Neural Network (ECAPA–TDNN) automatic speaker verification (ASV) system was trained by the participants on data anonymized using their anonymization system. Although this attack model was the most realistic considered to date, the provided attacker system is not its strongest possible instantiation, as it does not exploit similarities in spoken content, specific pseudo-speaker selection strategies, or more powerful ASV architectures, among other factors.

*Corresponding author
natalia.tomashenko@inria.fr (N. Tomashenko); xiaoxiao.miao@dukekunshan.edu.cn (X. Miao); emmanuel.vincent@inria.fr (E. Vincent); jyamagis@nii.ac.jp (J. Yamagishi)
https://sites.google.com/view/natalia-tomashenko (N. Tomashenko); https://xiaoxiaomiao323.github.io/ (X. Miao); https://members.loria.fr/EVincent/ (E. Vincent); https://yamagishilab.jp/ (J. Yamagishi)
ORCID(s): 0000-0002-7125-2382 (N. Tomashenko); 0000-0002-6645-6524 (X. Miao); 0000-0002-0183-7289 (E. Vincent); 0000-0003-2752-3955 (J. Yamagishi)

Table 1
Number of speakers and utterances in the attacker training, development, and evaluation sets

Subset		Type	#Speakers			#Utterances
			Female	Male	Total	
Training	LibriSpeech train-clean-360		439	482	921	104,014
Development	LibriSpeech dev-clean	Enrollment	15	14	29	343
		Trial	20	20	40*	1,978
Evaluation	LibriSpeech test-clean	Enrollment	16	13	29	438
		Trial	20	20	40*	1,496

*Includes 11 speakers who appear only in *different-speaker* trials.

Robust and trustworthy privacy evaluation requires identifying the most powerful possible attacker against each anonymization system. This need motivated a strategic shift in the VoicePrivacy Challenge's research focus. Rather than concentrating solely on the design of anonymization systems, the current challenge edition (Tomashenko et al., 2024b, 2025a) adopts an attacker perspective, explicitly focusing on developing attack systems that can expose vulnerabilities in voice anonymization approaches.

Instead of standardizing the anonymization pipeline and lightly constraining the attacker, we fixed a set of seven anonymization systems: three baselines and four top-performing submissions from VPC 2024 and invited participants to develop the strongest possible attacks against them under a *semi-informed* threat model concept. This setup enables a systematic study of (i) how far attacker performance can be advanced with current technology, (ii) how different anonymization approaches behave under strong and diverse attackers, and (iii) how privacy, utility, and fairness conclusions change once attacker strength is no longer artificially limited.

2. 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 (Tomashenko et al., 2024a) 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 3) for the ASV system; (4) a description of the voice anonymization system; (5) the software implementation of that system when available.

3. Data

3.1. Training, development, and evaluation resources

For each voice anonymization system, participants were provided with training, development, and evaluation data anonymized using that system. A detailed description of these datasets is presented below and in Table 1. The training set is the *train-clean-360* subset of the *LibriSpeech* (Panayotov et al., 2015) corpus. Besides the provided anonymized training data, participants were allowed to use the original *train-clean-360* data.

In addition, participants were allowed to propose other resources such as speech corpora and pretrained models before the deadline. Based on the suggestions received from the challenge participants, in this version of the evaluation plan, we published the final list of training data and pretrained models allowed for training attacker systems in the evaluation plan Tomashenko et al. (2024b). All the allowed resources are listed in Table 5 (Appendix A). From the provided list of resources, only a subset was eventually used by the teams to build their attack models. This subset includes the following pretrained models: *WavLM* (Chen et al., 2022), *ECAPA-TDNN* (Desplanques et al.,

2020), *wav2vec 2.0* (Baevski et al., 2020), *Encodec* (Défossez et al., 2023), *ResNet34*, *NVIDIA TitaNet-Large (en-US)* (Koluguri et al., 2022), and two undeclared models: *Whisper X* (Bain et al., 2023) and *BERT Base* (Devlin et al., 2019). Additional training datasets include the *train-other-500 LibriSpeech* subset and *VoxCeleb-1,2* (Chung et al., 2018). The toolkits used by participants comprise *WeSpeaker* (Wang et al., 2023a) and *SpeechBrain* (Ravanelli et al., 2021), and *PESTO* (Riou et al., 2023) for pitch estimation, with the latter being undeclared before the deadline.

The development and evaluation data sets comprise *LibriSpeech dev-clean* and *LibriSpeech test-clean*. Besides the provided anonymized enrollment data, the participants were allowed to use the original enrollment data.

3.2. Voice anonymization systems

The voice anonymization systems to be attacked include three baseline systems (**B3**, **B4**, and **B5**) and four selected systems developed by the VoicePrivacy 2024 Challenge participants (**T8-5**, **T10-2**, **T12-5**, and **T25-1**) (Tomashenko et al., 2026a, 2024a,b). The participants' systems were chosen according to their anonymization performance in the highest privacy category (EER $\geq$ 40%), excluding cascaded anonymization systems based on automatic speech recognition (ASR) followed by text-to-speech (TTS). Thus, a total of seven systems are to be attacked:

- **B3** – based on phonetic transcription, pitch and energy modification, and artificial pseudo-speaker embedding generation (Meyer et al., 2023; Tomashenko et al., 2024a).
- **B4** – based on neural audio codec language modeling (Panariello et al., 2024a; Tomashenko et al., 2024a).
- **B5** – based on vector quantized bottleneck (VQ-BN) features extracted from an ASR model and on original pitch (Champion, 2023; Tomashenko et al., 2024a).
- **T8-5** (team *JHU-CLSP*, system *"Admixture ($p = 0.4$)"* (Xinyuan et al., 2024)) – 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* (Radford et al., 2023) for ASR and *VITS* (Kim et al., 2021) for TTS and (2) a k-nearest neighbor (kNN) voice conversion (VC) system operating on *WavLM* (Chen et al., 2022) features.
- **T10-2** (team *NPU-NTU*, system *"C4"* (Yao et al., 2024)) – neural audio codec, with a specific disentanglement strategy for linguistic content, speaker identity, and emotional state.
- **T12-5** (team *NTU-NPU*, system *"3"* (Kuzmin et al., 2024)) – based on **B5**, with additional pitch smoothing.
- **T25-1** (team *USTC-PolyU*, system *"large: ESD+LibriTTS"* (Gu et al., 2024)) – disentanglement of content (VQ-BN as in **B5**) and style (global style token (GST) (Wang et al., 2018)) features and emotion transfer from target speaker utterances.

The code of **B3**, **B4**, and **B5** is available on GitHub[1] and could be used to develop attacker systems by, e.g., generating different or additional training data to train those systems.

4. 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. For every given pair of trial utterance and enrollment speaker, the attacker system outputs an ASV score from which a same-speaker vs. different-speaker decision is made by thresholding. Denoting by $P_{\text{fa}}(\theta)$ and $P_{\text{miss}}(\theta)$ the false alarm and miss rates at threshold θ, the EER metric corresponds to the threshold θ_{EER} at which the two detection error rates are equal, i.e., $\text{EER} = P_{\text{fa}}(\theta_{\text{EER}}) = P_{\text{miss}}(\theta_{\text{EER}})$. Lower EER values correspond to stronger attackers. The number of same-speaker and different-speaker trials in the development and evaluation datasets is given in Table 2. The attackers will be ranked separately for each voice anonymization system.

[1] https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2024

Table 2
Number of *same-speaker* and *different-speaker* trials considered for evaluation.

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

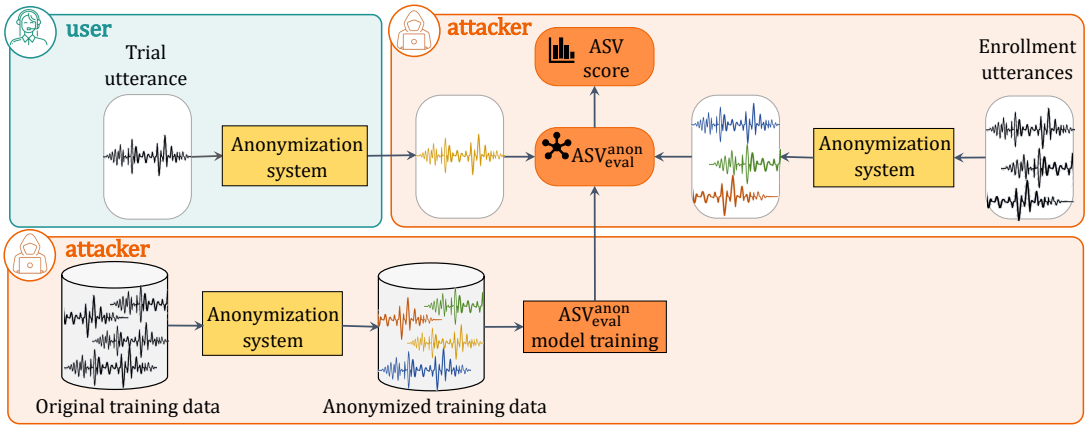


Figure 1: Baseline attacker: training $ASV_{\text{eval}}^{\text{anon}}$ on anonymized training data and using it to compare anonymized trial and enrollment data.

Table 3: Participating teams in the First VoicePrivacy Attacker Challenge

ID	Team Name	Affiliation	Country	Reference
A.1	HLTCOE	Human Language Technology Center of Excellence, Johns Hopkins University	USA	Xinyuan et al. (2025)
A.4	SpeechWorld	Vingroup Big Data Institute – VinBigData	Vietnam	Nhan Tri Do (2025)
A.5	GISP-HEU	Harbin Engineering University; Harbin Institute of Technology; University of Technology Sydney	China, Australia	Zhang et al. (2025a)
A.20	aluminumbox	Alibaba Inc., Shanghai	China	Lyu et al. (2025)
A.22	BlackPanther	Japan Advanced Institute of Science and Technology, Ishikawa	Japan	Mawalim et al. (2025)
A.24	Cross-Caps	IIIT Delhi	India	Barneet Singh (2025)
A.26	DSP521	National Chung Cheng University, Chiayi	Taiwan	Chen and Tsai (2025)
A.28	JHU-Shadow	Johns Hopkins University; University of Southern California	USA	Thebaud et al. (2025)
A.31	Voice Bouncers	Samsung R&D Institute Poland	Poland	Krzywdziak et al. (2025)
A.41	Q	Qifu Technology; Fudan University, Shanghai	China	Li et al. (2025)

Table 3: Participating teams in the First VoicePrivacy Attacker Challenge

ID	Team Name	Affiliation	Country	Reference
A.42	Textual Detectives	Friedrich-Alexander University of Erlangen-Nuremberg; Fraunhofer IIS; Trinity College Dublin	Germany, Ireland	Ünal Ege Gaznepoglu et al. (2025a)

5. Baseline attacker system

As a baseline, we consider the attacker system used in the VoicePrivacy 2024 Challenge (Tomashenko et al., 2024a) (see Figure 1). The ASV system (denoted $ASV_{\text{eval}}^{\text{anon}}$) is an ECAPA-TDNN (time-delay neural network with emphasized channel attention, propagation, and aggregation architecture (Desplanques et al., 2020)) with 512 channels in the convolution frame layers. It is implemented by adapting the *SpeechBrain* (Ravanelli et al., 2021) *VoxCeleb* recipe to *LibriSpeech* and 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[2]. The baseline attacker model configuration was kept fixed across all seven anonymization systems, as the optimal training hyperparameters depend on the specific anonymization algorithm and cannot be determined in advance under a semi-informed threat model.

6. Submitted attacker systems

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). The teams are listed in Table 3. The corresponding attacker systems are summarized in Table 4.

Team A.1

The team **A.1** (Xinyuan et al., 2025) proposes enhanced attack models building on a baseline ECAPA-TDNN classifier trained for speaker verification. To address poor convergence in the baseline, the authors retrain the model with extended epochs (20-30), reduced learning rates (LR=10^{-3}), and dropout applied to input features, improving classification robustness.

For ASV, the authors replace the standard comparison method, which uses cosine distance between an average speaker embedding computed from all enrollment utterances and speaker embeddings computed for trial utterances, with several alternatives. They argue that this standard metric may be vulnerable to multi-modal distributions, a phenomenon observable in systems such as **T8-5**. The proposed alternatives include: (1) max similarity scores, which ignore outlier enrollment samples; (2) the percentage of enrollment utterances exceeding a similarity threshold, classifying as target when this percentage exceeds 50%; and (3) probabilistic linear discriminant analysis (PLDA) scoring.

Additionally, **A.1** introduces kNN-VC normalization, which applies nearest-neighbor voice conversion to map all anonymized utterances back to a fixed proxy speaker before performing ASV training. This approach aims to eliminate target-specific confounding factors introduced by conditional anonymization systems.

These methods are combined adaptively per target anonymizer, with the strongest variant selected based on development set performance.

Team A.4

The team **A.4** (Nhan Tri Do, 2025) developed three attacker systems designed to attack different voice anonymization approaches using reconstruction-based methods:

*Attack on **B3*** processes anonymized audio through a pre-trained *WavLM Large* model to extract 1024-dimensional embeddings. A Multi-Head Factorized Attentive (MHFA) module reconstructs the original speaker embedding from

[2]The code to train the baseline attacker systems is available on GitHub: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2024

the stacked *WavLM* outputs. In parallel to this, prosody information (pitch and energy) is extracted and processed through Gradient Reversal Layers (GRL) to remove the influence of randomized prosodic patterns. The model is trained with Mean Squared Error (MSE) loss for embedding reconstruction, contrastive loss for speaker discrimination, and Additive Angular Margin Softmax (AAM-Softmax) for speaker classification.

*Attack on **B4*** targets neural audio codec (NAC) anonymization by reconstructing the original audio at the codec level rather than the embedding level. Both the original and anonymized audio are passed through the NAC encoder to extract discrete acoustic tokens. A Reconstruction Layer, composed of 4 stacked Retentive Network blocks, learns to map anonymized codecs to original codecs. The reconstructed codecs are then decoded to generate the original audio, from which speaker embeddings are extracted for verification using *WavLM*-based ECAPA and cosine similarity.

*Attack on **T12-5*** uses the same core architecture as the attack on **B3**.

The fundamental difference between these systems reflects their targets: **B3** and **T12-5** operate in the embedding space exploiting prosody modifications, while **B4** operates at the acoustic token level attempting to reconstruct the original codec representation.

Team A.5

The team **A.5** (Zhang et al., 2025a,b) developed two types of attack models: *ECAPA-PLDA-Mix* attack targeting **B3, B4, B5, T8-5, T12-5, T25-1** and *TitaNet-Cosine* attack for **T10-2**.

The *ECAPA-PLDA-Mix* combines an ECAPA-TDNN frontend with PLDA backend, where 16 kHz audio splits into 3-second chunks, converted to 80-dim mel-spectrograms with *utterance-level* mean normalization per VPC 2024 settings, then fed through a 921-dim linear layer optimized via AAM-Softmax loss in *SpeechBrain*. For vanilla use (**B5**, **T12-5**, **T25-1**), the front-end trains 10 epochs on mixed *LibriSpeech train-clean-360* original and anonymized data while PLDA trains solely on anonymized data; **T8-5** adds SpecAugment (Park et al., 2019) post-spectrogram (2 masks: $t_{mask} \sim U(0, 64)$, $f_{mask} \sim U(0, 8)$); and **B3/B4** employ a three-stage process where stage 1 follows vanilla training, stage 2 trains a separate ECAPA-TDNN on original train-clean-360 for 10 epochs to produce *utterance-level* to *speaker-level* embeddings, and stage 3 fine-tunes the front-end for 4 epochs on anonymized data using contrastive loss with positive/negative pairs from those *speaker-level* embeddings.

The *TitaNet-Cosine* attack targets only **T10-2** due to domain mismatch: *speaker-level* anonymization in train/enrollment vs. *utterance-level* in trials, making fine-tuning ineffective. It uses a pre-trained *TitaNet-Large* as the feature extractor and cosine similarity for scoring to improve performance.

Table 4: Summary of submitted attacker systems

ID	Attacker System	Anonymization Systems						
		B3	**B4**	**B5**	**T8-5**	**T10-2**	**T12-5**	**T25-1**
A.1	Improved ECAPA-TDNN classifier (20–30 epochs, dropout); alternative ASV scoring metrics (max similarity, PLDA); optional kNN-VC normalization to a fixed proxy speaker applied before ASV training The best combination is selected per target anonymization system based on development set performance.	-	-	-	∘	+	+	+
A.4	Reconstruction of original features with system-specific approaches: embedding-level (MHFA, *WavLM*-based ECAPA, GRL for prosody handling); codec-level (**B4**).	+	+	-	-	-	+	-
A.5	Data augmentation enhanced feature representations, speaker identity difference enhanced classifier, PLDA trained on anonymized data; data fusion and SpecAugment. For **T10-2**: *TitaNet-Large* with cosine similarity.	+	+	+	★	+	+	+
A.20	Adapting the ResNet34 ASV model pretrained on *VoxCeleb-2* from *WeSpeaker* using LoRA and *WavLM-Large* features on anonymized data.	★	★	★	+	★	★	★

Table 4: Summary of submitted attacker systems

ID	Attacker System	Anonymization Systems						
		B3	**B4**	**B5**	**T8-5**	**T10-2**	**T12-5**	**T25-1**
A.22	Fine-tuning *TitaNet-Large* model (final decoder layer) on anonymized data; speed perturbation; mixing anonymized data from multiple anonymization systems: **B5**, **B5**+**T12-5**; all.	+	+	+	+	+	+	+
A.24	Using *WavLM Base* model with multi-head factorized attention.	-	-	-	-	○	-	-
A.26	Employing CQT for spectral analysis, MCG-Res2Net-RH model with multi-scale branches and refined highway mechanism, GE2E loss.	-	-	-	-	-	-	+
A.28	Multi-stream de-anonymization framework combining x-vectors (*fw-SEResNet34*), prosody embeddings (*WhisperX* character-level alignments), and pitch statistics (*PESTO*). Different attack configurations are submitted per anonymization system: x-vectors are the main stream for all systems except **B4**, for which prosody features were used.	○	○	○	○	○	○	○
A.31	Individual anonymization models using: knowledge distillation, audio perceptual hashing, and data mixing augmentation.	+	+	+	-	+	+	-
A.41	Proposed SpecWav attack based on *wav2vec 2.0* feature extractor and spectrogram resizing.	-	-	-	+	+	+	+
A.42	Text-based attack using *BERT Base* fine-tuned on *train-clean-360*; cherry-picking attack for **T8-5**.	○	○	○	+	○	○	○

Notation	Description
-	Attack not submitted for this anonymization system
○	Submitted systems excluded from the official ranking: **A.1** – attack on their own anonymization system (**T8.5**); **A.28**, **A.42** – use of undeclared training data; **A.24** – incomplete system description
+	Submitted systems included in the official ranking
★	Best results in the official ranking
	Top five teams

Team A.20

The **A.20** attack model (Lyu et al., 2025) introduces a fast-adaptive ASV system initializing a ResNet34 backbone pretrained on *VoxCeleb-2* and enhancing it with Low-Rank Adaptation (LoRA (Hu et al., 2021)) modules at each Conv2d layer, augmented by re-parametrization for increased capacity, specifically to handle domain shifts in anonymized speech while leaving original audio unaffected. Pretrained *WavLM-large* features are integrated exclusively for anonymized inputs through residual connections in ResBlocks, providing richer representations to counter limited training data. A three-stage training strategy mitigates overfitting: (i) a *freeze* stage where the backbone is fixed and only the projection layer is trained for 10 epochs on *VoxCeleb-2 dev* and *LibriSpeech train* (original) data; (ii) a *pretrain* stage where the full model is trained for 70 epochs on mixed real-world and anonymized data; (iii) a per-system *finetune* stage of 10 additional epochs at learning rate 10^{-5} on each system's anonymized data, yielding

seven separate finetuned models. Final scoring uses cosine similarity with ASNorm and a top-$n = 300$ cohort from the anonymized training set, using the *WeSpeaker* toolkit [3].

Team A.22

The attacker systems proposed by Mawalim et al. (2025) leverage fine-tuned versions of the *TitaNet-Large* and ECAPA-TDNN models. The *TitaNet-Large*, featuring 25.3 million parameters, employs a ConvASR encoder-decoder architecture: the encoder processes normalized mel spectrograms using 1D depth-wise separable convolutions for local features and Squeeze-and-Excitation layers for global context, while the attentive statistics pooling decoder generates fixed-length t-vector embeddings via weighted channel statistics and linear layers for speaker verification scored by cosine similarity.

Fine-tuning focuses on the decoder layer using a 9:1 training-validation split of anonymized speech data, augmented with speed perturbation. This approach focuses on using **B5** and **T12-5** anonymized data and considers the following fine-tuning configurations: (1) only **B5**, (2) **B5** and **T12-5**, and (3) anonymized data from all systems (**B3, B4, B5, T8-5, T10-2, T12-5, T25-1**).

Team A.26

The team **A.26** (Chen and Tsai, 2025) proposes an attack system that processes speech signals using the Constant-Q Transform (CQT). This transform extracts fine-grained low-frequency features, such as fundamental frequency and formants, which are key to speaker identity, providing higher resolution in low frequencies compared to the traditional Fast Fourier Transform (FFT) and better alignment with human speech spectral distribution.

The core architecture (*MCG-Res2Net-1RH-E*) builds on Multi-group Channelwise Gated Res2Net (MCG-Res2Net (Chen et al., 2024a; Li et al., 2021)). It integrates multi-scale branches within residual blocks and Refined Highway (RH) connections to capture subtle speech features across varying receptive fields, improving distinction between speakers.

The Generalized End-to-End (GE2E) loss (Wan et al., 2018) serves as the cost function. It employs supervised contrastive learning to cluster embeddings by minimizing intra-speaker distances and maximizing inter-speaker distances via cosine similarity to centroids, optimizing identity verification on anonymized data.

Team A.28

The system **A.28** (Thebaud et al., 2025) proposes a multi-stream de-anonymization attack that fuses three complementary speaker identity features from anonymized speech. X-vectors are extracted from fine-tuned *fw-SEResNet34* models pretrained on *VoxCeleb-1,2*, then adapted using either all anonymized data or data corresponding to a target anonymization system. Prosody embeddings derive from *WhisperX* character-level alignments, tokenized and processed by a 2-layer transformer encoder with mean pooling to capture temporal dynamics, such as pauses and phoneme durations. Pitch features consist of a 8-dimensional vector of means and standard deviations from *PESTO*-estimated pitch trajectories, deltas, delta-deltas, and model confidence. These parallel streams exploit persistent leaks in anonymization systems, such as unmodified prosodic patterns or pitch dynamics. The features are integrated via a learnable weighted average (LDA-trained on development-set scores) to yield robust speaker verification performance across diverse anonymizers. Importantly, this team submits different attack configurations for different anonymization systems: for all systems except **B4**, the submitted attack relies primarily on the per-system fine-tuned x-vectors; the fusion of all three streams does not consistently outperform the best individual stream on anonymized data and is mainly beneficial on original (non-anonymized) speech. Per Table 2 of (Thebaud et al., 2025), pitch features alone are the weakest stream (EER $\approx$ 48–50% on most systems) and the score-level fusion even degrades performance relative to the x-vector-only configuration on several systems. This is therefore best understood as a proof-of-concept demonstrating that non-timbral features can be incorporated into a multi-stream attacker, rather than as evidence of their practical attack strength under the current architecture and training regime.

Team A.31

The team **A.31** (Krzywdziak et al., 2025) developed specialized attacker systems targeting anonymization methods **B3, B4, B5, T10-2,** and **T12-5** using ECAPA-TDNN models from *SpeechBrain* on the *LibriSpeech* dataset. Preprocessing aligns original and anonymized audio pairs by matching fragment lengths to account for speed changes,

[3] Full implementation details are available at `https://github.com/wenet-e2e/wespeaker`.

padding shorter segments with zeros, applying sentence-level normalization, chunking into 3-second segments, and using 80 mel coefficients.

Core methodologies include knowledge distillation, where a teacher model processes original audio to produce embeddings, and a student model on anonymized audio minimizes a combined AAM-Softmax loss (with *margin* 0.2, *scale* 30) plus scaled MSE between embeddings. Audio hashing enhances this by applying perceptual hashing to teacher embeddings for binary targets, enabling the student to match hashed features via sigmoid activation while preserving correlations. Data mixing progressively replaces portions of original audio with anonymized segments across epochs, with increments adapted to training length.

For **B3** and **T12-5**, attackers rely solely on data mixing over 20 epochs. **B4** and **T10-2** combine all three methods over 50 epochs, while **B5** uses distillation and hashing alone.

Team A.41

The team **A.41** proposed the *SpecWav*-attack (Li et al., 2025) that employs the ECAPA-TDNN speaker recognition architecture, replacing conventional fbank feature extraction, with 1024-dimensional *wav2vec 2.0* embeddings. This modification enables the model to capture deeper phonetic, contextual, and speaker-independent characteristics, enhancing sensitivity to subtle distortions introduced by anonymization techniques. The architecture incorporates components like attentive statistical pooling, batch normalization, Squeeze-Excitation (SE-)Res2Blocks, Conv1D layers with ReLU activation, and a final fully connected layer for speaker verification output.

A core innovation lies in the data augmentation pipeline, adapted from *FreeVC* (Li et al., 2023a) Spectrogram Resizing (SR) based method and applied to the *LibriSpeech train-clean-360* dataset during preprocessing. The process extracts Mel spectrograms from source waveforms, applies vertical spectral resampling (super-resolution upscaling via linear interpolation or downsampling with alignment, padding, silence, or noise addition), and reconstructs augmented waveforms using a neural vocoder, selectively distorting speaker-specific traits while preserving linguistic content. Training follows an incremental strategy: an initial phase of 10 epochs on clean *LibriSpeech train-clean-360* data, succeeded by 2-4 epochs on augmented anonymized variants (e.g., **T10-2**), promoting gradual adaptation and overfitting prevention.

Team A.42

The team **A.42** first analyzes the baseline attacker scores at *speaker-level* and observes that some speakers are consistently worse anonymized across different anonymization systems. They hypothesize that the attacker exploits similarity of linguistic content across enrollment and trial utterances for these vulnerable speakers, rather than only speaker characteristics and proposes a corresponding text-based attack.

*Text-based BERT attack **A.42-2****. To test the above hypothesis, they build an attacker that uses only text (transcriptions) and no audio, based on a *BERT Base* classifier with AAM-Softmax, fine-tuned on *LibriSpeech train-clean-360*.

*Cherry-picking attack **A.42-1** on **T8-5***. The team also develops an attacker against *AdMixture*-type systems, where anonymization systems are probabilistically mixed, as **T8-5**. They propose two modifications: (1) doubling the classifier output dimension and "*cherry-picking*" between paired units by taking, for each speaker, whichever logit is higher during training (a multi-label log-softmax wrapper); (2) replacing enrollment averaging by a 1-nearest-neighbor selection at test time: for each trial and candidate speaker, use the enrollment embedding that yields highest cosine similarity instead of mean pooling.

7. Results

A summary of the results for all attack models and anonymization systems on the test data in terms of EER (averaged across male and female speaker subsets) is presented in Figure 2. The black circles represent the EER results for the baseline attack models, while the colored markers correspond to the 55[4] attacks developed by the participants. More detailed EER results for the development and evaluation data are shown in Figure 3, and Figure 6 (see Appendix B.1) shows the EER values on the development and test datasets, separately for male and female speakers, reported with 95% confidence intervals calculated as suggested by Bengio and Mariéthoz (2004).

[4]In total, 60 colored marks include 55 different results from the submitted attack models (**A.42-2** provides the same result for all 7 anonymization systems; one submitted result, **A.4** on **B3** with EER > 50%, is not shown in the figure).

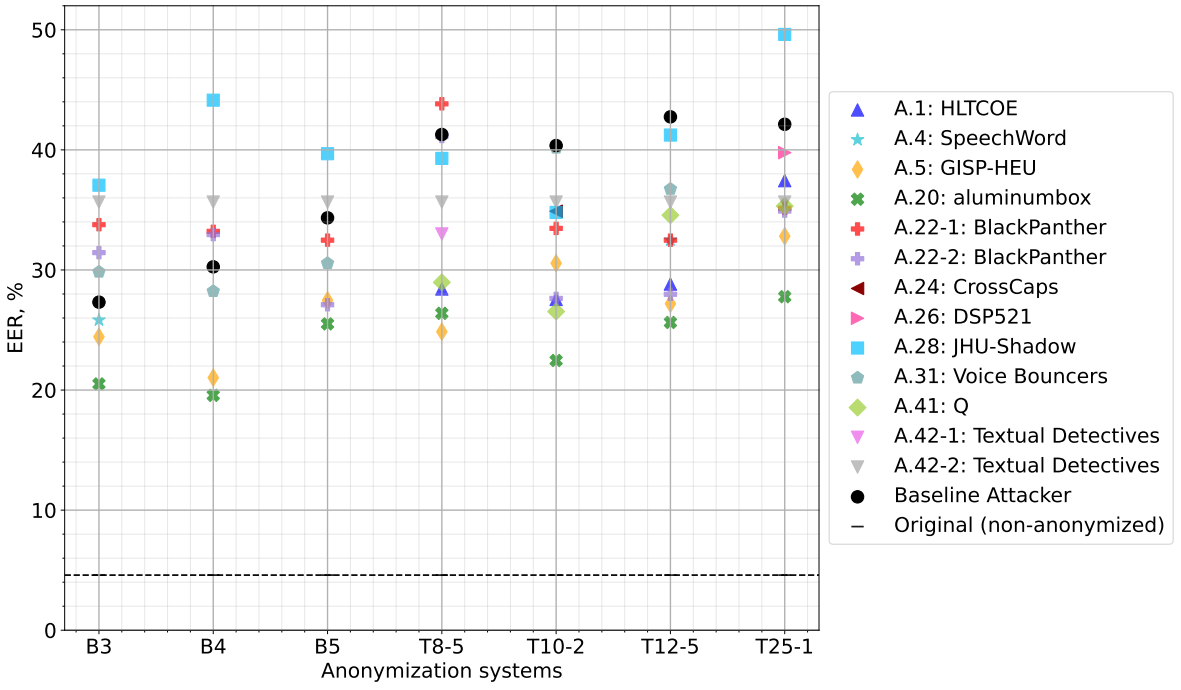


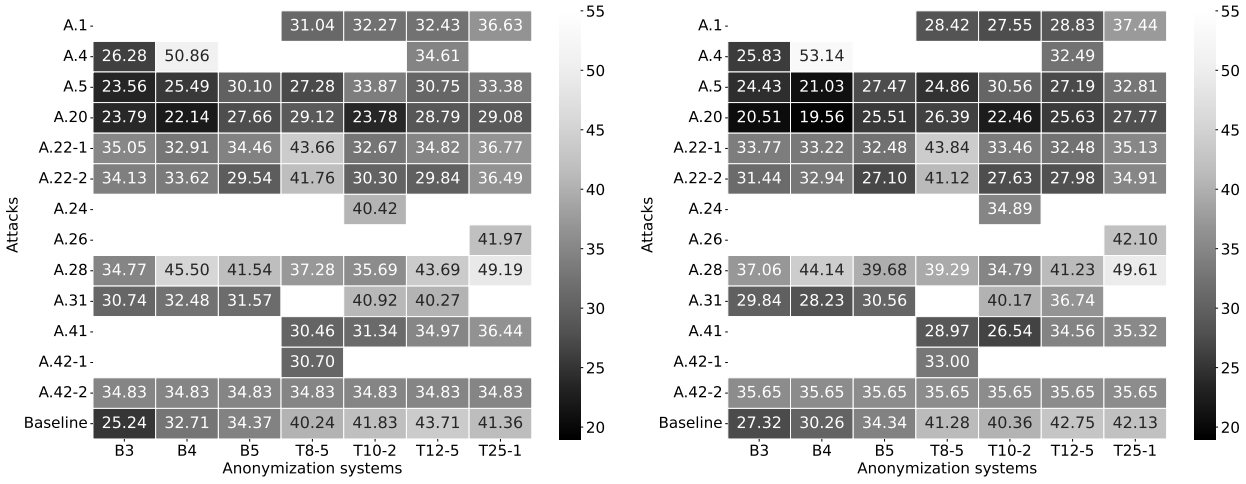
Figure 2: Challenge EER results on the test dataset.

Approximately 80% of the attack systems outperform the baseline attacks. The strongest attackers were developed by team **A.5** for anonymization system **T8-5** and by team **A.20** for all other systems. Team **A.20**'s LoRA-adapted *ResNet34* attacker from *WeSpeaker*, enhanced with *WavLM-Large* features on anonymized data, achieved top performance across six anonymization systems: **B3**, **B4**, **B5**, **T10-2**, **T12-5**, and **T25-1**. **A.5**'s *ECAPA-PLDA-Mix* attack model, which combines ECAPA-TDNN on mixed data, PLDA scoring on anonymized data, and SpecAugment, performed best on **T8-5**. As shown in Figure 4, the best attack systems reduce the EER on test by 7–18% absolute (25–44% relative), depending on the anonymization system. Figure 4 summarizes these absolute and relative EER reductions per anonymization system, making clear that gains are largest for **T10-2** and generally smaller for the VPC 2024 baseline anonymization systems than for the participant systems. The baseline attackers achieved significantly lower EERs on **B3**–**B5** (∼27–34%) than on VPC 2024 participant systems (∼40–43%). Top attackers narrowed this gap substantially. On average, EER improvements were smaller for VPC 2024 baseline anonymization systems than for participant systems.

The largest gain in best attacker performance among all anonymization systems was achieved for anonymization system **T10-2**. Team **A.5** also revealed non-compliance with the VPC 2024 rules for this system by using *speaker-level* rather than *utterance-level* anonymization on the training data, potentially weakening attacker performance. The most challenging anonymization systems to attack were **T25-1** and **T8-5**, while **B4** turned out to be the least resistant. These anonymization systems provide only moderate privacy protection: EER∼26–28%. Despite these results, best attacker performance remains substantially below that on original (non-anonymized) data.

Other effective attack strategies developed by the top-performing teams include: (**A.41**) a proposed *SpecWav* method built on the *wav2vec 2.0* feature extractor and spectrogram resizing; (**A.22-2**) fine-tuning the *TitaNet-Large* model on anonymized data; and (**A.1**) employing alternative distance metrics combined with *kNN-VC*-based voice normalization.

A change in the ranking of anonymization systems based on EER is observed when evaluated under the strongest attacker (**B4, B3, T10-2, T8-5, B5, T12-5, T25-1**) compared to the baseline attacker (**B3, B4, B5, T10-2, T8-5, T25-1, T12-5**), indicating that more advanced attack models can substantially alter the relative privacy performance of different systems. Moreover, when considering the overall privacy–utility evaluation and ranking of anonymization systems proposed in the official VPC 2024 setup (see Figure 7 in Appendix B.2), it can be concluded that all four



(a) Development (on original data: EER=5.72%)

(b) Test (on original data: EER=4.59%)

Figure 3: Challenge results (EER,%) on (a) development and (b) test datasets, averaged across female and male speakers.

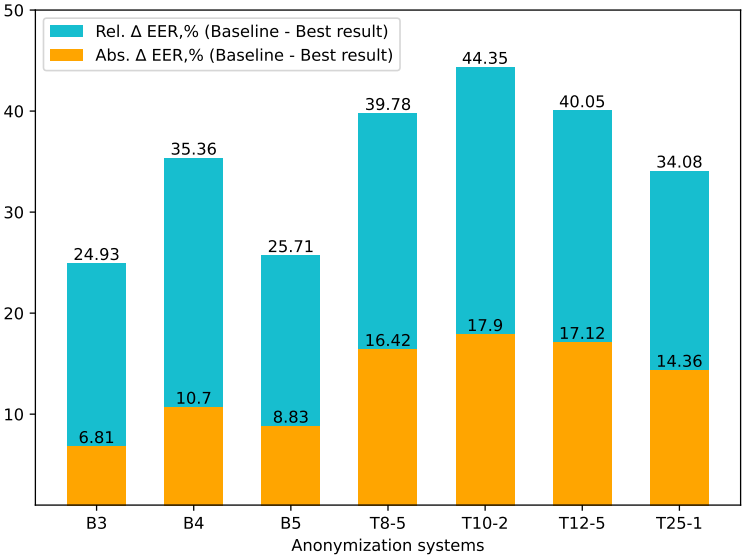


Figure 4: Relative and absolute improvements in ASV performance for the best attack models (**A.20** for **B3–B5, T10-2, T12-5, T25-1**, and **A.5** for **T8-5**) compared to the baseline on the test dataset.

VPC 2024 participants' anonymization systems consistently shift from the fourth privacy category (EER≥40%) to the second (EER≥20%), while preserving their mutual ranking. At the same time, a more global analysis that includes the baseline anonymization systems shows that the overall ranking of the anonymization systems changes when the strongest attack models are used for evaluation.

Figure 5 shows absolute difference in EER for female and male speakers on the test dataset. On original data, this gap is +8.34% absolute (EER=8.76% for female speakers; EER=0.42% for male speakers). The analysis of the performance gap on anonymized data reveals notable fairness concerns across the test set. However, the gap decreases after anonymization for nearly all systems, except for the **A.20** attack on **B3**. The mean gender gap is +1.07% with a standard deviation of 2.85%, indicating that females have, on average, higher EER than males.

The distribution of gender gaps ranges from a minimum of -6.28% (higher EER for male speakers, **A.28** attack on **T25-1**) to a maximum of +9.40% (higher EER for female speakers, **A.20** attack on **B3**), representing a total range of

15.68%. This indicates that while some system-attack combinations achieve near-perfect fairness, others demonstrate significant fairness failures.

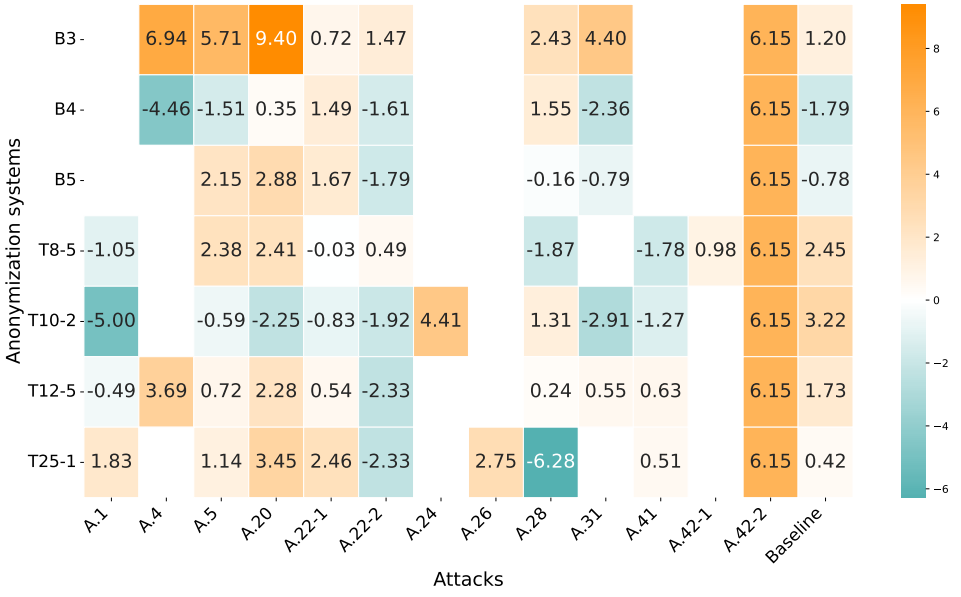


Figure 5: Difference in EER (abs.), % for female and male speakers on the test dataset.

8. Discussion and future directions

8.1. Limitations of EER-based privacy evaluation

While EER remains the standard metric in speaker verification and has been adopted across all VoicePrivacy Challenge editions, recent work highlights its limitations for privacy assessment. The EER can be insensitive to re-identification risks, remaining stable even when the probability of successful re-identification varies across users or use cases. It captures only average system performance at a single operating point, thereby masking worst-case scenarios and outlier vulnerabilities. This limitation can lead to an overly optimistic estimation of privacy protection, particularly for individuals who remain easily identifiable even after anonymization (Nautsch et al., 2020; Williams et al., 2024).

To better align privacy evaluation with both technical requirements and legal expectations, alternative frameworks have been proposed. One such framework, introduced by Vauquier et al. (2025), builds on the concepts of *linkability* and *singling out*, legally accepted interpretations of anonymity. Results derived from these metrics suggest that residual privacy risks after anonymization are often more substantial than implied by EER-based evaluations, underscoring the need for metrics that reflect both technical and legal perspectives. Appendix B.3 presents the challenge results in terms of the *linkability* metric. In particular, the summary *linkability* results on the test data in Figure 8 reveal that the ranking of several attacker systems differs from that obtained with the EER-based evaluation (see Figure 2). As detailed in the appendix, Figure 9 shows a strong correlation between EER and *linkability* scores despite ranking differences, suggesting that EER remains useful as a quick proxy but may be insufficient on its own for robust privacy assessment.

Interpreted through the legal lens of Vauquier et al. (2025), our challenge results reveal a more complex picture. Under the strongest attackers, the mean linkability of the best-performing anonymization systems remains substantially above chance (random-guess linkability $\approx 1/(|S|+1) \approx 0.077$ in our evaluation set-up ($|S| = 12$) for strong attack models, indicating that the average anonymized speaker can still be linked with non-trivial success. Critically, the per-speaker heterogeneity documented in Appendix B.4 means that a subset of speakers is consistently highly linkable across systems and attackers, providing these speakers with very weak anonymity guarantees This implies that systems previously classified in the highest VPC 2024 privacy category (EER$\geq$40% under the baseline attacker) can no longer claim that category under stronger attacks, and even at the lower EER values they achieve, their linkability remains substantially above chance — indicating that they would not satisfy GDPR-style anonymisation requirements under

Recital 26 and Article 29 Working Party Opinion 05/2014, where data are considered anonymous only when re-identification is not reasonably possible by reasonably likely means. For the VoicePrivacy community, these results have two concrete implications. First, challenge metrics should incorporate linkability-based thresholds aligned with legal notions of identifiability (e.g., a maximum acceptable linkability score), in addition to EER categories. Second, speaker-level worst-case linkability should be part of the official evaluation, as global averages mask the risk faced by the most vulnerable speakers. We plan to adopt such legally grounded evaluation criteria in future editions of the VoicePrivacy Challenge.

Other promising metrics include the *similarity rank disclosure* (SRD) measure introduced by Bäckström et al. (2025), which quantifies privacy in units of entropy (bits). SRD is defined as the reduction in entropy about a subject's true identity after observing their similarity rank when matching a query sample against a database of templates. By expressing how much personally identifiable information (PII) can be inferred from the similarity rank, the SRD metric offers a more intuitive and theoretically grounded means of evaluating privacy in voice anonymization.

8.2. Speaker-level privacy analysis

Anonymization effectiveness varies dramatically across individuals (Williams et al., 2024; Ünal Ege Gaznepoglu et al., 2025a; Noé et al., 2022). Analysis of per-speaker privacy risks using *linkability* across 68 attack-anonymization combinations (7 anonymization systems × 8–11 attack models per system, depending on the system), presented in Appendix B.4, reveals substantial heterogeneity in privacy protection. Figure 10 in Appendix B.4 presents per-speaker linkability distributions across all attack-anonymization combinations for male (Figure 10a) and female (Figure 10b) speakers. The between-speaker variance is significant: high-risk speakers show consistently elevated median linkability values (exceeding 0.6), while low-risk speakers maintain protection closer to chance-level performance (0.15–0.20) despite exposure to diverse attack strategies. This heterogeneity demonstrates that anonymization systems provide dramatically unequal protection across the speaker population.

Analysis of system-specific privacy risks and optimal anonymization system selection (in Appendix B.4) reveals that *no single anonymization system uniformly protects all speakers*. To quantify this heterogeneity, we identified for each speaker the anonymization system that minimizes linkability under worst-case attack conditions (the strongest attack targeting that speaker-system pair), as well as the system that maximizes vulnerability. Figure 11 in Appendix B.4 shows the distribution of optimal and worst-case system assignments across the speaker population. System **T25-1** represents the most robust anonymization approach, providing optimal protection for the largest number of test speakers (8 of 29, i.e. ≈28% of our test set). Conversely, system **B3** has the highest privacy risk, representing the worst-case system for 12 of 29 test speakers (≈41% of our test set). This speaker-dependent vulnerability pattern demonstrates that deploying a single anonymization system across all speakers may leave a substantial portion of the population under-protected.

We also compare per-speaker linkability for the same anonymization system across different attacks and observe substantial variability in per-speaker linkability depending on the attack model, even within a single anonymization system (see an example for selected attack models in Figure 12, Appendix B.4). This indicates that *speaker vulnerability (privacy risk) is not intrinsic but attack-dependent*, suggesting that certain attackers may exploit acoustic features (e.g., prosody, timbre, duration patterns, etc.) that others do not target, leading to dramatically different re-identification success rates for the same speaker under the same anonymization.

Future work will explore which speaker-related traits are leaked by current anonymization systems and investigate models that can predict the optimal anonymization system based on speaker acoustic embeddings or speaker characteristics, enabling real-time adaptive protection without exhaustive evaluation across all system–attack combinations. Such work will require increasing the number of speakers in the test data. We emphasise that all speaker-level results reported here are based on the 29 *LibriSpeech* test-clean speakers, who produce clean, read, English speech. These empirical proportions should be treated as indicative trends confined to this specific test set; generalisation to conversational, noisy, or non-English conditions remains an open question that motivates expanding the speaker population in future challenge editions.

8.3. Attacker advances

This subsection summarises recent advances in attacker models targeting voice anonymization systems, covering both submissions to the First VoicePrivacy Attacker Challenge and subsequent work in the broader literature.

Attacker advances within the challenge

From the defender's perspective, the goal of voice anonymization is to hide the source speaker identity while preserving other attributes, such as linguistic content, prosody, and emotion. It is well observed that these preserved attributes — such as context-dependent duration patterns, temporal speech dynamics, and non-timbral features — inevitably carry residual speaker-dependent cues, which can be highly informative for source speaker retrieval.

In the challenge, participants applied multiple attack strategies, including: reconstruction-based attacks with knowledge distillation (Nhan Tri Do, 2025); multi-system training and transfer learning from ASV models with advanced architectures (*ResNet34*, *TitaNet*, etc.) (pre-)trained on large corpora of original unprocessed data, as well as large semi-supervised models (such as *WavLM*) using target or mixed (Mawalim et al., 2025) anonymized data from multiple systems; advanced data augmentation (Li et al., 2025); and text-based attacks (Ünal Ege Gaznepoglu et al., 2025a). The text-based attacker submitted in the challenge is further analysed and extended in Ünal Ege Gaznepoglu et al. (2025b).

Within the challenge, several teams explored prosody-related features extracted using various pretrained models and evaluated them on anonymized speech (Thebaud et al., 2025). However, the **A.28** attack system did not rank among the top-performing attackers in the challenge, with EERs remaining above 40% in most conditions. As the authors themselves acknowledge (Thebaud et al., 2025), score-level fusion of x-vectors, prosody, and pitch fails to outperform the best individual stream on anonymized data, and the pitch sub-system alone performs near chance level. Consequently, the challenge results for **A.28** should not be interpreted as evidence that prosodic or pitch-based features are intrinsically weak or uninformative as attack signals: the modest performance reflects the specific model capacity and fusion strategy employed, rather than the ultimate potential of these feature classes. More powerful architectures exploiting non-timbral cues may yield substantially stronger attacks, as suggested by other works on duration-based representations (Tomashenko et al., 2025b,c, 2026b) and multimodal attackers (Aloradi et al., 2025).

Beyond improving re-identification performance, Franzreb et al. (2025) investigated how much training data is required for an attacker to reliably and efficiently assess an anonymizer's privacy. Their findings indicate that up to 20% of the training data can be discarded with only minimal degradation in attack effectiveness.

Most participants trained separate attacker models per anonymization system, optimising each configuration independently for a single target. This means that the challenge results do not, in general, permit a controlled comparison of backbone and design choices across teams. Team **A.1** represents a partial exception: by keeping the same ECAPA-TDNN backbone as the VPC 2024 baseline and modifying only training and scoring hyperparameters, they demonstrated that configuration-level choices alone can yield significant EER reductions on a subset of systems. A full controlled same-backbone comparison remains an open direction for future work.

Attacker advances beyond the challenge

Beyond the attacker challenge, recent attacker-side studies explore both data augmentation strategies and diverse feature representations to better capture residual speaker information from anonymized speech.

The *SegReConcat* data augmentation method, introduced in Arefeen et al. (2025), disrupts long-term contextual cues by segmenting anonymized speech at the word level and rearranging segments using random or similarity-based strategies. The rearranged segments are concatenated with the original utterance, thereby forcing the model to focus on extracting residual speaker information from multiple temporal perspectives.

Unlike most existing attacker models, which rely on commonly used speech features such as filterbanks or self-supervised learning (SSL) representations and implicitly exploit residual speaker-discriminative information, recent research investigates alternative directions from a feature-design perspective:

1) Textual feature modeling: Ünal Ege Gaznepoglu et al. (2025b) built a text-based attacker using a *BERT Base* model on ground-truth transcriptions, closely related to their text-based challenge submission (Ünal Ege Gaznepoglu et al., 2025a), and concluded that textual content alone can be indicative of speaker identity due to topic similarity. Subsequently, the same group proposed a dual-branch attacker architecture (Aloradi et al., 2025). One branch employs an ECAPA-TDNN to process anonymized speech and extract residual speaker-related characteristics, while the other branch uses a pretrained *BERT* model to encode ground-truth transcriptions and exploit idiolectal patterns and speaker-specific linguistic characteristics that remain unchanged after anonymization. The authors also found that SpecAugment data augmentation can further improve attacking performance.

2) Explicit temporal dynamics and prosodic feature modeling: Tomashenko et al. (2025b) pointed out that all baseline systems in the VoicePrivacy 2024 Challenge, as well as many state-of-the-art anonymization systems, modify various characteristics of the input (original) speech signal related to speaker identity, while largely preserving

speech rate and phoneme durations. They proposed simple metric-based approaches to measure distance statistics between speakers' average phoneme-duration statistics and showed that, when sufficient data per speaker is available, anonymization systems that do not explicitly alter phoneme durations still preserve a significant amount of speaker-specific information. This direction of work was extended in Tomashenko et al. (2025c) by proposing representations that encode each phoneme occurrence as a duration-weighted one-hot vector, capturing both phoneme identity and temporal variability at the sequence level. Unlike traditional semi-informed attack models, these representations can be trained directly on original speech data, eliminating the need for retraining when applied to anonymized data. These duration-based features replace conventional filterbank features as input to an ECAPA-TDNN speaker verification model, leading to significant improvements in performance for both original and anonymized data compared with simpler duration representations. Bakari et al. (2025) explored prosody-related features extracted using pretrained models and evaluated them on anonymized speech in settings complementary to the attacker challenge, further illustrating that non-timbral cues can be harnessed for re-identification attacks.

These advances demonstrate that attackers can exploit multiple information channels — acoustic embeddings, linguistic content, prosodic patterns, and codec-level representations — to defeat anonymization. The diversity of successful attack strategies underscores the necessity of comprehensive privacy evaluation that accounts for multi-modal information leakage rather than focusing solely on acoustic speaker embeddings.

In addition, the Source Speaker Tracing Challenge (SSTC 2024) (Li et al., 2024) introduced a dataset and benchmark for source speaker verification under 16 common voice conversion methods. While not specifically designed for anonymization evaluation, many anonymization systems are based on voice conversion techniques, making this benchmark directly relevant to the community.

8.4. Future directions

The findings from the First VoicePrivacy Attacker Challenge motivate several complementary research directions:

Generalised vs. system-specific attackers. Most challenge participants trained separate models per anonymization system. Future work should explore both directions: single attacker models that generalise across multiple anonymization systems without per-system fine-tuning, and conversely, ensembles of diverse attackers that simultaneously target a single system from multiple complementary perspectives, increasing the probability of exploiting any residual leakage channel.

Speaker-adaptive and speaker-specific attackers. Rather than optimising average EER across all speakers, future attackers could focus on specific target speakers, incorporating speaker metadata (linguistic profile, prosodic characteristics, demographic information) to maximise re-identification for the most vulnerable individuals under a targeted threat model.

Dynamic attacker selection. A mixture-of-experts style approach, in which different attacker models are dynamically selected depending on the speaker or anonymization system, could combine the strengths of system-specific and generalised attacks in a single adaptive framework. This is conceptually related to score-level fusion but more input-conditional and adaptive.

Speaker-adaptive anonymization. From the defender's side, speaker-adaptive anonymization should learn mappings that minimise worst-case linkability for each individual, and adapt transformation strength (e.g., spectral and prosodic distortion) to speaker-specific vulnerability rather than using global averages.

Richer vulnerability modeling. This includes analysis of acoustic and linguistic predictors of high residual linkability (pitch range, formant structure, speaking rate, prosodic variability, lexical patterns), clustering of per-speaker linkability vectors to discover latent vulnerability types, and longitudinal studies of how attacker performance evolves with prolonged exposure to a target's speech.

Stronger and more realistic attack models. This includes meta-attackers trained jointly across multiple anonymization systems to exploit cross-system consistency, attacks that incorporate metadata and more realistic privacy scenarios (e.g., when the anonymization algorithm is unknown), and attacks for more challenging scenarios such as target-speaker anonymization (Tomashenko et al., 2026c) in multi-speaker overlapping conversational recordings, or multi-speaker anonymization (Miao et al., 2025).

Legally grounded and speaker-level evaluation protocols. Current challenge evaluation relies on EER categories that do not map directly to legal notions of anonymity. Future editions should adopt evaluation frameworks grounded in legal definitions of identifiability, such as the linkability and singling-out metrics of Vauquier et al. (2025), include per-speaker worst-case linkability as a mandatory reported metric alongside global averages, and treat conversation length as a first-class evaluation parameter, given its strong influence on re-identification risk (Vauquier et al., 2025).

Complementary work on content anonymization for long conversational recordings (Aggazzotti et al., 2026) highlights additional attack surfaces arising from extended interaction context and conversation-level metadata, underscoring the importance of evaluation protocols that explicitly account for long-form, multi-session use cases rather than only short, isolated utterances. Recent analysis of speech privacy differences further emphasises that evaluation frameworks must jointly consider how speech is captured, transformed, and reused, together with heterogeneous privacy expectations across speakers and use cases (Welch et al., 2026).

Fairness-aware evaluation and anonymization. Current evaluation reports average EER across all speakers, masking large disparities in privacy protection. Future challenges should explicitly report privacy fairness metrics quantifying the gap between the best- and worst-protected speakers, and require anonymization systems to bound worst-case per-speaker linkability rather than only optimise average performance. From the design perspective, this motivates fairness-aware anonymization systems that adapt transformation strength to individual speaker vulnerability.

Cross-dataset and cross-domain generalisation. All challenge results are based on *LibriSpeech* read speech with a limited number of speakers in controlled recording conditions. Future work should evaluate attackers on spontaneous, conversational, or multilingual speech (Miao et al., 2026), where speaker characteristics and anonymization effectiveness may differ substantially, to assess whether the conclusions drawn from the present challenge generalise beyond the current evaluation set-up.

Attack transferability across anonymization systems. Most teams submitted system-specific attackers. Studying how well an attacker trained on one anonymization system transfers to unseen anonymization systems — zero-shot transfer — would reveal whether universal residual leakage channels persist regardless of the anonymization method, and inform the design of more robust anonymization approaches.

Joint optimisation of anonymization and attack robustness. Adversarial co-training between anonymization and attack models, in the spirit of adversarial training frameworks used in generative adversarial networks, would allow the anonymizer to be jointly trained to minimise speaker leakage under the strongest known attacker. This moves beyond the sequential challenge format toward a co-evolutionary evaluation framework that could yield substantially more robust anonymization systems.

Taken together, the growing range of advanced attack strategies, heterogeneous vulnerability across speakers, and emerging fairness concerns indicate that the next generation of voice anonymization systems must move beyond uniform, non-adaptive designs toward adaptive, fairness-aware architectures that provide more equitable privacy protection across diverse speaker populations and attack models.

9. Conclusions

The First VoicePrivacy Attacker Challenge demonstrates that the privacy protection of even top-performing anonymization systems from VoicePrivacy 2024 was substantially overestimated. Through systematic evaluation of 55 attacker submissions across seven anonymization systems, the challenge reveals that advanced attack methodologies can reduce baseline EER by 25–44% relative, with best attackers achieving EER values around 20%–28% compared to baseline attacker performance of approximately 27%–43%. This finding underscores a critical vulnerability: although the best anonymization methods maintain moderate protection (EER$\sim$26–28%), the re-identification risks are substantially higher than previously assumed, revealing potential real-world deployment threats when adversaries have access to system implementations and anonymized training data.

Key attack method categories include: (1) architecture enhancements, i.e., *TitaNet-Large* fine-tuning (**A.22**), *wav2vec 2.0* feature extraction and spectrogram resizing (**A.41**), and multi-stream fusion of x-vectors, prosody/rhythm features, and pitch (**A.28**); (2) data strategies, such as mixing original/anonymized data, knowledge distillation, perceptual hashing (**A.31**), speed perturbation, SpecAugment, and multi-system data pooling; (3) specialized attacks, which comprise embedding reconstruction using MHFA with gradient reversal for prosody handling and codec-level recovery tailored to **B4** (**A.4**), and cherry-picking for AdMixture (**A.42**); (4) enhanced scoring metrics, i.e. replacing average cosine with max similarity, thresholded percentage, or PLDA to handle multi-modal distributions (**A.1**, especially for **T8-5**); (5) system-independent, text-based attacks using *BERT* classification that exploit linguistic similarities in transcripts (**A.42**). Top-performing attack strategies include LoRA-adapted *ResNet34* with *WavLM*-Large features and three-stage training (pretraining on *VoxCeleb*, mixing original and anonymized data, and per-system fine-tuning) (**A.20**, best for six anonymization systems); and ECAPA-TDNN with PLDA on mixed data, enhanced by SpecAugment and multi-stage fine-tuning (**A.5**, best for one anonymization system). This diversity of

attack strategies reveals that speaker identity leaks occur not in isolated failure modes but across multiple dimensions of the anonymization design space.

Acknowledgment

This work was conducted in the context of the Inria–NII TrustedSpeech Associate Team. This work was supported by the French National Research Agency under the SpeechPrivacy project (https://anr.fr/Projet-ANR-23-CE23-0022) and the IPoP project funded by the French Cybersecurity PEPR (https://www.pepr-cybersecurite.fr/projet/ipop/). Experiments were carried out using the Grid'5000 testbed.

References

Adigwe, A., Tits, N., Haddad, K.E., Ostadabbas, S., Dutoit, T., 2018. The emotional voices database: Towards controlling the emotion dimension in voice generation systems. arXiv preprint arXiv:1806.09514 .

Aggazzotti, C., Garg, A., Cai, Z., Andrews, N., 2026. Content anonymization for privacy in long-form audio, in: ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 13412–13416.

Aloradi, A., Gaznepoglu, Ü.E., Habets, E.A., Tenbrinck, D., 2025. VoxATtack: A multimodal attack on voice anonymization systems. arXiv preprint arXiv:2507.12081 .

Aouf, A., 2020. Basic arabic vocal emotions dataset. URL: https://www.kaggle.com/ds/345828, doi:10.34740/KAGGLE/DS/345828.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M., Weber, G., 2019. Common voice: A massively-multilingual speech corpus. arXiv preprint arXiv:1912.06670 .

Arefeen, R., Miao, X., Tong, R., Ng, A.B., See, S., 2025. SegReConcat: A data augmentation method for voice anonymization attack. APSIPA ASC .

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., et al., 2021. XLS-R: Self-supervised cross-lingual speech representation learning at scale. arXiv preprint arXiv:2111.09296 .

Bäckström, T., Vali, M.H., Nguyen, M., Rech, S., 2025. Privacy disclosure of similarity rank in speech and language processing. IEEE Transactions on Audio, Speech and Language Processing 34, 196–205.

Baevski, A., Zhou, Y., Mohamed, A., Auli, M., 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems 33, 12449–12460.

Bain, M., Huh, J., Han, T., Zisserman, A., 2023. Whisperx: Time-accurate speech transcription of long-form audio, in: Proc. Interspeech 2023, pp. 4489–4493.

Bakari, R., Blouch, O.L., Evans, N., Gengembre, N., Panariello, M., Todisco, M., 2025. The influence of non-timbral cues in voice anonymisation and evaluation, in: 5th Symposium on Security and Privacy in Speech Communication, pp. 35–42. doi:10.21437/SPSC.2025-6.

Baveye, Y., Dellandrea, E., Chamaret, C., Chen, L., 2015. LIRIS-ACCEDE: A video database for affective content analysis. IEEE Transactions on Affective Computing 6, 43–55.

Bengio, S., Mariéthoz, J., 2004. A statistical significance test for person authentication, in: Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop.

Bu, H., Du, J., Na, X., Wu, B., Zheng, H., 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline, in: 2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA), IEEE. pp. 1–5.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B., 2005. A database of German emotional speech, in: Interspeech, pp. 1517–1520.

Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S., 2008. IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation 42, 335–359.

Champion, P., 2023. Anonymizing Speech: Evaluating and Designing Speaker Anonymization Techniques. Ph.D. thesis. Université de Lorraine.

Chen, O.T.C., Hsu, Y.C., Yang, T.S., 2024a. Detecting speech deepfakes through improved speech features and cost functions, in: 2024 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), IEEE. pp. 35–39.

Chen, O.T.C., Tsai, Y.C., 2025. System description for First VoicePrivacy Attacker Challenge, in: The First VoicePrivacy Attacker Challenge. URL: https://www.voiceprivacychallenge.org/attacker/docs/A26.pdf.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Wei, F., 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing 16, 1505–1518.

Chen, Y., Zheng, S., Wang, H., Cheng, L., , et al., 2024b. ERes2NetV2: Boosting short-duration speaker verification performance with computational efficiency .

Chen, Y., Zheng, S., Wang, H., Cheng, L., Chen, Q., Qi, J., 2023. An Enhanced Res2Net with Local and Global Feature Fusion for Speaker Verification, in: Interspeech 2023, pp. 2228–2232. doi:10.21437/Interspeech.2023-1294.

Chung, J.S., Nagrani, A., Zisserman, A., 2018. VoxCeleb2: Deep speaker recognition, in: Interspeech, pp. 1086–1090.

Défossez, A., Copet, J., Synnaeve, G., Adi, Y., 2023. High fidelity neural audio compression. Transactions on Machine Learning Research URL: https://openreview.net/forum?id=ivCd8z8zR2.

Delgado, H., Evans, N., Jung, J.w., Kinnunen, T., Kukanov, I., Lee, K.A., Liu, X., Shim, H.j., Sahidullah, M., Tak, H., et al., 2024. ASVspoof 5 Evaluation Plan. Technical Report. URL: https://www.asvspoof.org/file/ASVspoof5___Evaluation_Plan_Phase2.pdf.

Desplanques, B., Thienpondt, J., Demuynck, K., 2020. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification, in: Interspeech, pp. 3830–3834.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), pp. 4171–4186.

Ünal Ege Gaznepoglu, Leschanowsky, A., Aloradi, A., Singh, P., Tenbrinck, D., Habets, E.A.P., Peters, N., 2025a. FAU-Erlangen system description for the 1st VoicePrivacy Attacker Challenge, in: The First VoicePrivacy Attacker Challenge. URL: https://www.voiceprivacychallenge.org/attacker/docs/A42.pdf.

Ünal Ege Gaznepoglu, Leschanowsky, A., Aloradi, A., Singh, P., Tenbrinck, D., Habets, E.A.P., Peters, N., 2025b. You are what you say: Exploiting linguistic content for voiceprivacy attacks, in: Interspeech 2025, pp. 4238–4242. doi:10.21437/Interspeech.2025-681.

Franzreb, C., Das, A., Polzehl, T., Möller, S., 2025. Optimizing the Dataset for the Privacy Evaluation of Speaker Anonymizers, in: 5th Symposium on Security and Privacy in Speech Communication, pp. 18–26. doi:10.21437/SPSC.2025-4.

Gu, W., Liu, Z., Chen, L., Wang, R., Guo, C., Guo, W., Lee, K.A., Ling, Z.H., 2024. USTC-PolyU system for the VoicePrivacy 2024 Challenge.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al., 2014. Deep speech: Scaling up end-to-end speech recognition. arxiv 2014. arXiv preprint arXiv:1412.5567 .

Haq, S., Jackson, P.J., Edge, J., 2009. Speaker-dependent audio-visual emotion recognition, in: International Conference on Auditory-Visual Speech Processing (AVSP), pp. 53–58.

Hernandez, F., Nguyen, V., Ghannay, S., Tomashenko, N., Esteve, Y., 2018. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation, in: Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings 20, Springer. pp. 198–208.

Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A., 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing 29, 3451–3460.

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2021. LoRA: Low-rank adaptation of large language models, in: arXiv preprint arXiv:2106.09685.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132–7141.

Ju, Z., Wang, Y., Shen, K., Tan, X., Xin, D., Yang, D., Liu, Y., Leng, Y., Song, K., Tang, S., Wu, Z., Qin, T., Li, X.Y., Ye, W., Zhang, S., Bian, J., He, L., Li, J., Zhao, S., 2024. NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. arXiv preprint arXiv:2403.03100 .

Kim, J., Kong, J., Son, J., 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech, in: International Conference on Machine Learning (ICML), pp. 5530–5540.

Ko, T., Peddinti, V., Povey, D., Seltzer, M.L., Khudanpur, S., 2017. A study on data augmentation of reverberant speech for robust speech recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5220–5224.

Koluguri, N.R., Park, T., Ginsburg, B., 2022. TitaNet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 8102–8106.

Krzywdziak, J., Ryzhankow, I., Szmajdzinski, S., Sredniawa, W., Zak, J., Cieslak, A., Masztalski, P., 2025. SRPOL voice hacking: methods of attacking speaker verification anonymization algorithms, in: The First VoicePrivacy Attacker Challenge. URL: https://www.voiceprivacychallenge.org/attacker/docs/A31.pdf.

Kuzmin, N., Luong, H.T., Yao, J., Xie, L., Lee, K.A., Chng, E.S., 2024. NTU-NPU system for Voice Privacy 2024 challenge. arXiv preprint arXiv:2410.02371 .

Li, J., Tu, W., Xiao, L., 2023a. FreeVC: Towards high-quality text-free one-shot voice conversion, in: Proc. ICASSP, IEEE. pp. 1–5.

Li, L., Li, X., Jiang, H., Chen, C., Hou, R., Wang, D., 2023b. CN-Celeb-AV: A multi-genre audio-visual dataset for person recognition. arXiv preprint arXiv:2305.16049 .

Li, X., Wu, X., Lu, H., Liu, X., Meng, H., 2021. Channel-Wise Gated Res2Net: Towards robust detection of synthetic speech attacks, in: Proc. Interspeech 2021, pp. 4314–4318.

Li, Y., Zheng, Y., Guo, Z., Wang, Y., Yin, J., Fei, H., 2025. SpecWav-Attack: Leveraging spectrogram resizing and wav2vec 2.0 for attacking anonymized speech, in: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 1–2.

Li, Z., Lin, Y., Yao, T., Suo, H., Zhang, P., Ren, Y., Cai, Z., Nishizaki, H., Li, M., 2024. The database and benchmark for the source speaker tracing challenge 2024, in: 2024 IEEE Spoken Language Technology Workshop (SLT), IEEE. pp. 1254–1261.

Lin, Y., Cheng, M., Zhang, F., Gao, Y., Zhang, S., Li, M., 2024. Voxblink2: A 100k+ speaker recognition corpus and the open-set speaker-identification benchmark, in: Interspeech 2024, pp. 4263–4267. doi:10.21437/Interspeech.2024-1490.

Livingstone, S.R., Russo, F.A., 2018. 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 13, e0196391.

Lyu, X., Wang, Y., Zhao, T., Liu, H., 2025. Fast adaptation of pretrained speaker verification system for source speaker tracking, in: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 1–2.

Mawalim, C.O., Adila, A., Unoki, M., 2025. Fine-tuning TitaNet-Large model for speaker anonymization attacker systems, in: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 1–2.

Meyer, S., Lux, F., Koch, J., Denisov, P., Tilli, P., Vu, N.T., 2023. Prosody is not identity: A speaker anonymization approach using prosody cloning, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5.

Miao, X., Tao, R., Zeng, C., Wang, X., 2025. A benchmark for multi-speaker anonymization. IEEE Transactions on Information Forensics and Security 20, 3819–3833.

Miao, X., Tomashenko, N., Arefeen, R., Meyer, S., Panariello, M., Wang, X., Vincent, E., Evans, N., Yamagishi, J., Todisco, M., 2026. The VoicePrivacy 2026 Challenge evaluation plan.

Nautsch, A., Patino, J., Tomashenko, N., Yamagishi, J., Noé, P.G., Bonastre, J.F., Todisco, M., Evans, N., 2020. The Privacy ZEBRA: Zero evidence biometric recognition assessment, in: Interspeech, pp. 1698–1702.

Nhan Tri Do, 2025. Voice attacker leveraging multi-head factorized attentive reconstructor and gradient reversal for random prosody anonymization, in: The First VoicePrivacy Attacker Challenge. URL: https://www.voiceprivacychallenge.org/attacker/docs/A4.pdf.

Noé, P.G., Nautsch, A., Evans, N., Patino, J., Bonastre, J.F., Tomashenko, N., Matrouf, D., 2022. Towards a unified assessment framework of speech pseudonymisation. Computer Speech & Language 72, 101299.

Panariello, M., Nespoli, F., Todisco, M., Evans, N., 2024a. Speaker anonymization using neural audio codec language models, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4725–4729.

Panariello, M., Tomashenko, N., Wang, X., Miao, X., Champion, P., Nourtel, H., Todisco, M., Evans, N., Vincent, E., Yamagishi, J., 2024b. The VoicePrivacy 2022 Challenge: Progress and perspectives in voice anonymisation. IEEE/ACM Transactions on Audio, Speech, and Language Processing 32, 3477–3491. doi:10.1109/TASLP.2024.3430530.

Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015. LibriSpeech: an ASR corpus based on public domain audio books, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5206–5210.

Parada-Cabaleiro, E., Costantini, G., Batliner, A., Schmitt, M., Schuller, B.W., 2020. Demos: An italian emotional speech corpus: Elicitation methods, machine learning, and perception. Language Resources and Evaluation 54, 341–383.

Park, D.S., Chan, W., Zhang, Y., Chiu, C.C., Zoph, B., Cubuk, E.D., Le, Q.V., 2019. SpecAugment: A simple data augmentation method for automatic speech recognition, in: Interspeech 2019, pp. 2613–2617. doi:10.21437/Interspeech.2019-2680.

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., Collobert, R., 2020. MLS: A large-scale multilingual dataset for speech research, in: Interspeech 2020, pp. 2757–2761. doi:10.21437/Interspeech.2020-2826.

Qian, K., Zhang, Y., Chang, S., Yang, X., Hasegawa-Johnson, M., 2019. AutoVC: Zero-shot voice style transfer with only autoencoder loss, PMLR, Long Beach, California, USA. pp. 5210–5219. URL: http://proceedings.mlr.press/v97/qian19c.html.

Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., 2023. Robust speech recognition via large-scale weak supervision, in: International Conference on Machine Learning, PMLR. pp. 28492–28518.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.C., Yeh, S.L., Fu, S.W., Liao, C.F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R.D., Bengio, Y., 2021. SpeechBrain: A general-purpose speech toolkit. arXiv preprint arXiv:2106.04624 .

Riou, A., Lattner, S., Hadjeres, G., Peeters, G., 2023. Pesto: Pitch estimation with self-supervised transposition-equivariant objective, in: Proceedings of the International Society for Music Information Retrieval Conference.

Snyder, D., Chen, G., Povey, D., 2015. MUSAN: A music, speech, and noise corpus. arXiv:1510.08484.

Thebaud, T., Mehlman, N., Guan, Y., Moro-Velazquez, L., Lopez, J.V., Narayanan, S., Dehak, N., 2025. PPX-Anon: Prosody, Pitch and X-Vectors for De-Anonymization; our submission to the Voice Attacker Challenge 2024, in: 5th Symposium on Security and Privacy in Speech Communication, pp. 61–67. doi:10.21437/SPSC.2025-10.

Tomashenko, N., Miao, X., Champion, P., Meyer, S., Panariello, M., Wang, X., Evans, N., Vincent, E., Yamagishi, J., Todisco, M., 2026a. The third VoicePrivacy challenge: Preserving emotional expressiveness and linguistic content in voice anonymization. Computer Speech & Language 100, 101988. URL: https://www.sciencedirect.com/science/article/pii/S0885230826000513, doi:https://doi.org/10.1016/j.csl.2026.101988.

Tomashenko, N., Miao, X., Champion, P., Meyer, S., Wang, X., Vincent, E., Panariello, M., Evans, N., Yamagishi, J., Todisco, M., 2024a. The VoicePrivacy 2024 challenge evaluation plan. arXiv preprint arXiv:2404.02677 .

Tomashenko, N., Miao, X., Vincent, E., Yamagishi, J., 2024b. The First VoicePrivacy Attacker Challenge evaluation plan. arXiv preprint arXiv:2410.07428 .

Tomashenko, N., Miao, X., Vincent, E., Yamagishi, J., 2025a. The First VoicePrivacy Attacker Challenge, in: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 1–2.

Tomashenko, N., Srivastava, B.M.L., Wang, X., Vincent, E., Nautsch, A., Yamagishi, J., Evans, N., Patino, J., Bonastre, J.F., Noé, P.G., Todisco, M., 2020. Introducing the VoicePrivacy Initiative, in: Interspeech, pp. 1693–1697. doi:10.21437/Interspeech.2020-1333.

Tomashenko, N., Vincent, E., Tommasi, M., 2025b. Analysis of speech temporal dynamics in the context of speaker verification and voice anonymization, in: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 1–5.

Tomashenko, N., Vincent, E., Tommasi, M., 2025c. Exploiting context-dependent duration features for voice anonymization attack systems, in: Interspeech 2025, pp. 5128–5132. doi:10.21437/Interspeech.2025-2317.

Tomashenko, N., Vincent, E., Tommasi, M., 2026b. Learning temporal-dynamics metrics for attacks on voice anonymization, in: 6th Symposium on Security and Privacy in Speech Communication.

Tomashenko, N., Wang, X., Vincent, E., Patino, J., Srivastava, B.M.L., Noé, P.G., Nautsch, A., Evans, N., Yamagishi, J., O'Brien, B., Chanclu, A., Bonastre, J.F., Todisco, M., Maouche, M., 2022. The VoicePrivacy 2020 Challenge: Results and findings. Computer Speech and Language 74, 101362. doi:https://doi.org/10.1016/j.csl.2022.101362.

Tomashenko, N., Yamagishi, J., Wang, X., Liu, Y., Vincent, E., 2026c. Target speaker anonymization in multi-speaker recordings, in: ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 13417–13421.

Tong, F., Zhao, M., Zhou, J., Lu, H., Li, Z., Li, L., Hong, Q., 2021. ASV-Subtools: Open source toolkit for automatic speaker verification, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 6184–6188.

Vauquier, N., Srivastava, B.M.L., Hosseini, S.A., Vincent, E., 2025. Legally validated evaluation framework for voice anonymization, in: Interspeech 2025, pp. 3229–3233. doi:10.21437/Interspeech.2025-1699.

Vryzas, N., Kotsakis, R., Liatsou, A., Dimoulas, C.A., Kalliris, G., 2018. Speech emotion recognition for performance interaction. Journal of the Audio Engineering Society 66, 457–467.

Wan, L., Wang, Q., Papir, A., Moreno, I.L., 2018. Generalized end-to-end loss for speaker verification, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 4879–4883.

Wang, D., Deng, L., Yeung, Y.T., Chen, X., Liu, X., Meng, H., 2021. VQMIVC: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion, in: Proc. Interspeech 2021, pp. 1344–1348. doi:10.21437/Interspeech.2021-283.

Wang, H., Liang, C., Wang, S., Chen, Z., Zhang, B., Xiang, X., Deng, Y., Qian, Y., 2023a. Wespeaker: A research and production oriented speaker embedding learning toolkit, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 1–5.

Wang, H., Zheng, S., Chen, Y., Cheng, L., Chen, Q., 2023b. CAM++: A Fast and Efficient Network for Speaker Verification Using Context-Aware Masking, in: Interspeech 2023, pp. 5301–5305. doi:10.21437/Interspeech.2023-1513.

Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., Sahidullah, M., Vestman, V., Kinnunen, T., Lee, K.A., et al., 2020. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. Computer Speech & Language 64, 101114.

Wang, Y., Stanton, D., Zhang, Y., Ryan, R.S., Battenberg, E., Shor, J., Xiao, Y., Ren, F., Jia, Y., Saurous, R.A., 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis, in: International Conference on Machine Learning (ICML), pp. 5180–5189.

Welch, P., Tomashenko, N., Das, S., Williams, J., 2026. Differential performance and contextual integrity in speech privacy. IEEE Security & Privacy .

Williams, J., Pizzi, K., Tomashenko, N., Das, S., 2024. Anonymizing speaker voices: Easy to imitate, difficult to recognize?, in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 12491–12495.

Xinyuan, H.L., Cai, Z., Garg, A., Duh, K., García-Perera, L.P., Khudanpur, S., Andrews, N., Wiesner, M., 2024. HLTCOE JHU submission to the Voice Privacy challenge 2024. arXiv preprint arXiv:2409.08913 .

Xinyuan, H.L., Garg, A., Cai, Z., Duh, K., García-Perera, L.P., Khudanpur, S., Andrews, N., Wiesner, M., 2025. HLTCOE submission to the VoicePrivacy attacker challenge, in: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 1–2.

Yao, J., Kuzmin, N., Wang, Q., Guo, P., Ning, Z., Guo, D., Lee, K.A., Chng, E.S., Xie, L., 2024. NPU-NTU system for Voice Privacy 2024 challenge. arXiv preprint arXiv:2409.04173 .

Zadeh, A.B., Liang, P.P., Poria, S., Cambria, E., Morency, L.P., 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph, in: 56th Annual Meeting of the ACL (Volume 1: Long Papers), pp. 2236–2246.

Zhang, Y., Bi, Z., Xiao, F., Yang, X., Zhu, Q., Guan, J., 2025a. Attacking voice anonymization systems with augmented feature and speaker identity difference, in: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 1–2.

Zhang, Y., Bi, Z., Xiao, F., Yang, X., Zhu, Q., Guan, J., 2025b. GISP-HEU's submission for VoicePrivacy Attacker Challenge at ICASSP 2025, in: The First VoicePrivacy Attacker Challenge. URL: https://www.voiceprivacychallenge.org/attacker/docs/A5-system.pdf.

A. Training resources

Table 5 presents the list of allowed training resources for attack systems.

Table 5: List of models and data for training attacker systems.

#	Model	Link
1	WavLM (Chen et al., 2022) Base and Large	https://github.com/microsoft/unilm/tree/master/wavlm https://huggingface.co/microsoft/wavlm-large
2	Whisper (Radford et al., 2023)	https://github.com/openai/whisper https://huggingface.co/openai/whisper-large
3	HuBERT (Hsu et al., 2021)	https://github.com/facebookresearch/fairseq/blob/main/examples/hubert https://huggingface.co/facebook/hubert-large-ls960-ft
4	XLS-R (Babu et al., 2021)	https://github.com/facebookresearch/fairseq/blob/main/examples/wav2vec/xlsr https://huggingface.co/facebook/wav2vec2-large-xlsr-53
5	wav2vec 2.0 (Baevski et al., 2020)	https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec https://dl.fbaipublicfiles.com/voxpopuli/models/wav2vec2_large_west_germanic_v2.pt https://huggingface.co/facebook/wav2vec2-large-960h-lv60
6	ECAPA-TDNN (Desplanques et al., 2020)	https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb
7	NaturalSpeech 3 (Ju et al., 2024)	https://huggingface.co/amphion/naturalspeech3_facodec
8	Encodec (Défossez et al., 2023)	https://huggingface.co/facebook/encodec_24khz
9	Bark	https://huggingface.co/suno/bark https://huggingface.co/erogol/bark/tree/main
10	Resnet34	https://wenet.org.cn/downloads?models=wespeaker&version=voxceleb_resnet34.zip

11	Resnet34_lm	https://wenet.org.cn/downloads?models=wespeaker&version=voxceleb_resnet34_LM.zip
12	VGGVox	https://github.com/a-nagrani/VGGVox
13	Kaldi models	http://kaldi-asr.org/models.html
14	Conformer models	https://huggingface.co/models?search=conformer
15	NVIDIA TitaNet-Large (en-US) (Koluguri et al., 2022)	https://huggingface.co/nvidia/speakerverification_en_titanet_large

#	Dataset	Link
16	LibriSpeech (Panayotov et al., 2015): train-clean-100, train-clean-360, train-other-500	https://www.openslr.org/12
17	RAVDESS (Livingstone and Russo, 2018)	https://datasets.activeloop.ai/docs/ml/datasets/ravdess-dataset/ https://zenodo.org/records/1188976
18	SAVEE (Haq et al., 2009)	http://kahlan.eps.surrey.ac.uk/savee/
19	EMO-DB (Burkhardt et al., 2005)	http://emodb.bilderbar.info/download/
20	VoxCeleb-1,2 (Chung et al., 2018)	https://www.robots.ox.ac.uk/~vgg/data/voxceleb/index.html#about
21	CMU-MOSEI (Zadeh et al., 2018)	http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/
22	MUSAN (Snyder et al., 2015)	https://www.openslr.org/17/
23	RIR (Ko et al., 2017)	https://www.openslr.org/28/
24	CN-Celeb (Li et al., 2023b)	https://cnceleb.org/
25	Common Voice (Ardila et al., 2019)	https://commonvoice.mozilla.org/en/datasets
26	IEMOCAP (Busso et al., 2008)	https://sail.usc.edu/iemocap/iemocap_info.htm
27	Emo-DB (Burkhardt et al., 2005)	http://emodb.bilderbar.info/index-1280.html
28	Toronto Emotional Speech Database	https://tspace.library.utoronto.ca/handle/1807/24487
29	LIRIS-ACCEDE (Baveye et al., 2015)	https://liris-accede.ec-lyon.fr/database.php
30	AESDD (Vryzas et al., 2018)	http://m3c.web.auth.gr/research/aesdd-speech-emotion-recognition
31	ANAD	https://www.kaggle.com/suso172/arabic-natural-audio-dataset
32	BAVED (Aouf, 2020)	https://www.kaggle.com/a13x10/basic-arabic-vocal-emotions-dataset/
33	VoxForge	http://www.voxforge.org/
34	TED-LIUM 3 (Hernandez et al., 2018)	https://www.openslr.org/51/
35	AISHELL-1 (Bu et al., 2017)	https://www.openslr.org/33/
36	AISHELL-WakeUp-1	https://www.aishelltech.com/wakeup_data
37	CMU-MOSEI (Zadeh et al., 2018)	http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/
38	VoxBlink2 (Lin et al., 2024)	https://voxblink2.github.io/
39	Emotional Voices Database (Adigwe et al., 2018)	https://github.com/numediart/EmoV-DB
40	DEMoS (Parada-Cabaleiro et al., 2020)	https://zenodo.org/record/2544829
41	Multilingual LibriSpeech (MLS) (Pratap et al., 2020), English	https://www.openslr.org/94/

#	Software with pretrained models	Link
42	VITS (Kim et al., 2021)	https://github.com/jaywalnut310/vits/ Models: https://drive.google.com/drive/folders/1ksarh-cJf3F5eKJjLVWY0X1j1qsQqiS2
43	SASV fusion-based baseline from the SASV 2022 challenge (Wang et al., 2020)	https://github.com/sasv-challenge/SASVC2022_Baseline

44	Baseline of the ASVspoof 5 challenge Track 2, SASV (Delgado et al., 2024)	https://github.com/sasv-challenge/SASV2_Baseline/tree/asvspoof5
45	ASV-Subtools (Tong et al., 2021)	https://github.com/Snowdar/asv-subtools
46	WeSpeaker (Wang et al., 2023a)	https://github.com/wenet-e2e/wespeaker Models: https://github.com/wenet-e2e/wespeaker/blob/master/docs/pretrained.md
47	3D-Speaker *ERes2Net* (Chen et al., 2023) *ERes2NetV2* (Chen et al., 2024b) *CAM++* (Wang et al., 2023b)	https://github.com/modelscope/3D-Speaker
48	SpeechBrain (Ravanelli et al., 2021)	https://speechbrain.github.io/ Models: https://huggingface.co/speechbrain
49	ResNet (Hu et al., 2018)	https://github.com/ranchlai/speaker-verification
50	VQMIVC (Wang et al., 2021)	https://github.com/Wendison/VQMIVC
51	DeepSpeech (Hannun et al., 2014)	https://github.com/mozilla/DeepSpeech Models: https://github.com/mozilla/DeepSpeech/releases
52	AutoVC (Qian et al., 2019)	https://github.com/auspicious3000/autovc

B. Results

B.1. Equal error rate

Figure 6 shows the EER values on the *LibriSpeech* development and test datasets, separately for male and female speakers, reported with 95% confidence intervals calculated as suggested by Bengio and Mariéthoz (2004).

B.2. Privacy-utility tradeoff

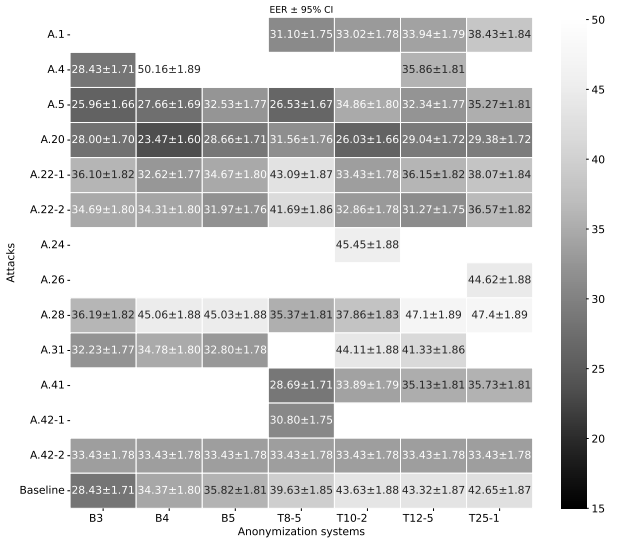
Figure 7 illustrates the performance of anonymization systems based on utility metrics — *word error rate* (WER), which measures linguistic content preservation, and *unweighted average recall* (UAR), which measures emotion preservation — as well as the privacy metric (EER) for the baseline (top) and best-attack (bottom) models on the test dataset. Privacy-utility tradeoff in VPC 2024 evaluation is assessed using four privacy categories (EER $\geq$ 10%, 20%, 30%, 40%). For each privacy category, submissions are ranked separately by utility (WER, UAR). The red vertical lines correspond to the privacy categories for the VPC 2024 participants' anonymization systems (**T8-5**, **T10-2**, **T12-5**, **T25-1**): EER $\geq$ 40% for baseline attacks and EER $\geq$ 20% for best attacks. All VPC 2024 participant systems consistently shifted from the fourth strongest privacy category (EER $\geq$ 40%) to the second (EER $\geq$ 20%) when evaluated with the strongest attack models compared to baseline attacks. Since the VPC 2024 ranking was performed independently by utility (WER, UAR) within each privacy category, the best attack models do not alter the relative utility ranking among the four VPC 2024 anonymization systems (**T8-5**, **T10-2**, **T12-5**, **T25-1**), which all shift from category 4 (EER $\geq$ 40%) to category 2 (EER $\geq$ 20%) while preserving order according to utility metrics. However, including the three baseline systems **B3**, **B4**, and **B5** changes the overall ranking of all seven systems, as they start in different privacy categories under baseline attackers (**B3** in category 2 with EER $\approx$ 27%, **B4/B5** in category 3 with EER $\approx$ 30$-$34%), but stronger attackers produce inconsistent shifts (larger EER reductions for participant systems). This category convergence and differential improvement suggests potential changes in the global utility-privacy ranking when re-evaluating all VPC 2024 systems (including baselines) under the strongest attackers.

B.3. Linkability

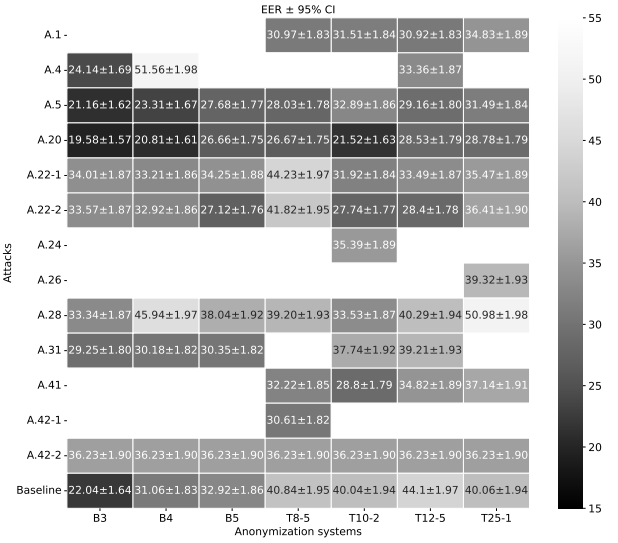
Figure 8 presents the results on the test data using the *linkability* metric, formatted consistently with Figure 2 for EER (the same style and notations). Compared to the EER-based evaluation, several attack models exhibit a different ranking when assessed using the *linkability* metric. In this experiment, the *linkability* metric (Vauquier et al., 2025) was computed from the same score distributions as those used for EER.

For each *trial* utterance u_s of *enrollment* speaker $s \in T$ in the test set T, we performed comparisons using a distance metric ρ[5] with *enrollment* speakers from subset $S \subseteq E$ (where E denotes the set of *enrollment* speakers and $s \notin S$),

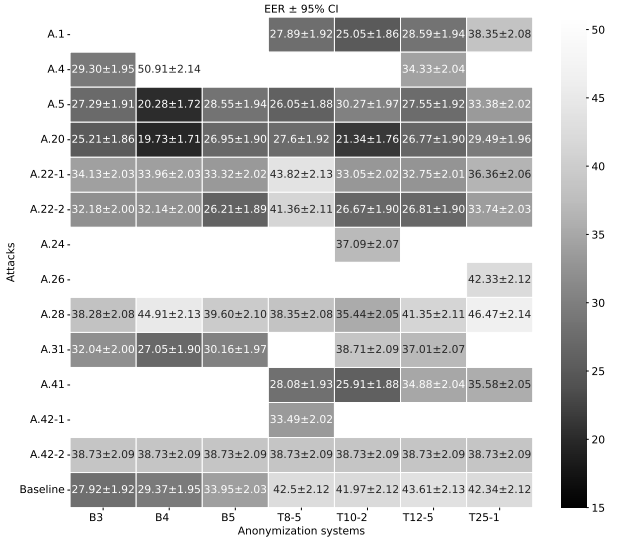
[5] Here, ρ represents the team-submitted scores or, for the baseline, cosine similarity between the *trial* utterance embedding and the corresponding average *enrollment* utterance embeddings.



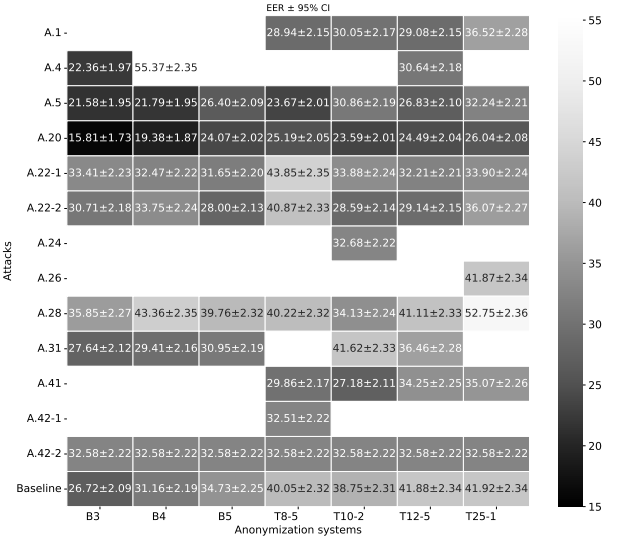
(a) Development female (original data: EER=10.51 ± 1.16%).



(b) Development male (original data: EER=0.93 ± 0.38%).



(c) Test female (original data: EER=8.76 ± 1.21%).



(d) Test male (original data: EER=0.42 ± 0.31%).

Figure 6: Challenge results (EER,% with 95% CI) on the development (top row) and test (bottom row) datasets for female (left) and male (right) speakers.

as well as one comparison with the same speaker in the *enrollment* set. The probability of linkage on the test set is then computed as:

$$P_{\text{link}}(T) = P\left(\rho(u_s, s) > \rho(u_s, s^*)\ \forall\ s^* \in S \mid s \in T\right). \tag{1}$$

P_{link} ranges in [0, 1], with lower values indicating stronger anonymization, i.e., a lower probability that the correct (same) speaker ranks above all impostors (different speakers). From the attacker's perspective, higher linkability scores correspond to stronger attacks, as the attacker can more reliably link anonymized trial utterances to the corresponding enrollment speakers. In this experiment, $|S| = 12$ for male speakers and $|S| = 15$ for female speakers.

Figure 9 provides a deeper analysis of the relationship between EER and linkability scores. The Pearson correlation is $r = 0.862$ ($p = 8.17 \times 10^{-21}$). Despite the expected strong correlation between EER and $(1 - P_{\text{link}})$ scores, some (anonymization system, attacker) pairs exhibit noticeably different behavior across the two metrics. In particular, for

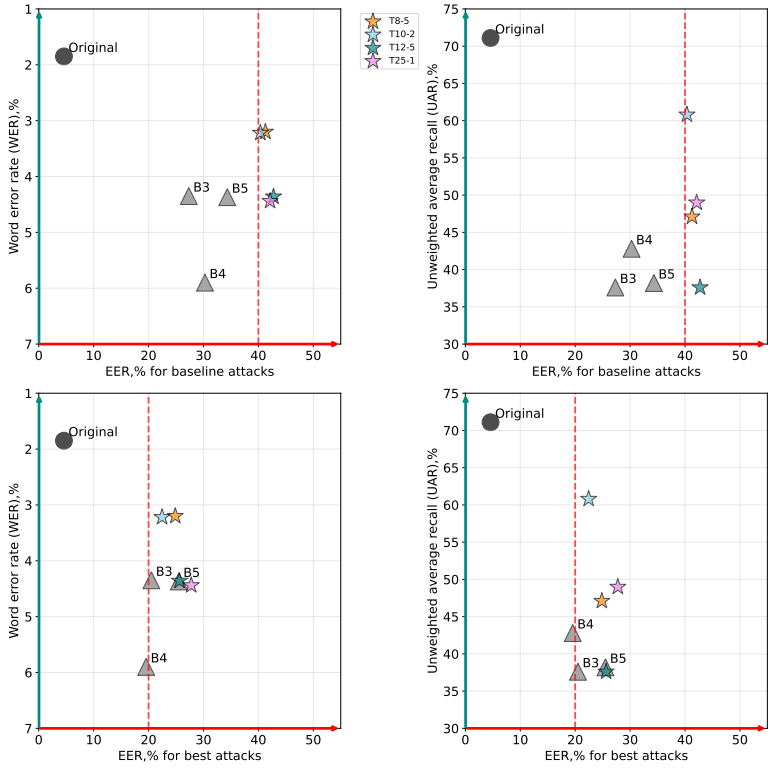


Figure 7: Differences in performance rankings of anonymization systems based on utility metrics (WER, UAR) and privacy (EER) for baseline (top) and best attack (bottom) models on the test dataset.

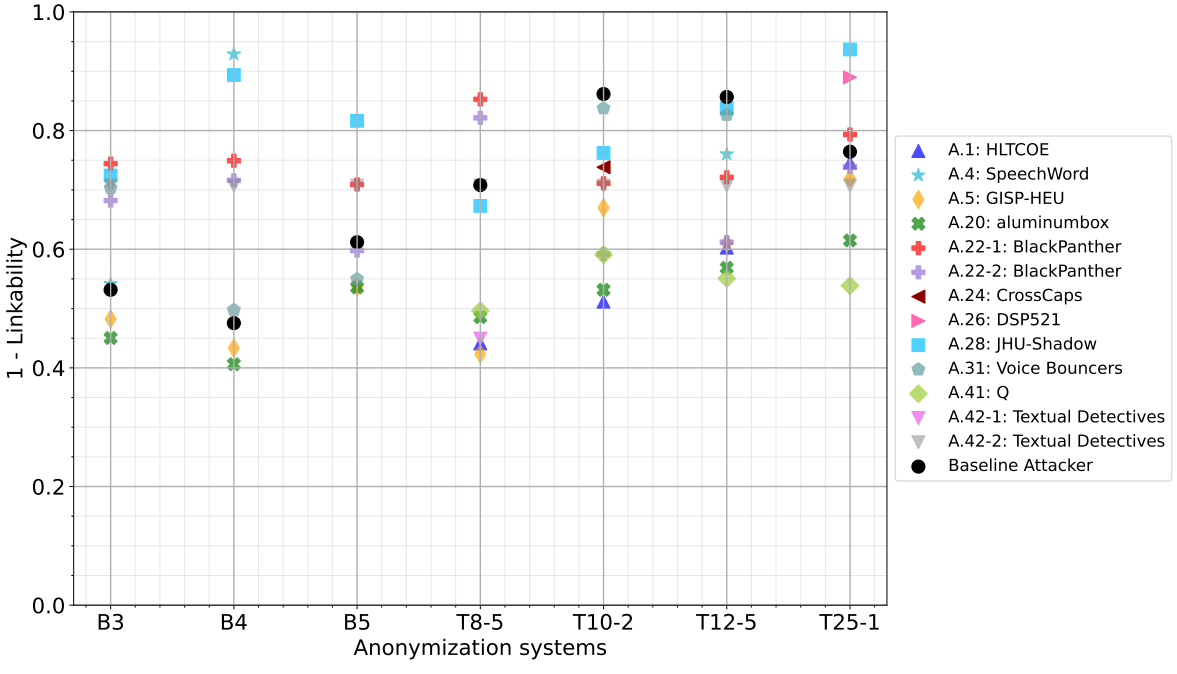


Figure 8: Linkability results on the test dataset.

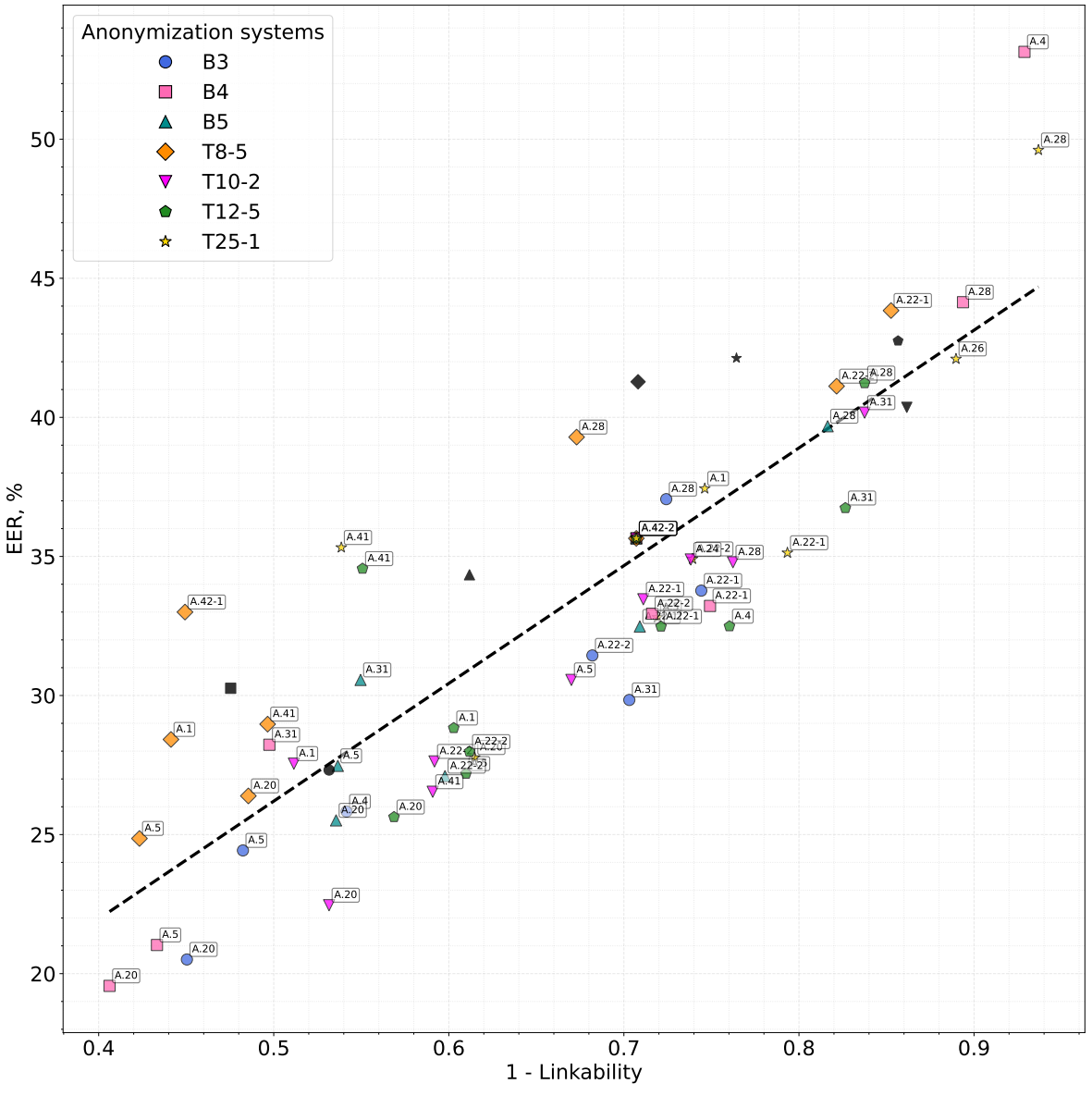


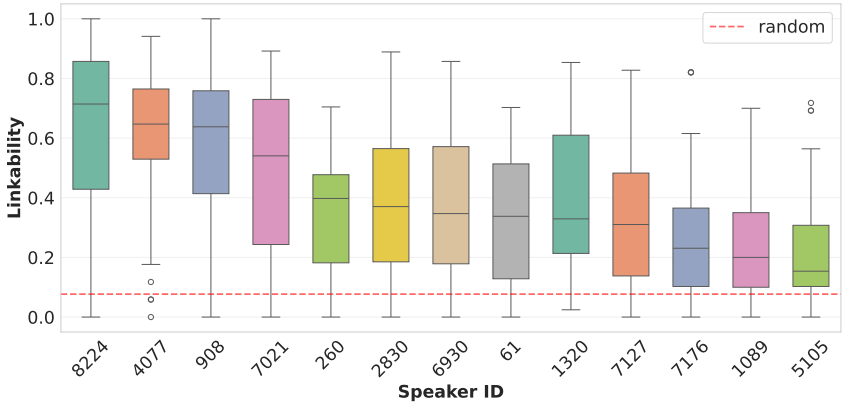
Figure 9: EER vs (1-Linkability) results on the test dataset. Black marks correspond to the baseline attacks.

anonymization system **T8-5**, attack models **A.42-1** and **A.1** yield very high linkability scores (0.55-0.56) while still achieving relatively high EER (28%-33%). This apparent inconsistency can be attributed to the specific AdMixture design of system **T8-5** and the way the targeted attack models exploit its properties.

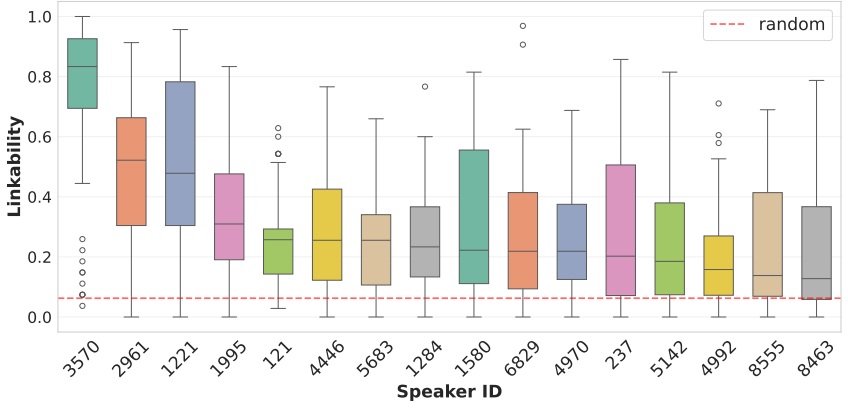
B.4. Speaker-level linkability analysis

Per-speaker linkability distributions over all 68 attacker–anonymization system combinations are shown as boxplots in Figure 10a for male speakers and in Figure 10a for female speakers. Linkability performance varies substantially across speakers: some remain highly linkable across systems and attackers, while others are much harder to link consistently. For example, some female speakers, such as 3570, 2961, and 6829, and some male speakers, such as 8224, 4077, and 908, show consistently high linkability across many attacker–anonymization system setups.

The between-speaker heterogeneity in median linkability is substantial. High-risk speakers, particularly 3570, 2961, and 6829, have median linkability values consistently in the upper half of the range, whereas lower-risk speakers



(a) Linkability for male speakers.



(b) Linkability for female speakers.

Figure 10: Linkability distributions for male (a) and female (b) speakers across all anonymization and attack systems. The red lines indicate chance-level performance ($\frac{1}{13}$ for males, $\frac{1}{16}$ for females). Boxplots show median, quartiles, and outliers. High-risk speakers (e.g., 8224, 3570) show consistently elevated median linkability, while low-risk speakers (e.g., 1089, 8555) maintain protection across diverse attacks.

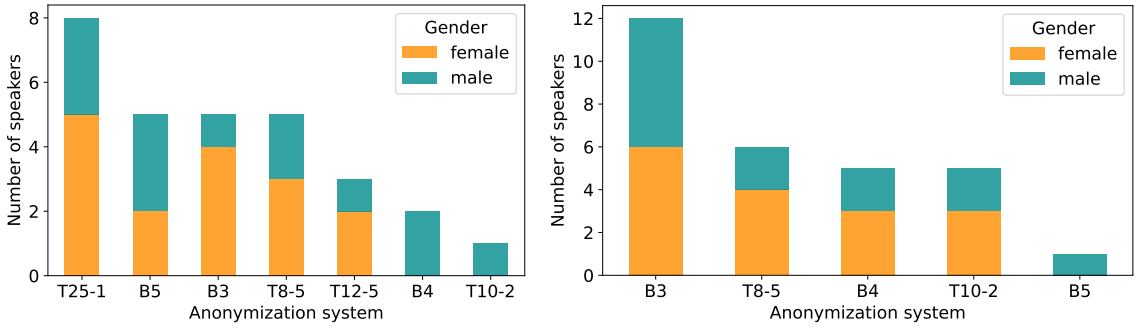


Figure 11: Distribution of speakers for which each anonymization system provides (a) optimal (lowest linkability under worst-case attack) or (b) worst-case (highest linkability) protection. System **T25-1** is optimal for 28% of speakers, while **B3** represents the worst choice for 41% of speakers.

such as 121 and 1284 are near 0.3–0.5. This wide separation indicates that speaker identity alone is a strong predictor of anonymization failure.

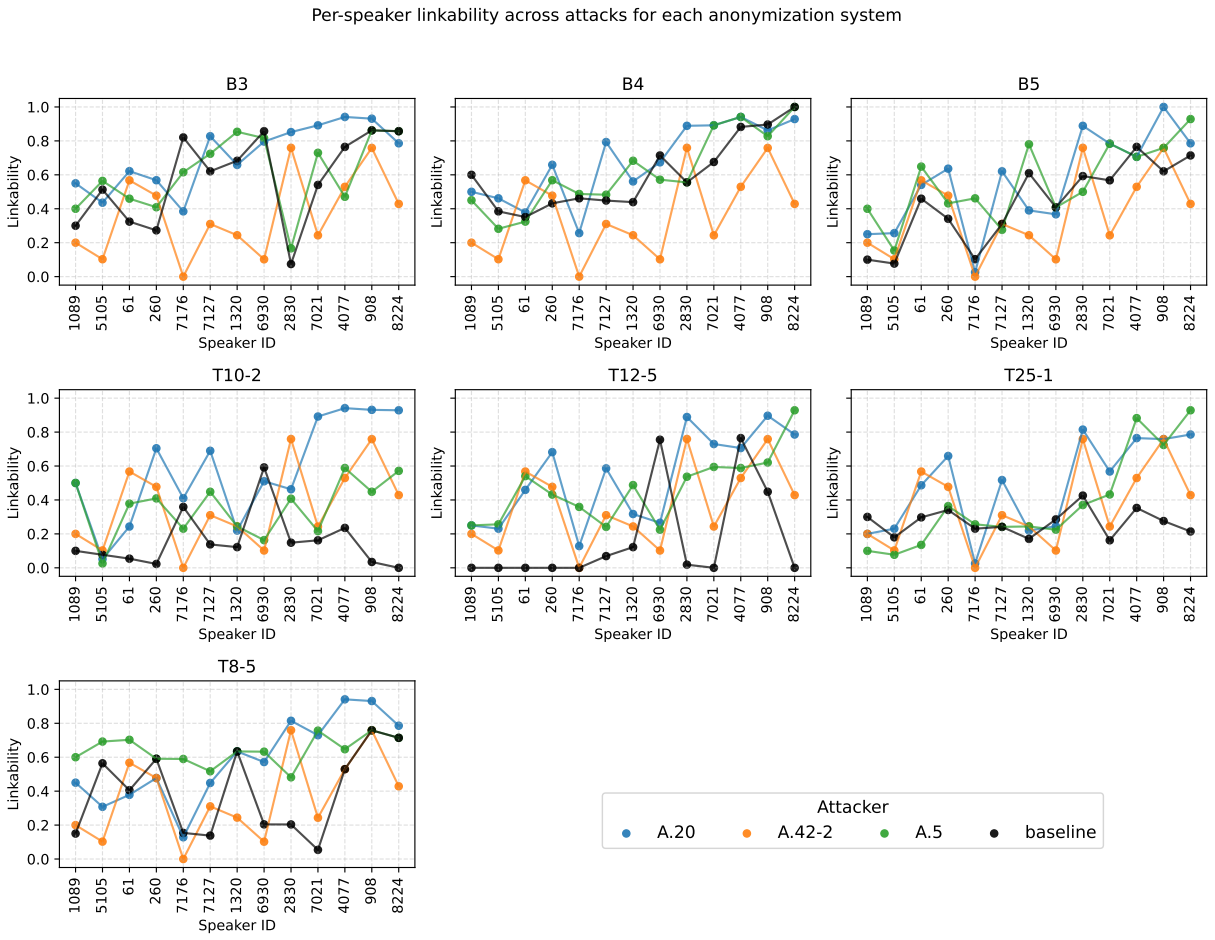


Figure 12: Per-speaker linkability scores for male test speakers across 4 selected attacks for every anonymization system.

Figure 11 illustrates how optimal and worst-case anonymization systems are distributed across the speaker population.

Figure 12 shows per-speaker linkability scores for male speakers across four selected attackers (the two strongest, **A.20** and **A.5**, the text-based attack **A.42-2**, and the baseline attacks) for every anonymization system.