The Third VoicePrivacy Challenge: Preserving Emotional Expressiveness and Linguistic Content in Voice Anonymization

Natalia Tomashenko, Miao Xiaoxiao, Champion Pierre, Meyer Sarina, Panariello Michele, Wang Xin, Evans Nicholas, Vincent Emmanuel, Yamagishi Junichi, Todisco Massimiliano

The Third VoicePrivacy Challenge: Preserving Emotional Expressiveness and Linguistic Content in Voice Anonymization

Natalia Tomashenko[a,*], Xiaoxiao Miao[b], Pierre Champion[c], Sarina Meyer[d], Michele Panariello[e], Xin Wang[f], Nicholas Evans[e], Emmanuel Vincent[a], Junichi Yamagishi[f] and Massimiliano Todisco[e]

[a] *Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France*
[b] *Duke Kunshan University, Kunshan, 215316, China*
[c] *Mission Défense et Sécurité, Inria, Grenoble, France*
[d] *Institute for Natural Language Processing, University of Stuttgart, Germany*
[e] *Audio Security and Privacy Group, EURECOM, Biot, Sophia Antipolis, F-06904, France*
[f] *National Institute of Informatics, Chiyoda-ku, Tokyo, 101-8340, Japan*

ARTICLE INFO

Keywords:
Privacy
Anonymization
Attack model
VoicePrivacy challenge
Speaker verification
Speech recognition
Emotion recognition
Voice conversion
Utility
Evaluation metrics

Abstract

We present results and analyses from the third VoicePrivacy Challenge held in 2024, which focuses on advancing voice anonymization technologies. The task was to develop a voice anonymization system for speech data that conceals a speaker's voice identity while preserving linguistic content and emotional state. We provide a systematic overview of the challenge framework, including detailed descriptions of the anonymization task and datasets used for both system development and evaluation. We outline the attack model and objective evaluation metrics for assessing privacy protection (concealing speaker voice identity) and utility (content and emotional state preservation). We describe six baseline anonymization systems and summarize the innovative approaches developed by challenge participants. Finally, we provide key insights and observations to guide the design of future VoicePrivacy challenges and identify promising directions for voice anonymization research.

1. Introduction

Speech data fall within the scope of privacy regulations such as the European General Data Protection Regulation (GDPR). This is because recordings of speech capture a wealth of personal and sensitive information such as the speaker's identity, age and gender, health status, personality, racial or ethnic origin, geographical background, social identity, and socio-economic status (Nautsch et al., 2019). The VoicePrivacy initiative and challenge series (Tomashenko et al., 2020b) were formed in 2020 to spearhead a community effort to develop privacy preservation solutions for speech technology.

Privacy preservation requires a combination of complementary solutions to obfuscate not only the speaker's identity, but also specific linguistic content, extra-linguistic traits, and background sounds which might also reveal the speaker's identity or other sensitive information. Thus far, the focus in VoicePrivacy has been upon the development of *voice anonymization* solutions which disguise the voice identity while preserving a set of additional attributes related to one or more downstream task (e.g. automatic speech recognition). Common datasets and experimental protocols have been identified and designed to support the fair benchmarking of competing anonymization solutions using a set of evaluation metrics. The first two editions of VoicePrivacy were held in 2020 (Tomashenko et al., 2020b, 2022c,

*Corresponding author
natalia.tomashenko@inria.fr (N. Tomashenko); xiaoxiao.miao@dukekunshan.edu.cn (X. Miao); pierre.champion@inria.fr (P. Champion); sarina.meyer@ims.uni-stuttgart.de (S. Meyer); michele.panariello@eurecom.fr (M. Panariello); wangxin@nii.ac.jp (X. Wang); evans@eurecom.fr (N. Evans); emmanuel.vincent@inria.fr (E. Vincent); jyamagis@nii.ac.jp (J. Yamagishi); massimiliano.todisco@eurecom.fr (M. Todisco)
https://sites.google.com/view/natalia-tomashenko (N. Tomashenko); https://xiaoxiaomiao323.github.io/ (X. Miao); https://www.eurecom.fr/en/people/evans-nicholas (N. Evans); https://members.loria.fr/EVincent/ (E. Vincent); https://yamagishilab.jp/ (J. Yamagishi); https://www.eurecom.fr/en/people/todisco-massimiliano (M. Todisco)
ORCID(s): 0000-0002-7125-2382 (N. Tomashenko); 0000-0002-6645-6524 (X. Miao); 0000-0003-1225-7957 (P. Champion); 0009-0004-1117-4783 (S. Meyer); 0009-0007-4154-5460 (M. Panariello); 0000-0001-8246-0606 (X. Wang); 0000-0002-8459-1041 (N. Evans); 0000-0002-0183-7289 (E. Vincent); 0000-0003-2752-3955 (J. Yamagishi); 0000-0003-2883-0324 (M. Todisco)

2021, 2020a, 2022b) and 2022 (Tomashenko et al., 2022a,b; Panariello et al., 2024b), resulting in substantial progress and the forming of a new research community in voice anonymization.

In this paper we describe the most recent edition of the VoicePrivacy Challenge (VPC) which concluded in 2024. In keeping with previous editions, the 2024 edition maintained the focus on voice anonymization and the preservation of linguistic content and paralinguistic attributes. New to the 2024 edition was a specific requirement to preserve the speaker's emotional state, a key paralinguistic attribute in many real-world application scenarios for voice anonymization, e.g., in call centers to enable the use of third-party speech analytics.[1]

Placing emphasis on preserving emotional attributes in voice anonymization raises the inherent difficulty of the task; it requires concealing the voice identity while maintaining an increasingly rich set of linguistic and paralinguistic characteristics. Successive challenge editions should therefore aim to broaden – and even shift – their scope, with a more explicit focus not only on what must be suppressed or obfuscated, but equally on what should be preserved after anonymization. Introducing such balanced and progressively demanding constraints can serve as a mechanism to stimulate scientific progress, advancing speech representation disentanglement, and promoting more robust and principled approaches to voice anonymization. The new focus also demands the adoption of additional emotion-labeled databases and experimental protocols, as well as new metrics to support the evaluation of emotion preservation.

The differences among the three VoicePrivacy Challenge editions are summarized in Table 1. It highlights the progressive evolution of objectives, attacker strength, evaluation strategies, and methodological diversity from 2020 to 2024: VPC 2020 focused on concealing speaker voice identity (under various attack scenarios: *ignorant*, *lazy-informed*, *semi-informed*) while preserving linguistic content, naturalness, and voice distinctiveness as a complementary objective. VPC 2022 strengthened defenses against *semi-informed* attackers and introduced intonation preservation; and VPC 2024 extended requirements to preserve both linguistic and emotional expressiveness. This progression reflects increasingly advanced attacker models, an expanding range of evaluation metrics from basic utility measures to prosody and emotion preservation, and significant growth in diversity of anonymization methods — with submitted systems evolving from incremental improvements over baselines to using neural codecs, advanced attribute disentanglement methods, and hybrid approaches that flexibly balance privacy-utility tradeoffs. These developments underscore the field's advancing demands for robust, multi-dimensional voice privacy solutions.

In this paper we describe the VPC 2024, present the challenge task, attack model, protocols, anonymization baselines and evaluation metrics. Furthermore, we provide a summary of the 36 anonymization systems designed by challenge participants, extending the analyses presented at the VoicePrivacy 2024 workshop held in conjunction with the 4th Symposium on Security and Privacy in Speech Communication (SPSC),[2] a joint event co-located with Interspeech 2024. We present new comparisons to systems that emerged from previous VPC editions, with additional insights into limitations and the reliability of privacy evaluation in the face of stronger attacks. We conclude with an overview of post-challenge results, and our perspectives on future research and likely directions for the next VPC edition tentatively scheduled for 2026.

2. Challenge design

In this section, we present an overview of the challenge setup: the anonymization task, attack model, and data.

2.1. Voice anonymization task

Privacy protection is formulated as a game between a *user* who shares data with the purpose of accomplishing some desired downstream task and an *attacker* who accesses and uses this data, or data derived from it, to infer information about the data subject (Qian et al., 2018; Srivastava et al., 2020b; Tomashenko et al., 2020b). Here, we consider the scenario where the user shares anonymized utterances for downstream automatic speech recognition (ASR) and speech emotion recognition (SER) tasks, whereas the attacker seeks to identify the speaker using the same anonymized utterances (see Figure 1).

Utterances shared by the user are referred to as *trial* utterances. In order to hide his or her identity, the user treats each utterance with a voice anonymization system prior to sharing. The resulting utterance sounds as if it was uttered by another speaker, referred to as a *pseudo-speaker*. The pseudo-speaker might, for instance, be an artificial voice not corresponding to the voice of any real individual. The task of challenge participants is to develop this voice

[1] This requirement reflects a selective anonymization approach: while speaker identity cues are suppressed to ensure GDPR compliance, certain paralinguistic attributes such as emotion are deliberately preserved to support downstream utility. Preserving emotion is generally considered compatible with GDPR provided that it cannot be linked back to the individual or reveal sensitive health information.

[2] https://spsc-symposium.de/2024

Table 1
Comparison of VoicePrivacy Challenges (2020–2024)

VoicePrivacy 2020	VoicePrivacy 2022	VoicePrivacy 2024
Task		
remove speaker id keep linguistic content keep naturalness & intelligibility	remove speaker id keep linguistic content **keep intonation** keep naturalness & intelligibility	remove speaker id keep linguistic content **keep emotional state**
Training data		
LibriSpeech: train-clean-100, train-other-500 LibriTTS: train-clean-100, train-other-500 VoxCeleb-1,2	LibriSpeech: train-clean-100, train-other-500 LibriTTS: train-clean-100, train-other-500 VoxCeleb-1,2	**open resources (data, models)** **(based on participant proposals)**
Development and evaluation data		
LibriSpeech dev-/test-clean VCTK	LibriSpeech dev-/test-clean VCTK	LibriSpeech dev-/test-clean **IEMOCAP**
Anonymization level of the test data		
speaker	speaker	**utterance**
Objective assessment		
Attack models		
ignorant, *lazy-informed*, *semi-informed*	***semi-informed***	*semi-informed*
Privacy metrics		
Primary: equal error rate (EER), Post-eval: C_{llr}, $C_{\text{llr}}^{\text{min}}$, linkability, ZEBRA (Nautsch et al., 2020), de-identification (DeID) (Noé et al., 2020, 2022)	EER	EER
Utility metrics		
WER (ASR trained on orig. & anon. data) Gain of voice distinctiveness (G_{VD})	WER (ASR trained on **anonymized** data) Gain of voice distinctiveness (G_{VD}) **Pitch correlation** ($\rho_{F0} > 0.3$)	WER (ASR trained on **original** data) **UAR (SER trained on original data)**
Subjective assessment		
Privacy metrics		
Primary: speaker verifiability Post-eval: linkability (O'Brien et al., 2021)	speaker verifiability	—
Utility metrics		
naturalness & intelligibility	naturalness & intelligibility	—
Ranking policy		
—	**ranking by utility within privacy categories (EER $\geq$ 15%, 20%, 25%, 30%)**	ranking by utility within privacy categories (**EER $\geq$ 10%, 20%, 30%, 40%**)
Baselines		
– *Neural voice conversion:* B1 (x-vector + neural vocoder) – *Signal processing:* B2 (McAdams fixed coeff.)	– *Neural voice conversion:* **B1.a, B1.b** – *Signal processing:* B2 (McAdams **random** coeff.)	B1 (B1.b from VPC2022) B2 (McAdams random coeff.) **B3 (phonetic transcriptions + TTS and GAN)** **B4 (neural codecs)** **B5, B6 (ASR-BN with VQ)**
Submitted systems		
– *Neural voice conversion:* Attribute disentanglement (focus on x-vector anonymization, mainly incremental improvements over the baselines) – *Signal processing:* Formant, F0, and speaking rate modification	– *Neural voice conversion:* Attribute disentanglement (**modify all components in baseline**: content, speaker, and prosody feature extraction; speaker and prosody anonymizers; speech generation models) – *Signal processing:* Formant, pitch, and speaking rate modification	– *Neural voice conversion:* Attribute disentanglement (speaker, linguistic, pitch, **emotion**; **multiple emotion preservation strategies** (**pretrained encoders**, **distillation**); **kNN-VC** – ***Neural codecs*** – ***ASR+TTS cascades*** – ***Hybrid methods***

anonymization system which should (a) output a speech waveform; (b) conceal the voice identity at the *utterance level*; (c) preserve the linguistic and emotional content.

Requirement (b) for *utterance-level* anonymization means that the voice anonymization system must replace the voice identity with that of a pseudo-speaker independently of any other utterance. The pseudo-speaker assignment process must be identical across all utterances and not rely on speaker labels. When this process involves use of a random number generator, the random number(s) generated for each utterance must be different, thereby resulting in a different pseudo-speaker for each utterance. Alternatively, a single pseudo-speaker may be assigned to *all* utterances to satisfy this requirement. The achievement of requirement (c) is assessed via *utility* metrics. Specifically, we measure

Table 2
Statistics of the *LibriSpeech* development and evaluation sets for ASV and ASR evaluation.

	Subset	Type	#Speakers			#Utterances
			Female	Male	Total	
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.

the preservation of linguistic and emotional content using the word error rate (WER) and unweighted average recall (UAR) using automatic speech recognition (ASR) and speech emotion recognition (SER) systems, both trained using original (unprotected) data.

2.2. Attack model

For each speaker of interest, the attacker is assumed to gain access to utterances spoken by that speaker, referred to as *enrollment* utterances. Using these utterances, attackers deploy automatic speaker verification (ASV) to re-identify the speaker whose voice is contained in each anonymized *trial* utterance.

We assume that the attacker has access to: (a) several enrollment utterances for each speaker and (b) the same voice anonymization system employed by the user. Using these resources, the attacker anonymizes the enrollment utterances to reduce the mismatch with the trial utterances, and trains an ASV system adapted to that specific anonymization system. Like the user, the attacker performs voice anonymization at the *utterance level*. This attack model was, to date, conceptually the strongest known; therefore, we considered it for privacy assessment.

2.3. Data and pretrained models

The training, development and evaluation of voice anonymization systems are all performed using publicly available resources. The development and evaluation data are fixed, while the choice of training resources is flexible, subject to the constraints outlined below.

2.3.1. Training resources

In addition to the training data used for previous challenge editions (*VoxCeleb-1,2*; *LibriSpeech train-clean-100*, *train-other-500*; *LibriTTS train-clean-100*, *train-other-500*) as well as data and pretrained models used to build and develop baseline anonymization systems (corpora *RAVDESS*, *ESD*; models *HuBERT Base*, *EnCodec*, *Bark*, and *wav2vec2_large_west_germanic_v2*) — see Section 3.4 for more details — participants were permitted to propose additional training resources, including both data and pretrained models. From the suggestions received, a list of permitted training data and pretrained models was finalized and published in an update to the evaluation plan (Tomashenko et al., 2024a). The list is shown in Table 8 in Appendix A.

2.3.2. Development and evaluation data

Development and evaluation datasets come from two corpora: *LibriSpeech*[3] (Panayotov et al., 2015) and *IEMOCAP* (Busso et al., 2008), as presented in Tables 2 and 3. *LibriSpeech* contains 960 hours of read English speech (16 kHz) in the form of audiobooks and is used commonly for ASV and ASR evaluation. The *LibriSpeech* evaluation and development sets are the same as for previous challenge editions. *IEMOCAP* contains 12h of emotional speech (16 kHz) collected from two-speaker conversations between 10 English actors (5 female, 5 male). Following Pappagari et al. (2020), we focus on a subset of 4 emotions for SER evaluation: *neutral*, *sadness*, *anger*, and *happiness* (with *happiness* and *excitement* classes merged). Due to data limitations, we implement leave-one-conversation-out cross-validation, using four conversations (8 speakers) for SER training and the remaining one conversation (2 speakers) for development and evaluation.

[3]LibriSpeech: `http://www.openslr.org/12`

Table 3
Construction and statistics of the *IEMOCAP* development and evaluation sets for SER evaluation. *Train* subsets refer to the training data for the SER evaluation system.

Conversation		#Utterances	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Session 1	Female	528	Dev	Train	Train	Train	Train
	Male	557	Eval				
Session 2	Female	481	Train	Eval			
	Male	542		Dev			
Session 3	Female	522		Train	Dev		
	Male	629			Eval		
Session 4	Female	528			Train	Eval	
	Male	503				Dev	
Session 5	Female	590				Train	Eval
	Male	651					Dev

Table 4
Statistics of the *LibriSpeech* training sets for ASV and ASR evaluation models.

System	Subset	Size,h	#Speakers			#Utterances
			Female	Male	Total	
ASV	LibriSpeech train-clean-360	363.6	439	482	921	104,014
ASR	LibriSpeech train-960	960.9	1128	1210	2338	281,241

3. Privacy and utility evaluation

Three metrics are used for the objective ranking of submitted systems: the equal error rate (EER) as the privacy metric and two utility metrics: word error rate (WER) and unweighted average recall (UAR). These metrics rely on automatic speaker verification (ASV), automatic speech recognition (ASR), and speech emotion recognition (SER) systems.

3.1. Training data for evaluation models

The ASV system is trained on *LibriSpeech-train-clean-360* and the ASR system on the full *LibriSpeech-train-960* dataset, whose statistics are presented in Table 4. The SER system for each *IEMOCAP* cross-validation fold is trained on the corresponding *IEMOCAP* training subset, whose statistics are reported in Table 3. Training and evaluation are performed with the provided recipes and models.[4] Models for privacy evaluation are trained by participants on their anonymized training data with the provided training scripts, while the models for utility evaluation are provided by the organizers.

3.2. Privacy metric: the equal error rate (EER)

The ASV system used for privacy[5] evaluation is an ECAPA-TDNN model (Desplanques et al., 2020) with 512 channels in the convolution frame layers, implemented by adapting the *SpeechBrain* (Ravanelli et al., 2021) recipe for the *VoxCeleb* dataset to the *LibriSpeech* dataset. As illustrated in Figure 1, the attacker is assumed to have access

[4]Evaluation scripts: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2024

[5]In this context, *privacy* refers specifically to speaker de-identification (measured by ASV EER increase), while *utility* denotes content and emotion preservation (ASR WER, SER UAR); together they constitute *anonymisation performance*. More broadly, *privacy* encompasses the protection of all personal and sensitive information that could identify or link an individual, including metadata, contextual details, and behavioral patterns beyond voice characteristics. Therefore, even perfect speaker de-identification may not guarantee complete privacy if other identifiable information, such as a speaker's name, location, or interaction patterns, is exposed or disclosed.

Table 5
Number of *same-speaker* and *different-speaker* pairs 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

to the anonymization system under evaluation (Srivastava et al., 2020b; Tomashenko et al., 2022c). The attacker uses the system to anonymize the enrollment utterance(s) so as to reduce the mismatch between enrollment and trial utterances. In addition, the attacker uses the same system to anonymize the *LibriSpeech-train-clean-360* dataset and retrains the ASV system (denoted $ASV_{\text{eval}}^{\text{anon}}$) so that it is adapted to operate upon similarly anonymized input. In all cases, anonymization is applied at the *utterance level*, using the same pseudo-speaker assignment process used for anonymization of trial utterances. For speakers with multiple enrollment utterances, embeddings are extracted for each utterance and then averaged.

The cosine similarity score is computed for each pair of embeddings extracted from enrollment and trial utterances and thresholded to produce same-speaker vs. different-speaker decisions. Denoting the ASV false alarm and miss rates at threshold θ by $P_{\text{fa}}(\theta)$ and $P_{\text{miss}}(\theta)$ respectively, the EER is defined by the decision threshold θ_{EER} for which the two detection error rates are equal, i.e., $\text{EER} = P_{\text{fa}}(\theta_{\text{EER}}) = P_{\text{miss}}(\theta_{\text{EER}})$. The higher the EER, the better the anonymization. The number of same-speaker and different-speaker enrollment/trial pairs used to estimate the EER is given in Table 5 for the development and evaluation partitions of each dataset.

3.3. Utility metrics

Whereas for the 2022 VPC there were three utility metrics, there are only two for the 2024 edition. While the measure of linguistic content preservation remains unchanged, measures of pitch correlation and gain of voice distinctiveness were replaced with a measure of emotion preservation. Measures of pitch correlation were deemed to be insufficiently informative or reliable, while the gain of voice distinctiveness pertains to an unnecessarily narrow use case scenario. A speaker's emotional state is a key paralinguistic attribute that is relevant to a broader range of application scenarios.

3.3.1. Word error rate (WER)

The preservation of linguistic content is assessed using an ASR system[6] (denoted ASR_{eval}) fine-tuned using the *LibriSpeech-train-960* dataset and *wav2vec2-large-960h-lv60-self*[7] representations using a *SpeechBrain* recipe. The ASR evaluation model is trained using original (unprocessed) data. In contrast to previous VPC editions, however, it is then fixed rather than being adapted to anonymized data.

The motivation for using original data to train the ASR for utility estimation is to minimize acoustic distortions and better approximate the natural speech distribution, enabling utility estimates that reflect real-world usability and listener experience with anonymized speech. The broader perspective involves using anonymized data for training or fine-tuning downstream models that can perform well on both original and anonymized speech.

For every utterance in the *LibriSpeech* development and evaluation sets, the ASR system outputs a word sequence. The WER, estimated using either original or anonymized data, is calculated according to

$$\text{WER} = \frac{N_{\text{sub}} + N_{\text{del}} + N_{\text{ins}}}{N_{\text{ref}}},$$

where N_{sub}, N_{del}, and N_{ins} are the number of substitution, deletion, and insertion errors, respectively, and N_{ref} is the number of words in the reference. The lower the WER, the better the preservation of linguistic content, and hence the greater the utility.

[6] https://huggingface.co/speechbrain/asr-wav2vec2-librispeech
[7] https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self

3.3.2. Unweighted average recall (UAR)

The preservation of emotional content is assessed using an SER system (denoted SER_{eval}) trained using the *SpeechBrain* recipe for the *IEMOCAP* database. The wav2vec2-based model is trained separately for each of the training folds shown in Table 3. SER performance is measured using the *IEMOCAP* development and evaluation sets using either original or anonymized utterances. We calculate the unweighted average recall (UAR) according to

$$\text{UAR} = \frac{\sum_{i=1}^{N_{\text{class}}} R_i}{N_{\text{class}}}.$$

where the recall R_i for each class i is computed as the number of true positives divided by the total number of samples in that class and where $N_{\text{class}} = 4$ is the number of classes. UARs are averaged across the five folds. The higher the UAR, the better the preservation of emotional state and the greater the utility. Unweighted averaging (in UAR) rather than weighted averaging is used to prevent high-frequency emotion classes from dominating the metric, which is particularly important given class imbalance in *IEMOCAP*.

3.4. Privacy-utility tradeoff

As in the 2022 edition, multiple evaluation conditions (privacy categories) specified with a set of minimum target privacy requirements are considered. For each minimum target privacy requirement, submissions that meet this requirement are ranked according to the resulting utility for each utility metric separately. The goal is to measure the privacy-utility trade-off at multiple operating points, e.g. when systems are configured to offer better privacy at the cost of utility and vice versa. This evaluation framework better aligns with real-world user privacy expectations and provides participants with clear optimization targets, creating a more comprehensive assessment using our established privacy and utility metrics. Privacy requirements are defined by N minimum target EERs: $\{\text{EER}_1, \ldots, \text{EER}_N\}$. Each minimum target EER constitutes a separate evaluation category. Submissions to category i had to achieve an average EER on the VoicePrivacy 2024 evaluation set greater than EER_i. The challenge includes $N = 4$ categories with minimum target EERs of 10%, 20%, 30%, and 40%. Systems are ranked by WER (lower is better) and UAR (higher is better) within each privacy (EER) category. A depiction of example results and system rankings according to this methodology is shown in Figure 2.

4. Baseline voice anonymization systems

In this section we provide an overview of the baseline voice anonymization systems which were made available to participants as inspiration for developing their own solutions. To gauge progress, we included two baseline systems (**B1** and **B2**) established with previous challenge editions. We also made four new baseline systems available, namely (**B3**, **B4**, **B5**, and **B6**), which offer enhanced anonymization performance with varying compromises to utility. New baseline systems are trained using different resources. A description of each baseline is provided in the following. Further details can be found in the evaluation plan (Tomashenko et al., 2024a) and in the publicly available baseline recipe scripts[8].

4.1. Anonymization using x-vectors and a neural source-filter model: B1

The **B1** baseline anonymization system (Srivastava et al., 2020a, 2022), shown in Figure 3, is based on the VPC 2022 **B1.b** system, but applies anonymization at the *utterance level*. The system first extracts three features: 256-dimensional bottleneck (BN) features for linguistic content, fundamental frequency (F0), and 512-dimensional speaker x-vectors. BN features are obtained from a factorized time delay neural network (TDNN-F) (Povey et al., 2018) model trained on *LibriSpeech (train-clean-100, train-other-500)*, while x-vectors are from a TDNN trained on *VoxCeleb-1,2*. An anonymized x-vector is then computed by averaging a random subset of 100 among the 200 farthest candidate x-vectors from a disjoint speaker pool (*LibriTTS train-other-500*). Finally, the anonymized speech is synthesized with the anonymized x-vector, BN, and F0 features using a neural source-filter (NSF) model trained on *LibriTTS train-clean-100* with a HiFi-GAN (Kong et al., 2020) architecture.

[8]The VoicePrivacy 2024 Challenge anonymization baselines and evaluation scripts: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2024

4.2. Anonymization using the McAdams coefficient: B2

The **B2** baseline anonymization system (Patino et al., 2021) uses only signal processing techniques and requires no use of training data. It employs the McAdams coefficient (McAdams, 1984) to shift pole positions derived from linear predictive coding (LPC) analysis of input speech signals. Frame-by-frame LPC analysis is used to extract coefficients and residuals. LPC coefficients are converted to z-plane pole positions, with each pole representing a spectral peak corresponding to a formant position. The McAdams' transformation is used to modify only poles with non-zero imaginary parts by raising their phase ϕ (0 to π radians) to the power of a coefficient α, resulting in shifted phases of ϕ^α. For each utterance, α is sampled from $U(0.5, 0.9)$, causing contraction or expansion around $\phi = 1$ radian ($\approx$2.5 kHz). Complex conjugate poles are shifted in opposite directions. The full set is converted back to LPC coefficients, which are combined with residuals to synthesize a frame of anonymized speech.

4.3. Anonymization using phonetic transcriptions and generative adversarial network: B3

Baseline **B3**, shown in Figure 4, is a speech synthesis system conditioned to retain linguistic and general prosodic information while substituting the speaker voice (Meyer et al., 2023a,b). At its core is a generative adversarial network (GAN) that generates an artificial pseudo-speaker embedding (Meyer et al., 2023c). **B3** performs speech anonymization through a three-step process. First, speaker embeddings, phonetic transcriptions, the fundamental frequency (F0), energy, and phone duration are extracted from the input. The speaker embedding is extracted using an adapted global style tokens model (Wang et al., 2018) trained using the *LibriTTS train-clean-100* dataset. The phonetic transcription is obtained using an end-to-end ASR model with a hybrid CTC-Attention architecture, a Branchformer (Peng et al., 2022) encoder and a Transformer decoder trained using the *LibriTTS train-clean-100* and *train-other-500* datasets. Second, anonymization is performed by replacing the original speaker embedding with an artificial embedding generated using a Wasserstein GAN (Arjovsky et al., 2017) ensuring that the cosine distance between the artificial and original embeddings exceeds 0.3. The GAN model is trained using *LibriTTS train-clean-100*, *RAVDESS* (Livingstone and Russo, 2018), and *ESD* (Zhou et al., 2021) datasets. Additionally, the F0 and energy values of each phone are multiplied by random values sampled uniformly and independently from $U(0.6, 1.4)$ to remove speaker-specific prosodic patterns while still retaining broader prosody dynamics. Random values are chosen for each phone individually. Finally, the anonymized speaker embedding, modified F0 and energy features, original phonetic transcripts and original phone durations are fed into a speech synthesis system based on *FastSpeech2* (Ren et al., 2020) and a HiFi-GAN (Kong et al., 2020), as implemented in the *IMS-Toucan* Toolkit (Lux et al., 2021), both trained using LibriTTS train-clean-100, to synthesize an anonymized output.

4.4. Anonymization using neural audio codec language modeling: B4

Baseline **B4** (Panariello et al., 2024a) illustrated in Figure 5, is based upon the technique of neural audio codec (NAC) language modeling (Borsos et al., 2023; Wang et al., 2023). A NAC is an encoder-decoder neural network using which an input signal can be encoded into a sequence of discrete *acoustic tokens* $\mathbf{a} \in \{1, \dots, N_Q\}^{Q \times T_A}$ and subsequently decoded to a waveform. T_A is the number of time frames into which the input is segmented, and Q is the number of tokens associated to each time frame, each drawn from one of Q different token dictionaries. Tokens are integers ranging from 1 to N_Q. Acoustic tokens are assumed to capture the characteristics of an individual's speech. Several sets of acoustic tokens are extracted from the speech of a pool of pseudo-speakers, obtaining a *pool of acoustic prompts* $\mathbf{A}$. A *semantic extractor* is used to extract a sequence of discrete *semantic tokens* $\mathbf{s} \in \{1, \dots, N_S\}^{T_S}$ which encode the spoken content, where T_S is the number of time frames. They are concatenated with a randomly chosen sequence of acoustic tokens $\tilde{\mathbf{a}} \in \mathbf{A}$ to form sequence $(\mathbf{s}, \tilde{\mathbf{a}})$. A GPT-like, decoder-only Transformer uses this sequence as a prompt to generate acoustic tokens $\mathbf{a}$ that reflect the semantics encoded in $\mathbf{s}$ and style encoded in $\tilde{\mathbf{a}}$. The NAC decoder module then converts $\mathbf{a}$ to a waveform containing the semantic content of the input, but the voice of a different pseudo-speaker. The NAC is *EnCodec*[9] (Défossez et al., 2023), which is trained with speech segments from the DNS Challenge (Dubey et al., 2022) and the *Common Voice* (Ardila et al., 2020) dataset, with additional non-speech segments from the *AudioSet* (Gemmeke et al., 2017), *FSD50K* (Fonseca et al., 2022), and *Jamendo* (Bogdanov et al., 2019) datasets. The semantic extractor is composed of a HuBERT feature extractor (Hsu et al., 2021) trained using the *LibriSpeech-train-960* dataset, and a LSTM back-end that predicts token indices from feature vectors. The decoder-only model is a publicly available checkpoint from the *Bark* TTS model.[10]

[9] https://github.com/facebookresearch/encodec

[10] https://github.com/suno-ai/bark

4.5. Anonymization using ASR-BN with vector quantization (VQ): B5 and B6

The pipeline for **B5** and **B6** baselines, developed by Champion (2023) and illustrated in Figure 6, shares similarities with **B1**. However, it differs by incorporating vector quantization (VQ) to enhance the disentanglement of linguistic and speaker attributes. The pipeline extracts fundamental frequency (F0) and acoustic VQ bottleneck (VQ-BN) features from an ASR acoustic model trained to identify left-biphones. VQ-BN features, combined with F0 features and target speaker representation in the form of a one-hot vector corresponding to a distinct speaker in the training dataset, are then used with a HiFi-GAN network to synthesize an anonymized output. The use of two different ASR AMs corresponds to two different baselines: **B5** for which the AM is a combination of a pretrained wav2vec 2.0 model with three additional TDNN-F layers; and **B6** for which the AM consists of 12 TDNN-F layers. In the final TDNN-F layer of both models, VQ is applied following the first activation (inner bottleneck of the TDNN-F with 256 dimensions). This process approximates a continuous vector with an equivalent-dimensional vector from a finite set. The incorporation of VQ minimizes speaker information encoded within the bottleneck features. **B5** and **B6** baselines employ multiple pre-trained models. The wav2vec 2.0 model is pre-trained using 24.1k hours of unlabeled multilingual West Germanic speech from the VoxPopuli[11] dataset, then fine-tuned using the *LibriSpeech train-clean-100* dataset. The **B6** ASR-BN extractor operates upon Mel filterbank input feature and is trained using *LibriSpeech train-other-500* and *train-clean-100* datasets. Finally, the HiFi-GAN model is trained using *LibriTTS-train-clean-100*, giving 247 one-hot speaker vectors.

5. Submitted anonymization systems

The VPC 2024 attracted more than 100 participants from academic, industrial, and governmental organizations across 16 countries, represented by 40 teams. In total, we received 36 anonymization systems from 13 teams (listed in Table 6) who successfully submitted their results. Each team submitted 1 to 6 systems that targeted different privacy categories and utility performance. The voice anonymization solutions developed by the participants include several technical approaches. The summary of the systems is provided in Table 7. Noncompliance with the VPC 2024 requirements is indicated by a red asterisk (*) and further discussed in the text, where the reasons for noncompliance are highlighted in red.

5.1. Neural voice-conversion (VC)-based anonymization systems

Most submitted anonymization systems are based on neural voice conversion techniques, which focus on explicit attribute disentanglement. They extract separate representations for linguistic content (e.g., bottleneck or self-supervised learning features like *WavLM* or *HuBERT*), speaker identity (x-vectors or ECAPA embeddings), prosody (F0), and emotion. Then, they replace speaker features with pseudo-speaker embeddings while preserving or modifying other attributes, and synthesize the speech signal via vocoders such as HiFi-GAN. Variants include k-nearest neighbors VC (kNN-VC), which performs frame-level nearest-neighbor replacement in *WavLM* feature space, and emotion-enhanced approaches.

5.1.1. k-nearest neighbors voice conversion (kNN-VC)

Two teams, **T8** and **T19**, use k-nearest neighbor voice conversion (kNN-VC), originally proposed by Baas et al. (2023). Instead of using disentangled speaker embeddings and speaker-independent representations, kNN-VC-based methods use a *WavLM* model to extract a single representation sequence from the input utterance and replace it with a 'similar' representation but extracted from the utterances of another speaker selected at random from a pool. The similar representation is the average of the top k nearest representations to the original. The assumption is that the similar representation encodes similar linguistic content to the input but a different speaker identity. An anonymized output is generated using a waveform generator conditioned on the resulting representations.

T8-2 *kNN-VC system on WavLM features* (Xinyuan et al., 2024): A modification of the kNN-VC system which operates on *WavLM* features (Chen et al., 2022). Source and target speaker utterances are converted into *WavLM* feature space. kNN regression is then applied at the frame level, before the output is synthesized using a HiFi-GAN vocoder. The **T8-2** implementation uses $k = 4$ and a vocoder trained using *LibriSpeech train-clean-100*. In pursuit of improved anonymization, target speaker selection is performed at random, drawn from the *VoxCeleb* corpus.

[11] https://dl.fbaipublicfiles.com/voxpopuli/models/wav2vec2_large_west_germanic_v2.pt

Table 6
Participant teams.

Team name (Ref.)	Team ID	Affiliations
Anemone (Hua et al., 2024)	T7	Institute of Acoustics, Chinese Academy of Sciences; University of the Chinese Academy of Sciences, China
JHU CLSP (Xinyuan et al., 2024)	T8	Johns Hopkins University, USA
LongYuan (Tan et al., 2024)	T9	Nanjing University of Posts and Telecommunications; Nanjing Longyuan Information Technology Co. Ltd, China
NPU-NTU (Yao et al., 2024a)	T10	Audio, Speech and Language Processing Group (ASLP@NPU); School of Computer Science, Northwestern Polytechnical University; Nanyang Technological University; The Hong Kong Polytechnic University, China
NTU-NPU (Kuzmin et al., 2024)	T12	Nanyang Technological University, China; Institute for Infocomm Research, A*STAR, Singapore; Audio, Speech and Language Processing Group (ASLP@NPU), China; The Hong Kong Polytechnic University, China
ADRES (Leang et al., 2024)	T14	University Grenoble Alpes, CNRS, France; Institute of Digital Research and Innovation, Phnom Penh, Cambodia
NKU HLT Lab (He et al., 2024)	T17	College of Computer Science, Nankai University, China
Q (Li et al., 2024)	T18	Qifu Technology; Fudan University, Shanghai, China
DFKI_SLT (Das et al., 2024)	T19	Speech and Language Technology, German Research Center for Artificial Intelligence (DFKI); Technische Universität Berlin, Germany
USTC-PolyU (Gu et al., 2024)	T25	NERC-SLIP, University of Science and Technology of China; The Hong Kong Polytechnic University, Hong Kong, China
V-Beam (Lee et al., 2024b,a)	T30	Sogang University, Seoul; Ewha Womans University, Seoul, Korea
KIT-ISL (Akti et al., 2024a,b)	T33	Karlsruhe Institute of Technology (KIT), Germany; Carnegie Mellon University (CMU), USA
Orange_Shiva (Le Blouch et al., 2024a,b)	T38	Orange, France

T19-1 *Single-layer kNN-VC (KNNS)* (Das et al., 2024): Based on kNN-VC using a single-layer *WavLM* feature extraction (6th Transformer layer). The source utterance is processed through *WavLM Large* to extract frame-level features. A random target speaker is selected from the English speakers in the emotional speech database (*ESD*). All utterances of the target speaker are processed by *WavLM* to create a matching set of features. For each source frame, the top 4 nearest neighbors from the target speaker's feature pool are found; their average replaces the original frame. A pretrained HiFi-GAN vocoder reconstructs the audio waveform.

T19-2 *Multi-layer kNN-VC (KNND)* (Das et al., 2024): Extends **T19-1** by using outputs from both the 6th and 12th *WavLM* layers simultaneously. The 12th layer captures emotional cues better, enhancing emotion preservation. Query and matching sets double in size due to this double-layer feature concatenation. HiFi-GAN vocoder is augmented with a convolutional PreNet module, pretrained jointly for this dual-input format.

5.1.2. Attribute disentanglement (linguistic content, speaker, pitch, emotion, etc.)

T7-1,2 *Anonymization using fusion of VC speaker models at the model parameter level* (Hua et al., 2024): Anonymization is performed using the fusion of multiple fine-tuned, single-target speaker end-to-end VC models integrated at the parameter level. Emotion is transferred using two different emotion encoders: **T7-1** employs an emotion encoder based on the *wav2vec 2.0* model which outputs emotion embeddings that added to the acoustic model

as a priori encodings; **T7-2** leverages a global style tokens (GST) encoder (Wang et al., 2018) composed of multi-head Attention modules.

T9 *GMM-blender* (Tan et al., 2024): A *WavLM* model (layer 6 output) is used to extract speech features. The *GMM-Blender* method generates a rich array of anonymous templates through feature interpolation from the anonymization pool. The *GMM-Blender* operates by first randomly selecting M speakers from the *LibriSpeech* corpus to form a dynamic anonymization pool. For each speech utterance to be anonymized, four speakers are randomly selected from this pool, contributing 50 speech samples each (200 samples total). *WavLM* embeddings are extracted from these samples and subjected to Gaussian Mixture Model (GMM) clustering, which identifies 2000 cluster centers. At each cluster center, the four nearest feature embeddings are selected and blended using random weights, creating a composite anonymized embedding template. A wav2vec2-based emotion feature extractor (*wav2vec2-large-robust-12-ft-emotion-msp-dim* (Wagner et al., 2023)) is incorporated to preserve emotional characteristics during anonymization. Emotion features extracted from the source speech are fused with anonymized speech features. This fused representation is then provided to a modified HiFi-GAN vocoder for waveform synthesis. The training objective combines the standard HiFi-GAN adversarial loss with an emotional similarity term that constrains the synthesized speech to maintain the emotional characteristics of the original speech.

T12-5 ***B5****-based anonymization with mean reversion pitch smoothing* (Kuzmin et al., 2024): This system modifies the fundamental frequency (F0) in the **B5** baseline by applying a "mean reversion" mechanism, where the F0 estimate for each frame is interpolated with a moving average of adjacent frames. This reduces pitch variability and smooths the pitch contour to help anonymization.

T12-6 ***B5****-based anonymization with mean reversion pitch smoothing and additive Gaussian noise* (Kuzmin et al., 2024): Similar to **T12-5**, this system additionally applies additive white Gaussian noise (AWGN) explicitly with controlled power to the mean-reverted F0 during inference. This further enhances privacy by increasing difficulty of speaker identification while maintaining reasonable speech utility.

T14-1,2 *VQ variational auto-encoder (VAE) with prosody parameters* (Leang et al., 2024): A custom convolutional module is used to extract content features from the input utterance which are quantized using VQ to disentangle residual speaker information from content information. The log-normalised F0 curve of the input utterance is estimated using YAAPT (Kasi and Zahorian, 2002), then fed into a bidirectional gated recurrent unit (GRU) model (Cho et al., 2014). The voice identity is encoded with an ECAPA-TDNN embedding which is anonymized by means of the well-known pool-based anonymization function (similar to **B1**). A HiFi-GAN is used to synthesize the output. **T14-1** uses standard normalized F0 and energy prosody features extracted from inputs. **T14-2** introduces random scaling of F0 during synthesis to further obscure speaker identity. This scaling uses a random factor in range $[0.8, 1.2]$. The rest of the architecture remains consistent with **T14-1**.

T18-1,2 *Emotion-enhanced anonymization based on FreeVC* (Li et al., 2024): Both systems are based on the FreeVC (Li et al., 2023) voice conversion framework, which uses a conditional variational autoencoder (CVAE) augmented with GAN training for one-shot voice conversion without the need for text annotations. Both systems integrate emotional feature extraction to condition the decoder, aiming to preserve emotional content during anonymization. **T18-1** uses a wav2vec 2.0 large-robust model fine-tuned on the *MSP-Podcast* dataset as the emotion encoder. After conditioning on emotion information, the system functions in the same way as *FreeVC*. For **T18-2**, handcrafted features extracted using *openSMILE* (Eyben et al., 2010) are processed with an LSTM neural network to create an emotion embedding.

T19-3 *Vector quantized mutual information-based VC (VMC)* (Das et al., 2024): A system based on the vector quantization and mutual information-based VC (VQMIVC) approach (Wang et al., 2021). The content extractor uses VQ and a contrastive-predictive-coding-based training loss on quantized content features (Wang et al., 2021). The training criterion acts to minimize the mutual information between the content, speaker, and pitch representations. The content and speaker extractors, both of which are convolutional DNNs, are trained jointly from scratch. The pitch extractor is based on *WORLD*, while the waveform generator is an independently trained parallel WaveGAN.

T25-1,2* *and* ***T25-p1,p2*** *Emotion-aware anonymization via content and non-content disentanglement with GST modeling* (Gu et al., 2024): Both systems share a common framework comprising: (1) A content encoder adapted from baseline **B5** using ASR-BN features extracted via wav2vec 2.0 combined with vector quantization (VQ), generating discrete content representations; (2) A non-content learner leveraging the GST framework to model non-content attributes like speaker identity, prosody, and emotion via a reference encoder and attention module; (3) A HiFi-GAN vocoder that synthesizes waveform conditioned on combined content and non-content embeddings; (4) An emotion recognition module that predicts the emotion of the input speech, used to select a reference utterance with matching emotion from a speech pool. The anonymization pipeline works by encoding content and emotion from the original speech, selecting a reference utterance with the same emotion for pseudo-speaker embedding extraction, and synthesizing anonymized speech with the combination of original content and pseudo non-content information. The two submitted systems mainly differ in composition of the emotion-labeled corpora for non-content learner: In **T25-1**, *LibriTTS train-clean-100* datasets supplemented with English utterances from the *ESD*; in **T25-2**, only the *ESD* part was used. SER (fine-tuned on *IEMOCAP*) is used at test time to select utterances with matching emotions. However, post-evaluation analysis by this team showed that the rule violation was not critical. Specifically, the authors evaluated these anonymization systems using an alternative SER model: the MMER model (a multimodal multi-task learning approach for speech emotion recognition (Ghosh et al., 2023)). The MMER model was pre-trained on the *MSP-Podcast* corpus and fine-tuned on the *ESD* dataset. The corresponding systems **T25-p1** and **T25-p2**, which replace the *IEMOCAP* SER with MMER, comply with challenge rules. **T25-p1** (like **T25-1**) uses *ESD* as the reference pool, while **T25-p2** uses both *ESD* and *LibriTTS train-clean-100*. Results for **T25-p1** and **T25-p2** show that MMER improves UAR without significantly degrading other metrics (see Table 9).

T33-1,2 *FreeVC-based VC systems with HuBERT and ASR models* (Akti et al., 2024a,b): A zero-shot VC-system based on *FreeVC* (Li et al., 2023): Content features are extracted from the final layer of a transformer-based ASR system and combined with a target speaker embedding and F0 estimates before waveform generation. The target speaker embedding is selected at random from the VCTK training set. **T33-1** is similar to **T33-2** but uses a pre-trained *HuBERT Base* model instead of the transformer-based ASR. The outputs from the 9th transformer block of the HuBERT are used as the content features.

T38-1,2,3,4,5 *DISSC-based systems* (Le Blouch et al., 2024a,b): All based on the use of DIscrete units for Speaking Style Conversion (DISSC) (Maimon and Adi, 2023a), an approach to VC which uses disentangled SSL representations. **T38-1** is the original DISSC system. Discrete HuBERT representations, quantized using k-means to encode linguistic content, are input to predictors that estimate the duration of each discrete unit and F0. Content features and a target speaker embedding randomly selected from the VCTK training set are used by HiFi-GAN to synthesize the output. **T38-2** uses durations and F0 extracted from the input waveform. Extending the same system and to retain expressiveness, **T38-3** uses a HiFi-GAN fine-tuned using the *MSP-Podcast* training set. **T38-4** adds random perturbations to F0 estimates. **T38-5**, a variant of **T38-1**, uses *MSP-Podcast* data to fine-tune duration and F0 prediction modules. Finally, **T38-6** extends **T38-5** by concatenating speaker embeddings with emotion embeddings extracted from the source waveform using a pre-trained SSL model.

5.2. Anonymization systems based on neural audio codecs

Neural codec-based voice anonymization systems represent a modern approach leveraging the powerful representation capabilities of neural audio codecs (NACs). These codecs transform speech into quantized codes that effectively bottleneck speaker-related information while capturing linguistic content and prosodic features. By encoding speech into discrete bottleneck representations, these systems inherently disentangle speaker identity, linguistic content, and prosody. Anonymization is achieved by manipulating or replacing the speaker identity component at the bottleneck level while preserving content and prosodic cues to maintain speech naturalness and intelligibility. Compared to most traditional VC methods that rely on explicit feature extraction and modification, neural codec systems operate within a unified codec framework that naturally separates and recombines speech components.

T10-1,2* *Disentangled neural codec anonymization with serial disentanglement and multi-level distillation* (Yao et al., 2024a): A pair of neural codec systems that use disentangled representations of speaker identity, linguistic content, and fundamental frequency. Emotion cues are preserved through frame-level emotion distillation, while linguistic content and speaker information are disentangled via semantic teacher and self-supervised speaker distillation. Anonymization is performed using a weighted sum of an average voice identity derived from a speaker pool and a

random identity selected from a Gaussian distribution. The parameter α in the anonymization system is a weight that controls the mixture between an averaged speaker identity vector and a randomly sampled speaker identity vector during the anonymization process of speaker identity: $\mathbf{s}_{\text{anon}} = \alpha \cdot \bar{\mathbf{s}} + (1 - \alpha) \cdot \hat{\mathbf{s}}$, where: $\mathbf{s}_{\text{anon}}$ is the anonymized speaker identity vector, $\bar{\mathbf{s}}$ is the averaged speaker identity vector computed from a selected set of candidate speaker identities, $\hat{\mathbf{s}}$ is a randomly generated speaker identity vector, $\alpha \in [0, 1]$ is the mixing weight controlling the degree of influence of the averaged and random vectors. A higher α puts more weight on the averaged speaker identity, typically resulting in less anonymity but better utility preservation, while a lower α increases randomness, enhancing anonymity but potentially decreasing utility. The parameter α allows adjustment of the privacy category. Both systems differ in their weights and belong to different privacy categories: **T10-1** with $\alpha = 0.9$ belongs to the third privacy category (EER $\geq 30\%$), and **T10-2** with $\alpha = 0.8$ belongs to the last one (EER $\geq 40\%$). These systems fail to fully comply with challenge requirements for anonymizing training data used in ASV model training. Specifically, training data should be anonymized at the *utterance level* using the same randomization strategy as applied to test data. Deviations from this protocol may result in a weaker attack model and consequently overestimate the actual level of privacy protection achieved by these systems.

T12-1 *FACodec-based anonymization with pool averaging, cross-gender conversion, and Gaussian noise* (Kuzmin et al., 2024): A factorized neural speech codec (FACodec) (Ju et al., 2024) from NaturalSpeech3 is used to disentangle prosody, content, and other acoustic detail. The codec is conditioned on a speaker embedding generated by a timbre extractor. Anonymization is performed by replacing the original speaker embedding extracted by the codec with an anonymized one. The anonymization strategies experimented with include: pool-based averaging, cross-gender conversion, and the addition of white Gaussian noise (AWGN) to the speaker embedding with varying noise power levels.

T17* *FACodec-based anonymization with enhanced prosody transformation* (He et al., 2024): A FACodec is modified with an additional prosody anonymization module which operates upon a Mel spectrogram and a target speaker embedding. Speaker-specific features contained within latent prosodic features generated using a prosody decoder is suppressed using a speaker embedding extractor paired with a gradient reversal layer. A prosody decoder generates a new Mel spectrogram which contains the prosodic information of the target speaker. The timbre embedding extracted from the original utterance is also replaced with that of the target speaker. Computing a center of speaker embeddings across the entire dataset before anonymization, where embeddings are extracted by the FACodec timbre extractor, may violate the VPC 2024 requirement that trial utterances be processed independently of each other.

5.3. Cascaded ASR+TTS anonymization systems

This section covers anonymization techniques that employ a sequential pipeline of ASR followed by text-to-speech (TTS) synthesis. These cascaded systems transcribe speech into word or phonetic sequences as an intermediate representation, which is then resynthesized to produce anonymized voice outputs.

5.3.1. Conventional cascaded ASR-TTS

These systems follow the traditional approach of passing text transcripts from ASR directly to TTS without explicitly preserving prosodic or paralinguistic attributes.

T8-1 *Cascaded ASR (Whisper) and TTS (VITS)* (Xinyuan et al., 2024): Text transcripts are derived from the input using the *medium English Whisper* model. A VITS (Kim et al., 2021) TTS model trained using the *LibriTTS* dataset is then used to generate anonymized outputs.

5.3.2. Cascaded ASR-TTS systems with paralinguistic preservation

These advanced cascaded systems integrate mechanisms to preserve paralinguistic features such as emotion, intonation, and speaking style alongside speech content.

T12-2,3 ***B3****-based system with integrated emotion embeddings and cross-gender averaging and selecting anonymization* (Kuzmin et al., 2024): Both systems are based on the **B3** baseline system and utilize the *FastSpeech2* model with HiFi-GAN vocoder to synthesize speech. They integrate emotion embeddings extracted with a fine-tuned *wav2vec 2.0* model to enhance emotion preservation. Speaker anonymization in both cases involves cross-gender related strategies. The differences lie in their speaker anonymization mechanisms and embedding usage, and prosody modification

module. System **T12-2** anonymizes the speaker by averaging 100 speaker embeddings randomly selected from a pool of 200 farthest embeddings from the source speaker, combined with cross-gender selection. System **T12-3** employs random speaker selection from a cross-gender pool on a per-utterance basis rather than embedding averaging. In addition, **T12-3** modifies prosody through pitch and energy features normalized and adjusted by multiplicative random factors in the range [0.7, 1.3].

T12-4 ***B3****-based anonymization with cross-gender selection-based anonymization* (Kuzmin et al., 2024): This system is based on the **B3** baseline but does not modify prosody. Instead of using GAN-generated pseudo-speaker embeddings, it applies cross-gender selection-based anonymization for speaker embedding anonymization.

T30-1 ***B3****-based anonymization with embedding pool perturbation and k-means clustering* (Lee et al., 2024b,a): The GAN-based speaker embedding generation process of **B3** is replaced with a pool-based anonymization algorithm. The speaker embedding from the pool with the largest cosine distance to the input speaker embedding is selected. A random selection of embedding dimensions is then replaced with the corresponding dimensions from a speaker embedding selected at random from the 500 pool cluster centroids.

T30-2 *Cascaded ASR+TTS anonymization with emotionally enriched feature integration* (Lee et al., 2024b,a): The concatenation of Whisper ASR and variational inference with adversarial learning for end-to-end text-to-speech (VITS) (Kim et al., 2021) TTS systems. F0 estimates extracted from the input waveform are transformed into latent features which are then combined with a set of emotion prototype vectors using cross attention. Combined features are input to a Feature-wise Linear Modulation (FiLM) layer, the output of which is multiplied with Whisper ASR text embeddings. The resulting features are used with VITS for waveform generation.

5.4. Hybrid anonymization systems

Hybrid systems combine multiple anonymization techniques to balance privacy and utility. Different approaches exhibit distinct trade-offs: cascaded ASR+TTS achieves strong anonymization but degrades emotion preservation, while VC preserves paralinguistic content at the cost of weaker privacy. Hybrid systems address this by enabling dynamic method selection, such as the randomized *admixture* approach proposed in challenge submissions, allowing fine-grained control over the privacy-utility trade-off.

T8-3,4,5 *Admixture* (Xinyuan et al., 2024): Anonymization is carried out according to one of two methods selected at random. The first is the cascaded ASR-TTS system **T8-1** whereas the second is the kNN-VC system **T8-2**. This hybrid method provides a balance between anonymization strength and the preservation of utility, with the **T8-1** system better suppressing voice information at the cost of reduced utility and the **T8-2** system better preserving linguistic and emotional content at the cost of reduced anonymization.

Table 7: Summary of the submitted systems. Colors in the first column highlight the privacy category of the corresponding systems: EER≥ 10%, EER≥ 20%, EER≥ 30%, EER≥ 40%. Noncompliance with VPC 2024 requirements is marked by*.

ID	ID (paper)	Training resources	Description
T7-1	ppg-w2vF0-fusion	LibriTTS train-clean-360, LibriTTS train-other-500, LJSpeech, ESD, wav2vec 2.0 FT on MSP-Podcast	anonymization: fusion at the parameter level of pseudo-speakers (several pretrained end-to-end VC models fine-tuned with single speaker data); emotion encoder: based on wav2vec 2.0; F0: Yaapt; linguistic content: PPG-features from a hybrid Transformer-CTC ASR
T7-2	ppg-gstF0-fusion	LibriTTS train-clean-360, LibriTTS train-other-500, LJSpeech, ESD	similar to **T7-1** except for emotion encoder: GST
T8-1	Whisper-VITS TTS	LibriTTS, medium English Whisper	cascaded ASR (Whisper) and TTS (VITS)

Table 7: Summary of the submitted systems. Colors in the first column highlight the privacy category of the corresponding systems: EER≥ 10%, EER≥ 20%, EER≥ 30%, EER≥ 40%. Noncompliance with VPC 2024 requirements is marked by*.

ID	ID (paper)	Training resources	Description
T8-2	kNN-VC	LibriSpeech train-clean-100, VoxCeleb; WavLM	kNN-VC system on WavLM features
T8-3	Admixture ($p = 0.2$)	as in {**T8-1** ∪ **T8-2**}	random selection of one of two methods for each utterance (with probability p for the second method): **T8-1** (cascaded ASR-TTS) and **T8-2** (kNN-VC)
T8-4	Admixture ($p = 0.325$)		
T8-5	Admixture ($p = 0.4$)		
T9	GMM-Blender	ESD, LibriSpeech train-clean-360, train-other-500, WavLM, wav2vec2-large-robust-12-ft-emotion-msp-dim	anonymization: *GMM-Blender* based on GMM speaker clustering to generate anonymous templates; emotion encoder: wav2vec2-large-robust-12-ft-emotion-msp-dim
T10-1*	C3	LibriSpeech, LibriTTS; WavLM	neural audio codec, with a specific disentanglement strategy; anonymization: weighted sum of an averaged speaker identity (from speakers in a speaker pool) and a randomly generated)
T10-2*	C4		similar to **T10-1**, but with smaller weight for an averaged speaker identity
T12-1	1a	LibriTTS train-clean-100, LibriLight train, NaturalSpeech3 FACodec	relies on FACodec, anonymization: pool-based, cross-gender selection, addition of white Gaussian noise.
T12-2	1b	LibriTTS train-clean-100, train-other-500, MSP-Podcast; wav2vec 2.0	modifies **B3**; emotion encoder: wav2vec-based; anonymization: pool-based, selection+averaging, cosine distance, cross-gender)
T12-3	2a	LibriTTS train-clean-100, train-other-500, MSP-Podcast; wav2vec 2.0	modifies **B3**; emotion encoder: FT wav2vec 2.0 prosody: value-wise F0 and energy multiplication by random values; anonymization: random selection, cross-gender
T12-4	2b	LibriTTS train-clean-100, train-other-500	modifies **B3**; anonymization: random selection, cross-gender
T12-5	3	VoxPopuli, Librispeech train-clean-100, LibriTTS train-clean-100	modifies **B5**; prosody: replaces the pitch curve with a weighted average between the curve itself and its moving average.
T12-6	4		similar to **T12-5**, adds white Gaussian noise to the resulting pitch curve.
T14-1	v1	LibriSpeech, CREMA-D	VQ VAE; prosody: normalized F0 and energy from inputs; anonymization: selection+averaging as in **B1**
T14-2	v2		similar to **T14-1** with F0 scaling by random factors in range [0.8, 1.2].

Table 7: Summary of the submitted systems. Colors in the first column highlight the privacy category of the corresponding systems: EER≥ 10%, EER≥ 20%, EER≥ 30%, EER≥ 40%. Noncompliance with VPC 2024 requirements is marked by*.

ID	ID (paper)	Training resources	Description
T17*	FACodec+PANO	LibriLight, LibriSpeech train-other-500; VITS (for prosody encoder); NaturalSpeech3 codec	adds a prosody anonymization module to FACodec.
T18-1	EESA(Wv)	MSP-Podcast, RAVDESS, VCTK; FreeVC, WavLM, wav2vec 2.0, wav2vec2-large-robust12-ft-emotion-msp-dim	FreeVC integrated with emotion embeddings generated by a fine-tuned version of wav2vec 2.0.
T18-2	EESA(OS)	MSP-Podcast, RAVDESS; VCTK, FreeVC, WavLM, wav2vec 2.0	FreeVC enhanced with emotion embeddings extracted via an LSTM network operating on openSMILE features
T19-1	KNNS	LibriSpeech train-clean-360, ESD, RAVDESS, CREMA-D	kNN-VC system using WavLM features (from 6th Transformer block)
T19-2	KNND		kNN-VC system using WavLM features (concatenated from the 6th & 12th Transformer blocks)
T19-3	VMC		anonymization: randomly selected speaker vector from ESD speakers; F0: PyWORLD; linguistic content: CNN + vector quantization.
T25-1*	large: ESD+LibriTTS	ESD, LibriTTS train-clean-100, IEMOCAP (test time); wav2vec 2.0 content encoder from **B5**	disentanglement of content (VQ-BN as in **B5**) and style features; emotion transfer from target speaker utterances; non-content learner: trained on LibriTTS train-clean-100 datasets supplemented with English utterances from the ESD;
T25-2*	small: ESD only		similar to **T25-1**, but the non-content learner is trained on the ESD subset only
T30-1	V-Beam_sttts-tj-method3	LibriTTS, RAVDESS, ESD	modifies **B3**; anonymization: selection-based from a pool created by k-means clustering
T30-2	V-Beam_method2	LJspeech; Whisper, VITS	cascaded ASR (Whisper) + TTS (VITS using LJspeech) with integration of emotion class prototype features with F0
T33-1	freevc-hubert-base-ss	VCTK; HuBERT Base L (km500 quantizer)	anonymization: randomly selected from VCTK; F0: not explicitly extracted or converted; linguistic content: features extracted from the 9th block of HuBERT
T33-2	freevc-asr-bottleneck-f0-ss	VCTK, LibriSpeech	similar to **T33-1**, but content features are extracted from the final layer of a CTC-Transformer-based ASR
T38-1	M0 - original DISSC	DISSC model trained on VCTK	DISSC using speaker embeddings randomly selected from the VCTK speakers

Table 7: Summary of the submitted systems. Colors in the first column highlight the privacy category of the corresponding systems: EER≥ 10%, EER≥ 20%, EER≥ 30%, EER≥ 40%. Noncompliance with VPC 2024 requirements is marked by*.

ID	ID (paper)	Training resources	Description
T38-2	M1 - prosody preservation	as in **T38-1**	**T38-1** but using duration and F0 of input waveform
T38-3	M2 - dataset for expressiveness	as in **T38-1** + MSP-Podcast	**T38-1** + HiFi-GAN trained on MSP-Podcast
T38-4	M3 - add randomeness to prosody		**T38-3** + random noise in F0
T38-5	M4 - prosody prediction with MSP-Podcast		**T38-1** + prosody predictors trained on MSP-Podcast
T38-6	M5 - emotion embeddings	MSP-Podcast; DISSC model trained on VCTK, wav2vec2-large-robust-12-ftemotion-msp-dim	**T38-5** + emotion embeddings

5.5. Trends

5.5.1. Anonymization

Cascade-based approaches: Systems like **T8-1** that cascade ASR+TTS achieve near-perfect anonymization with an EER of approximately 48%. However, while linguistic content is largely preserved, there are significant compromises on emotion preservation. To achieve a better balance, *admixture* strategies were employed whereby the specific anonymization method applied to each utterance is selected at random, allowing a flexible privacy-utility trade-off. The benefit of such an approach is seen in results for hybrid systems **T8-3**, **T8-4**, **T8-5** that apply cascaded ASR-TTS and kNN-VC methods to each utterance at random.

Speaker embedding anonymization methods:

- *Weighted averaging and random mixing:* **T10-1,2** systems combine an averaged speaker identity derived from a speaker pool with a randomly generated speaker identity sampled from a Gaussian distribution. A mixing weight parameter controls the balance between averaged and random components. Higher weights better preserve utility by favoring averaged identities, while lower weights increase randomness for enhanced privacy.
- *Pool-based selection with optional averaging and cross-gender conversion:* Multiple systems employ pool-based speaker embedding selection strategies with varying degrees of averaging and cross-gender constraints:
  - *Averaging-based strategies*: **T12-1** implements the FACodec-based system with pool-based speaker embedding averaging combined with cross-gender speaker selection and the injection of additive white Gaussian noise to speaker embeddings with varying noise power levels. **T14-1,2** employ selection and averaging as in baseline **B1**. Also similar to **B1**, **T12-2** averages 100 speaker embeddings selected at random from a pool of 200 farthest embeddings based on cosine distance, combined with cross-gender selection.
  - *Single speaker selection strategies:* **T12-3** employs random speaker selection from a cross-gender pool on a per-utterance basis. **T30-1** uses selection-based anonymization from a pool created by k-means clustering of speaker embeddings. **T33-1,2** selects single pseudo-speaker embeddings at random from the VCTK speaker corpus. **T38-1,2,3,4,5,6** implements DISSC using speaker embeddings selected at random from VCTK speakers.
- *Emotion-matched pseudo-speaker selection:* **T25-1,2** systems employ an emotion-aware anonymization strategy with which pseudo-speakers are selected from emotion-labeled pools. An emotion recognition module is used

to estimate the emotion of the input speech so that a reference utterance with matching emotion is selected from the pool. The strategy helps to better maintain emotional consistency while still replacing the voice identity.

F0 manipulation for prosody alteration: **T12-3** modifies prosody through value-wise F0 and energy scaling by random multiplication factors. **T12-5,6** achieve prosodic anonymization through mean-reversion pitch smoothing, where F0 estimates are interpolated with the moving average of adjacent frames. **T12-6** additionally applies additive white Gaussian noise to the smoothed F0 estimates. **T14-2** introduces random F0 scaling by factors in the range 0.8–1.2 during synthesis to obscure pitch-related speaker cues. **T17** integrates an additional prosody anonymization module.

5.5.2. Emotion preservation

Diverse strategies were explored in VPC 2024 submissions to improve upon the preservation of emotion cues during voice anonymization.

- *Pretrained SSL-based emotion encoders:* **T7-1** and **T18-1** employ *wav2vec 2.0* models fine-tuned on emotional corpora (MSP-Podcast, ESD) to extract emotion embeddings that condition the decoder during synthesis. **T19-2** extends single-layer *WavLM* features (6th Transformer block) to dual-layer concatenation of the 6th and 12th blocks, with the 12th layer capturing emotional cues more effectively.
- *Handcrafted feature-based emotion encoding*: **T18-2** uses *openSMILE* to extract acoustic features which are processed with an LSTM network to generate emotion embeddings.
- *Global Style Token (GST) approaches*: **T7-2** and **T25-1,2** leverage GST frameworks composed of multi-head Attention modules to model emotion and style features as learnable style embeddings extracted through reference encoders.
- *Emotion distillation*: **T10**'s neural codec systems use frame-level emotion distillation as part of a multi-stage disentanglement strategy, explicitly preserving emotional information alongside linguistic content and speaker identity.
- *Emotion recognition-guided reference selection:* **T25-1,2** use a SER model trained on *IEMOCAP* to estimate the emotion in the input speech. Reference utterances with matching emotion labels are then selected from emotion-annotated pools (*ESD* or *ESD+LibriTTS*), to better preserve emotional expressiveness.

5.5.3. Linguistic content preservation

Content preservation strategies vary based on the anonymization approach.

- *Vector quantization (VQ) bottleneck features:* Multiple systems (**T14-1,2**, **T19-3**, **T12-5,6**, and **T25-1,2**) employ vector quantization-based disentanglement to separate speaker identity from linguistic content. **T14-1,2** use a convolutional module to extract content features, which are subsequently quantized via VQ to remove residual speaker information while preserving content representation. Similar to baselines **B5** and **B6**, **T12-5,6** and **T25-1,2** apply VQ to ASR-BN features. Continuous bottleneck representations are discretized into tokens to reduce speaker leakage. The principle underlying these approaches is that quantization into discrete tokens acts as a bottleneck that attenuates speaker-specific variations while retaining the discrete linguistic content necessary for speech recognition and synthesis.
- *Pretrained SSL representations*: Several systems leverage pretrained self-supervised speech models as content encoders, exploiting their robust capture of linguistic information across diverse acoustic conditions. **T33-1** extracts content features from the *HuBERT Base* model; **T8-2** and **T19-1,2** employ *WavLM* features. **T8-2** and **T19-1,2** use *WavLM* features.
- *Phonetic transcriptions*: **T12-2,3,4**, **T30-2** employ cascaded ASR systems (*Whisper*, hybrid CTC-Attention ASR) to obtain phonetic transcriptions or intermediate representations that are then fed to TTS systems to encourage the preservation of linguistic information at the transcription level.

6. Results

A summary of results in terms of utility (WER,% or UAR,%) against privacy (EER,%) is illustrated in Figure 7 for submitted systems (orange stars) described in Section 5 and baseline systems (blue triangles) described in Section 3.4. For all cases, EER results are derived using ASV models re-trained by the organizers using anonymized data provided by participants. Results for original, unprotected data are indicated with black circles. Each EER estimate is an average of EERs derived separately using male and female speaker subsets computed using the *LibriSpeech test-clean* dataset. Detailed results for development and test datasets can be found in Table 9.

6.1. Privacy achievement across categories

Results and rankings for utility metrics (WER,% and UAR,%) against the EER are illustrated in Figure 8 across four different privacy thresholds (10, 20, 30, and 40% of EER). WER values are given with 95% confidence intervals. The EER for original, unprotected test data is 4.6%. For the first category (EER $\geq$ 10%) there are 28 submitted systems and 4 baselines. Of these, 21 submitted systems and all 4 baselines meet the requirements of the second category (EER $\geq$ 20%). For the third category (EER $\geq$ 30%) there are 15 submitted systems and 2 baselines, while only 6 submitted systems (**T10-2**, **T8-5**, **T25-1**, **T12-5**, **T30-2**, and **T8-1**) meet the criteria for the fourth strongest category (EER > 40%). However, two of these (**T10-2** and **T12-5**) are non-compliant with the challenge rules. The highest EER, close to 50%, is achieved by the cascaded ASR-TTS system **T8-1**.

6.2. Comparative performance by system type

6.2.1. Neural codec-based systems (T10, T12-1, T17)

Figure 8 demonstrates that some neural codec systems achieve strong performance in the highest privacy category. **T10-2** (EER=40.4%), which combines a codec with serial disentanglement and frame-level emotion distillation, achieves the best overall trade-off between privacy protection (EER $\geq$ 40%) and preservation of both linguistic content (WER=3.2%) and emotional expressiveness (UAR=60.8%). **T17**, a FACodec-based system, is in the second category (EER=27%) and demonstrates is also effective in preserving utility WER=3%, UAR=51.4%). **T12-1**, which leverages a FACodec with pool-based averaging, cross-gender conversion, and AWGN, offers weaker privacy protection (EER=9.7%) and moderate utility characteristics.

6.2.2. Cascaded ASR+TTS systems (T8-1, T12-2,3,4, T30-2)

These systems achieve near-perfect speaker anonymization with EER values for one (**T8-1**) approaching 50%. However, as seen in Figure 7, cascaded systems systematically show substantially lower UAR values compared to neural codec and VC-based alternatives. **T8-1**, a cascaded ASR-TTS system using *Whisper-VITS*, achieves the strongest privacy (EER=49.5%) but is among the worst for emotion preservation (UAR=30.6%). However, linguistic content preservation remains strong (WER=3.75%, the third-ranked system in the highest privacy category). The asymmetric preservation of content through the transcription bottleneck but degradation of paralinguistic attributes highlights the main limitations of text-based anonymization.

Among cascaded ASR-TTS systems with paralinguistic preservation, **T30-2** in the same highest category (EER=46.3%), achieves worse results than **T8-1** for both utility metrics (WER=4.78%, UAR=29.4%), indicating that the incorporation of latent features based on original F0 and emotion prototype vectors may lead to a minor reduction in privacy while not improving emotion preservation. **T12-2,3,4** – **B3**-based anonymization systems – fall into the first (**T12-2**) and second (**T12-3,4**) privacy categories. **T12-2** and **T12-3** use emotion encoders that improve UAR (up to 43%), which is the best result for this type of system. For comparison, similar systems without emotion encoders (i.e., **T12-4, B3**) achieve UAR values not exceeding 38%.

6.2.3. Hybrid admixture systems (T8-3,4,5)

T8's *admixture* approaches, which randomly select between cascaded ASR+TTS (**T8-1**) and kNN-VC (**T8-2**) on a per-utterance basis, achieve above-average performance across privacy-utility objectives. The mixing probability allows flexible adjustment of the privacy-utility tradeoff within each privacy category: UAR changes proportionally to EER, while WER remains similar across all systems (approximately 3.2–3.3%). **T8-5**, which ranks among the top systems, achieves EER=41.3% while maintaining UAR closer to the levels achieved by pure VC systems (UAR=47%), demonstrating that hybrid approaches effectively enable context-dependent privacy-utility optimization.

6.2.4. VC-based systems (T7, T8-2, T9, T12-5,6, T14, T18, T19-1,2,3, T25, T33, T38)

Pure VC approaches generally struggle to achieve the highest privacy categories but in some cases show good utility preservation. kNN-VC variants (**T8-2**, **T19-1,2**) achieve relatively strong WER and UAR values but typically their EER does not exceed 10%. Emotion-aware VC systems such as **T25-1,2** and **T9** demonstrate improved emotion preservation compared to most baseline VC systems but remain below neural codec performance (e.g., **T10-1,2**) in the highest privacy categories. **T9** ranks among the top systems in the first three privacy categories (EER=34.5%, WER=2.4%, UAR=61%).

6.3. Utility metrics analysis

6.3.1. Linguistic content preservation (WER)

The WER on unprocessed test data is 1.85%. In the first three privacy categories, the best system for content preservation, **T9**, achieves WER=2.37% on the test data, corresponding to 0.5% absolute and 28% relative WER increase. In the highest privacy category, the two systems with the lowest WER (3.2%) – the cascaded system **T8-5** and the codec-based system **T10-2** – demonstrate more noticeable ASR performance degradation: approximately 1.4% absolute and 73–74% relative. Within the highest privacy category, WER ranges from 3.2% to 4.8%.

6.3.2. Emotion preservation (UAR)

The UAR on the original (unprocessed) test data, marked with a black circle in Figure 7, is 71%. In the first three privacy categories, the top systems (**T10-1,2**, **T9**) achieve UAR values of approximately 61%–65%, while rule-compliant systems (**T9**, **T7-1**, **T8-4**) achieve approximately 49%–61% in the third privacy category and 51%–61% (**T9**, **T8-3**, **T7-1**) in the first two privacy categories. In the highest privacy category, top-performing systems (**T10-2**, **T25-1**, **T8-5**) achieve UAR values of approximately 47%–61%, and the best rule-compliant system (**T8-5**) achieves 47%. The spread between best and worst rule-compliant systems (**T8-5** and **T30-2**) in the highest privacy category exceeds 17% (and 31% for all submitted systems in this category) in UAR, underscoring substantial performance differences and highlighting the importance of emotion-aware architecture choices. Systems explicitly incorporating emotion preservation mechanisms (**T9**, **T10-1,2**, **T25-1,2**) rank higher in UAR across all privacy categories compared to systems without emotion-specific designs.

Top systems (**T9**, **T10**, **T25-2**) achieving high privacy categories (EER$\geq$30%,40%) also rank among the best performers in lower privacy categories (EER$\geq$10%,20%) for emotion preservation. This consistency across privacy thresholds indicates that these systems benefit from principled architectural designs, such as multi-stage emotion-aware disentanglement, emotion-matched pseudo-speaker selection, and hybrid anonymization approaches, that do not trade off emotion preservation for privacy gains.

Figure 9 shows emotion-specific accuracy. For original speech, accuracy is balanced across all four classes. Anonymization impacts accuracy differently for specific emotion classes. In particular, many systems fail to recognize *sadness*, which is the hardest class to preserve and has the lowest accuracy across all anonymized systems. Balanced systems include **T10-1**, which preserves all emotions well (60–70% accuracy for each). *Happiness*-biased systems include **T18-1**, **T14-1,2**, and **T30-2**, which achieve excellent *happiness* accuracy (80–94%) but poor *sadness* accuracy (29–45%). Systems **T8-1**, **T30-1**, **T30-2**, **B3**, and **T33-2** have near 0% accuracy for *sadness*.

6.4. Privacy-utility trade-off

The VPC 2024 reveals fundamental trade-offs across all evaluation metrics, with no single system simultaneously maximizing privacy (EER), linguistic content preservation (WER), and emotion preservation (UAR). The challenge results demonstrate a non-uniform operating space where the relationship between privacy and utility varies substantially across different metrics and operating points.

Figure 10 presents scatterplots of relative metric changes (ΔWER, ΔUAR, and ΔEER) for anonymization systems with respect to results on original data. The Pareto frontiers shown across these plots reveal distinct trade-off patterns. The first plot (*Privacy vs. Utility (UAR)*) identifies systems on the frontier, with **T10-1** achieving the best balance (ΔUAR = 9%, ΔEER = 704%), followed by **T10-2** (ΔUAR=14%, ΔEER = 779%) and **T25-1** (ΔUAR=31%, ΔEER=818%). In contrast, **T8-1** achieves the strongest privacy (ΔEER = 979%) but at significant utility cost (ΔUAR=57%). The second plot (*Privacy vs. Utility (WER)*) demonstrates that **T9** achieves a good trade-off (ΔWER = 28%, ΔEER = 652%), while **T10-1** again performs competitively (ΔWER = 39%, ΔEER = 704%). The third plot (*Both Utilities (UAR vs. WER)*) identifies systems that preserve both utility metrics simultaneously, with **T9** (ΔUAR

= 14%, ΔWER = 28%) and **T10-1** (ΔUAR = 9%, ΔWER = 39%) representing the most effective approaches. These Pareto frontiers underscore the critical role of system design choices in achieving balanced privacy-utility trade-offs.

Figure 13 presents a normalized performance matrix displaying all systems ranked by overall combined score, enabling direct comparison of privacy-utility trade-offs across all submission approaches. The challenge's multi-threshold evaluation shows that optimal system selection depends on application requirements. *Admixture* systems (**T8-3,4,5**) demonstrate that dynamic method selection on a per-utterance basis enables flexible operating points, with mixing probability controlling positions along the privacy-utility spectrum.

6.5. Gender impact on anonymization

Figure 11 displays EER results for the test set for female and male speaker subsets, ordered by the difference ($EER_{male} - EER_{female}$) in descending order. Gender-based performance disparities vary significantly across systems. The original ASV system on unprotected data exhibits substantial male bias (EER=8.8% for female speakers and EER=0.4% for male speakers), indicating that male speaker verification is inherently easier. The most gender-biased anonymisation systems include **T17**, which exhibits a 14% absolute difference (EER_{female}=34%, EER_{male}=20%), and **T33-2**, which shows females bias (EER_{female}=24%, EER_{male}=37%). Gender-balanced systems include **B6**, **T8-4**, **T25-1**, **T38-3**, and **T33-1**, which maintain differences below 0.5%, demonstrating uniform anonymization effectiveness across genders.

To further assess gender fairness in anonymization, we employ the *modified fairness discrepancy rate* (mFDR), evaluated at the EER operating point, which captures the relative magnitude of performance difference between genders. This simplified metric, adapted for speaker verification systems and VPC 2024 evaluation conditions, is defined as follows [12]:

$$\text{mFDR} = \frac{2 \cdot |\text{EER}_{\text{female}} - \text{EER}_{\text{male}}|}{(\text{EER}_{\text{female}} + \text{EER}_{\text{male}})} \cdot 100.$$

Figure 12 displays gender fairness analysis for all anonymization systems using the mFDR metric on both test and development datasets. Systems are categorized by mFDR performance, ranging from *excellent* (mFDR<5%) for systems such as **B6**, **T8-4**, **T25-1**, **T10-1**, **B5**, **T8-1**, and **T38-5** to *poor* for systems **B2**, **T14-2**, **T8-2**, and **T14-1**. The maximum mFDR value is observed on the original data (182%). Figure 12 (2) shows consistency of mFDR across the development and test sets. For most of the anonymization systems, mFDR improves on the test set.

6.6. Baseline influence

Table 7 presents the baseline influence on submitted systems. The baseline systems (**B1**–**B6**) established performance boundaries that participant systems aimed to improve upon. The x-vector anonymization strategy based on pool selection and averaging, introduced in **B1**, was employed in **T14-1** and **T14-2**. Multiple cascaded ASR-TTS systems with paralinguistic preservation are based on **B3**, which uses phonetic transcriptions and GAN-based pseudo-speaker generation. **T12-2**, **T12-3**, **T12-4**, and **T30-1** build directly upon **B3**, introducing emotion awareness and alternative anonymization strategies. Performance for these systems is similar to **B3**, with slight improvements in overall utility for **T12-2** and **T12-3**. **T12-5** and **T12-6** modify **B5** (ASR-BN with VQ-based content features) by replacing the pitch curve with a weighted average between the original curve and its moving average. **T12-6** additionally applies white Gaussian noise to the resulting pitch curve. These systems achieve performance close to **B5**. In contrast, **T25-1** and **T25-2** employ VQ-BN for content features while introducing emotion transfer mechanisms, achieving significantly better privacy and UAR metrics than **B5**. VQ-based approaches are also employed in **T19-3**. The success of neural codec baseline **B4** motivated substantial community adoption, with **T10**, **T12-1**, and **T17** representing independent neural codec approaches that achieve competitive performance. **T10** and **T17** significantly outperform **B4** in both WER and UAR, and **T10** systems also achieve higher EER values. Overall, the best submitted systems significantly outperform baselines across all metrics. EER improves from 34% (**B5**) to 49.5% (**T8-1**), UAR improves from 54% (**B2**) to 65% (**T10-1**), and WER improves from 2.9% (**B1**) to 2.37% (**T9**).

[12] A similar approach to assess gender impact in the speech domain was used in (Zanon Boito et al., 2022). A more general definition of fairness discrepancy rate (FDR) in biometric systems can be found in (Freitas Pereira and Marcel, 2021), which is calculated for operational conditions where a single threshold is set for all demographic groups. Such a metric, in application to the VPC 2024 evaluation setup (EER operational point, two gender groups), would correspond to a weighted sum of the absolute differences in FRR and FAR between female and male speakers. The mFDR, while using normalization of the gap by the overall performance level, addresses issues of difference-based bias measures that lose sensitivity when base error rates are small or differ in magnitude across conditions. From this perspective, mFDR is closer in spirit to *Gini aggregation rate for biometric equitability* (GARBE) (Howard et al., 2022; Chouchane et al., 2024).

6.7. Summary

Cascaded approaches clearly dominate in absolute privacy (EER = 48%) but fail in emotion preservation. Neural codec methods achieve the best overall balance, with **T10-2** representing state-of-the-art privacy-utility compromise. Hybrid approaches successfully bridge different optimization targets through dynamic method selection. Advanced VC systems with explicit attribute disentanglement and self-supervised learning (SSL) representations demonstrate substantially improved performance compared to traditional VC, reflecting the field's adoption of deep learning advances. The challenge results demonstrate that while strong privacy protection remains achievable, it requires sophisticated architectural choices that balance multiple competing objectives. The evolution from simple baseline approaches to complex multi-stage disentanglement systems reflects rapid methodological development in voice anonymization research.

7. Discussion and future perspectives

In this section we present key insights from the VPC 2024 and previous editions, examine evaluation methodology and data limitations, survey emerging research directions, and provide strategic recommendations for future voice anonymization research.

7.1. Evaluation

Reliable evaluation is key to the fostering of progress in anonymization while also helping to ensure real-world applicability. However, current evaluation methodologies present limitations that require careful analysis.

7.1.1. Reliability of privacy evaluation

EERs approaching 50% under the *semi-informed* attack model (see Section 2.2) indicate ideal privacy protection. Many systems submitted to the VPC appear to achieve such high levels of privacy. However, several recent analyses, call into question whether such encouraging estimates of anonymization performance are reliable. Results from the **First VoicePrivacy Attacker Challenge**[13] (Tomashenko et al., 2025a) and other works (Panariello et al., 2025) show that exaggerated EERs can be the result of poor attacker generalization rather than strong anonymization. Meanwhile, the current attack model imposes a number of constraints which can also contribute to the overestimation of privacy. These include:

- *Data*: the attacker trains an ASV system using anonymized data produced by each system. In more realistic scenarios, attackers may access arbitrary external corpora, including other sets of anonymized speech, to strengthen the attack.
- *Features*: the attacker uses Fbank input features and trains a speaker encoder following a typical ASV pipeline, while other studies (Tomashenko et al., 2025c,b; Bakari et al., 2025) show that context-dependent duration features, temporal speech dynamics, and non-timbral cues can all effectively reveal speaker-specific information after anonymization.
- *Model and hyperparameters*: attacks are performed using a fixed architecture (i.e. ECAPA-TDNN) and untuned hyperparameters, whereas there is evidence that stronger attacks can be implemented using other models (Bakari et al., 2025), e.g. *WavLM*.

These limitations were addressed in the First VoicePrivacy Attacker Challenge (Tomashenko et al., 2025a, 2024b), an ICASSP 2025 SP Grand Challenge which evaluated attacks against a selection of voice anonymization systems. The use of strong attacks is crucial to ensure reliable privacy evaluation. Seven anonymization systems were considered. They were the best three VPC 2024 baselines, namely **B3**, **B4** and **B5** (see Section 3.4), and a selection of four participant systems which (initially) achieved the highest level of privacy (EER≥40%, category 4 in Figure 2) when evaluated using the baseline attack model, namely **T8-5**, **T10-2**, **T12-5**, and **T25-1**. When evaluated with the baseline attacker, these systems achieved the 4th privacy category (EER≥40%). However, when tested against the best attack models developed by Attacker Challenge participants, EERs for all systems fall substantially (EER≥20%, category 2). Results also show that system ranking based on EERs changes when evaluation is performed using stronger attack models. These finding highlight the importance to reliable and trustworthy privacy evaluation of using the strongest possible attack models.

[13] https://www.voiceprivacychallenge.org/attacker/

7.1.2. Metrics

Since the 2020 VPC, a wide range of evaluation metrics have been proposed, covering both privacy and utility aspects. Nevertheless, several limitations and concerns remain:

Privacy metrics. Use of the EER fails to reflect legal privacy requirements such as *linkability*, *singling out*, or *inference* (Article 29 Working Party, 2014). The EER can be insensitive to re-identification risks, remaining stable even as the probability of successful re-identification varies for specific users or use cases. It focuses only on average performance at a single operating point, which masks worst-case risks and outlier vulnerabilities, potentially giving an overly optimistic view of privacy protection in the case of individuals who remain readily identifiable even after anonymisation (Nautsch et al., 2020; Williams et al., 2024).

To better align with both technical and legal expectations, alternative evaluation frameworks have been proposed. One such approach, based on *linkability* and *singling out*, and inspired by the legally accepted interpretation of the definitions of anonymous information was proposed in Vauquier et al. (2025). Results derived using these metrics indicate that residual privacy risks after anonymization are more substantial than suggested by EER estimates, underscoring the need for evaluation metrics that are better aligned with both technical and legal perspectives. Alternative metrics also include the *similarity rank disclosure* (SRD) metric proposed by Bäckström et al. (2025), measured in units of entropy (bits). It is defined as the reduction in entropy (uncertainty) about the true identity after observing the similarity rank. The metric quantifies how much personally identifiable information (PII) about the true identity can be inferred from the similarity rank when comparing a query sample to a database of templates and shows promise as a metric for the evaluation of privacy in voice anonymisation (Chandra et al., 2025). The adoption in future VPC editions of metrics that reflect performance for the most challenging or privacy-sensitive samples might provide estimates of anonymization performance that are better aligned with expectations in real-world settings.

Utility metrics. One limitation of current utility metrics is the disconnect between automatic metrics and perceptual quality. While more powerful ASR models generally yield lower WERs, this does not necessarily indicate improved quality, or even intelligibility. Studies demonstrate that WER estimates are not always positively correlated with subjective intelligibility (Panariello et al., 2024b), highlighting a fundamental gap between automatic evaluation and human perception. Additionally, while estimates of the WER reflect the preservation of speech content, they neglect other important aspects of utility. Speech naturalness, prosody preservation, and acoustic quality are underrepresented by current metrics.

Proposed solutions include developing multi-dimensional metrics that assess both objective and subjective utility, integrating evaluation criteria for specific use cases, and designing utility metrics that jointly evaluate intelligibility, naturalness, and downstream task utility.

7.1.3. Data and benchmarking

All previous VPC editions have relied primarily on the *LibriSpeech* datasets for the evaluation of privacy and content preservation, with VCTK also having been used for the 2020 and 2022 editions. These corpora contain clean, single-speaker, read English speech. Since the latter lack emotional annotations, the *IEMOCAP* corpus, which features spontaneous and emotionally expressive speech, was introduced for the 2024 edition to support the evaluation of emotion preservation. However, due to the limited number of speakers and size, *IEMOCAP* is itself not suited to the evaluation of privacy and content preservation.

There are some limitations of the current datasets and experimental protocols: (1) *Dataset inconsistency in privacy and utility evaluation.* The estimation of privacy and content preservation from one dataset, and emotion preservation from another dataset might introduce some bias in the results, raisings concerns about whether comparisons of system performance are fair. (2) *Language and accent coverage.* Speech data are of English language, with minimal representation of multilingual, code-switched, or heavily accented speech. (3) *Speech characteristics. LibriSpeech* contains only clean, read speech from a limited demographic; the dataset lacks spontaneous speech, child speech, pathological speech, and speaker variations reflecting real-world diversity. (4) *Recording conditions.* There is no systematic evaluation of anonymization robustness under noisy, reverberant, or compressed acoustic conditions typical of many real-world applications (e.g., phone calls, live streaming, video conferencing). (5) *Conversational scenarios. LibriSpeech* and *IEMOCAP* are single-speaker, and current solutions explored in the challenge are limited to single-speaker anonymization. These methods are not yet directly applicable to multi-speaker or overlapping speech scenarios Miao et al. (2025a); Tomashenko et al. (2025d).

Future voice anonymization research demands datasets that are not only larger but also far more diverse, encompassing data collected from speakers of different ages, accents, languages, speech styles, emotional states, and acoustic conditions. To ensure anonymization techniques are robust and universally applicable, there is also a need for datasets which reflect more challenging and realistic use-case scenarios, such as those involving overlapping speech and low-resource languages; and establishing fairness benchmarks with assessment whether anonymization performance is consistent across speaker demographics (age, gender, accent, language background, speech disorders).

7.2. Emerging and advanced approaches to anonymization

In this section we survey recent advances in anonymization highlighting both methodological innovations and application-specific developments. The majority of recent work continues to build upon disentanglement-based methods, which explicitly decompose speech into independent factors to enable controlled anonymization. Several primary approaches have emerged:

Multi-attribute disentanglement extends single-attribute approaches to enable simultaneous disentanglement of multiple speech attributes (Miao et al., 2025b; Yao et al., 2025). This advancement enables more fine-grained, context-aware anonymization that can simultaneously control voice identity, emotional expression, demographic attributes, and other speaker characteristics, offering greater flexibility for domain-specific applications.

Speaker-embedding-free approaches (Cai et al., 2024; Franzreb et al., 2025; Tang et al., 2025) avoid explicit speaker embeddings entirely. Instead, these approaches learn anonymized representations directly from speech signals, potentially simplifying the anonymization pipeline and improving generalization to unseen acoustic conditions and speaker populations.

Recent work has addressed other, diverse applications and challenging scenarios. It is organized along three primary dimensions: deployment constraints, target populations, and paradigm variants:

Language and cross-lingual generalization: Emerging work in multilingual anonymization (Miao et al., 2022b,a; Deng et al., 2023; Miao et al., 2023; Meyer et al., 2024; Yao et al., 2024b) and code-switching scenarios (Meyer et al., 2025) shows that anonymization models can generalize to other languages and mixed-language speech. Cross-lingual generalization broadens real-world applicability without the need for multilingual training data or any retraining/adaption.

Conversational scenarios and multi-speaker recordings: Many natural use cases for anonymisation involve multiple, concurrent speakers, and complex conversational dynamics that complicate anonymization (Miao et al., 2025a). Recent work (Tomashenko et al., 2025d) addresses the challenge of anonymizing a single target speaker in multi-speaker overlapping conversational recordings, and proposes new methods and improved evaluation strategies to enhance privacy protection and utility assessment in these complex scenarios.

Recording conditions and noisy speech: Only few works address the evaluation of anonymization robustness under noisy, reverberant, or compressed acoustic conditions which typify many real-world applications (e.g., phone calls, live streaming, video conferencing). Work in Miao et al. (2024) shows that the application of anonymization techniques to noisy, spontaneous speech from the *VoxCeleb-2* dataset can distort speech content.

Real-time and low-latency anonymization: is critical for live applications such as streaming, video conferencing, and telephony. Recent work (Quamer and Gutierrez-Osuna, 2024) addresses this challenge by replacing computationally intensive models (e.g., *HuBERT*, HiFi-GAN) with lightweight convolutional neural network architectures and chunk-wise speech processing, achieving anonymization under strict latency constraints without substantial utility degradation.

Target populations and specialized speech domains: Recent work has extended anonymization techniques to underrepresented populations and specialized speech domains. Techniques tailored to children's voices should address unique acoustic characteristics and developmental variations (Kulkarni et al., 2025). Methods for pathological speech (dysarthria, dysphonia, etc.) explored by Tayebi Arasteh et al. (2024) maintain intelligibility and clinical diagnostic utility while protecting voice identity.

Synchronous and asynchronous anonymization: Two complementary anonymization paradigms have emerged, each with distinct use cases (Chen et al., 2025). *Synchronous* anonymization modifies both machine-perceived and human-perceived voice identity, altering speech signals so that both automatic systems and human listeners perceive a different voice identity. This approach aligns with the VPC definition of voice anonymization. *Asynchronous* anonymization modifies only machine-discernible speaker attributes while preserving human perception of the original voice identity.

User control and flexibility: Recent work (Hui et al., 2025) introduces prompt-based speech anonymization leveraging large language model (LLM)-based ASR-TTS frameworks. This approach enables flexible, user-controllable anonymization through natural language prompts that specify desired speaker attributes, while simultaneously addressing content privacy through named entity recognition and substitution. This framework opens new possibilities for context-dependent and task-specific anonymization strategies.

Beyond voice identity: multi-attribute privacy protection: In addition to voice-identity privacy, other attributes may require protection in different scenarios. For example, background environmental sounds (e.g., announcements at a train station) can reveal a speaker's location, necessitating anonymization that considers scenario-dependent audio cues alongside voice identity. Content privacy (Williams et al., 2021, 2022; Sinha et al., 2024; Hui et al., 2025) is also a major concern. Even if the latter is now attracted attention, most approaches treat named entities as the primary privacy-sensitive information and simply remove them — a solution that is not always practical. An attacker may still infer sensitive meaning from surrounding context even if named entities are removed.

Reversibility and recovery requirements: To ensure strong privacy guarantees, most anonymization systems are designed to be irreversible. However, reversibility is required in certain use cases. For example, in forensic or security contexts, the original content and speaker-identity information should be recoverable to facilitate investigations. Recent work (Zhang et al., 2025) explores recoverable anonymization methods enabling the selective restoration or user-customized modification of anonymized speech. This direction addresses forensic and authorized investigation scenarios where conditional decryption and voice recovery are necessary, while maintaining privacy under normal operation. A practical implementation could involve allowing users or authorized parties to maintain a secure mapping between original and anonymized speaker/content, enabling recovery when necessary.

8. Conclusions

The VoicePrivacy 2024 Challenge represents a significant milestone in voice anonymization research, demonstrating substantial progress in privacy-preserving speech technologies which simultaneously preserve linguistic and emotional content. The challenge attracted 36 submissions from 13 teams across 16 countries, representing both academic and non-academic organizations, significant international engagement, and substantial methodological diversity – more than doubling participation compared to the previous challenge edition. Many submitted systems evolved far beyond incremental baseline improvements, introducing state-of-the-art neural architectures combining neural audio codecs, attribute disentanglement, self-supervised learning models, and emotion-aware design.

Results show that strong privacy protection can often only be delivered using complex neural architectures and careful consideration of multiple competing objectives. Only six of the submitted systems achieved the highest privacy target of an EER $\geq$ 40% while only four met the full complement of other challenge requirements, illustrating the fundamental difficulty of protecting privacy while simultaneously preserving utility. Results show that cascaded ASR+TTS systems achieve near-perfect anonymization but at the cost of emotion preservation. Through multi-stage disentanglement and frame-level emotion distillation, neural codec-based approaches achieve a better balance between privacy and utility requirements. Systems without explicit emotion-aware components mostly underperform those with dedicated emotion distillation, encoder integration, or emotion-matched pseudo-speaker selection.

Five dominant technical trends emerged. They include: the adoption of neural codecs providing frameworks for attribute disentanglement; the use of self-supervised learning and pretrained models for feature extraction; multi-stage attribute disentanglement of speaker identity, linguistic content, prosody, and emotion anonymization techniques which enable flexible privacy-utility trade-off through dynamic method selection; emotion-aware design using emotion encoders, distillation mechanisms and emotion-conditioned synthesis.

The best submitted systems significantly outperform baselines across all metrics. Specifically, for individual metrics: EER improves from 34% (**B5**) to 49.5% (**T8-1**), UAR improves from 54% (**B2**) to 65% (**T10-1**), and WER reduces from 2.9% (**B1**) to 2.37% (**T9**).

Some challenges and other concerns persist. Other work Tomashenko et al. (2025a) reveals that privacy evaluation is vulnerable to overestimation: four top-performing systems that initially achieved an EER $\geq$ 40% were later shown to be vulnerable to attacks that reduced EERs to less that 20%. This finding emphasizes the importance of independent and continuous two-stage evaluation in which defenses and attacks are developed and revised by competing teams. This challenge emphasizes the importance of independent or continuous two-stage evaluation frameworks in which anonymization systems and attackers are developed by different teams, ensuring that privacy assessments are based on independently verified threat models rather than single standardized attackers.

Emerging research directions include the study of multi-attribute and application-specific anonymization, real-time and computationally efficient solutions, and semantic-level approaches leveraging prompt-based and LLM-driven methods for flexible, user-controllable anonymization. The evolution from simple baseline approaches to complex multi-stage disentanglement systems reflects rapid methodological development in voice anonymization research. As the field continues to advance toward the next VoicePrivacy Challenge edition tentatively scheduled for 2026, addressing these evaluation limitations, expanding application scenarios, and ensuring fairness across diverse populations will be critical to developing robust, trustworthy, and widely deployable voice anonymization solutions.

Acknowledgment

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/). This work was partially supported by JST, PRESTO Grant Number JPMJPR23P9, Japan. Experiments were carried out using the Grid'5000 testbed. The challenge organizers thank *Ünal Ege Gaznepoğlu* for his help with the code base.

References

Akti, S., Nguyen, T.N., Liu, Y., Waibel, A., 2024a. Investigating voice conversion architecture with different bottleneck features for the voice privacy challenge 2024. The VoicePrivacy 2024 Challenge .

Akti, S., Nguyen, T.N., Liu, Y., Waibel, A., 2024b. Voice privacy-investigating voice conversion architecture with different bottleneck features, in: Proc. SPSC 2024.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M., Weber, G., 2020. Common Voice: a massively-multilingual speech corpus, in: 12th Language Resources and Evaluation Conference (LREC), p. 4218-4222.

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein generative adversarial networks, in: International Conference on Machine Learning (ICML), pp. 214–223.

Article 29 Working Party, 2014. Opinion 05/2014 on anonymisation techniques. https://ec.europa.eu/justice/article-29/. URL: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216en.pdf. endorsed by the European Data Protection Board (EDPB).

Baas, M., van Niekerk, B., Kamper, H., 2023. Voice conversion with just nearest neighbors. arXiv preprint arXiv:2305.18975 .

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. arXiv preprint arXiv:2508.05250 .

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.

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.

Bogdanov, D., Won, M., Tovstogan, P., Porter, A., Serra, X., 2019. The MTG-Jamendo dataset for automatic music tagging, in: ICML 2019 Machine Learning for Music Discovery Workshop.

Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., Roblek, D., Teboul, O., Grangier, D., Tagliasacchi, M., Zeghidour, N., 2023. AudioLM: a language modeling approach to audio generation. arXiv preprint arXiv:2209.03143 .

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. Journal of Language Resources and Evaluation 42, 335–359.

Cai, Z., Xinyuan, H.L., Garg, A., García-Perera, L.P., Duh, K., Khudanpur, S., Andrews, N., Wiesner, M., 2024. Privacy versus emotion preservation trade-offs in emotion-preserving speaker anonymization, in: Proc. SLT, IEEE. pp. 409–414.

Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R., 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE Transactions on Affective Computing 5, 377–390.

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

Chandra, S., Petteno, M., Evans, N., Bäckström, T., Kolossa, D., Martin, R., 2025. Voice anonymisation evaluation using similarity rank disclosure, in: in preparation.

Chen, L., Guo, C., Wang, R., Lee, K.A., Ling, Z.H., 2025. Any-to-any speaker attribute perturbation for asynchronous voice anonymization. IEEE Transactions on Information Forensics and Security .

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.

Cho, K., van Merriënboer, B., Bahdanau, D., Bengio, Y., 2014. On the properties of neural machine translation: Encoder–decoder approaches, in: Wu, D., Carpuat, M., Carreras, X., Vecchi, E.M. (Eds.), Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Association for Computational Linguistics, Doha, Qatar. pp. 103–111. URL: https://aclanthology.org/W14-4012/, doi:10.3115/v1/W14-4012.

Chouchane, O., Busch, C., Galdi, C., Evans, N., Todisco, M., 2024. A comparison of differential performance metrics for the evaluation of automatic speaker verification fairness, in: Proc. Odyssey 2024, pp. 209–216.

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

Chung, Y.A., Zhang, Y., Han, W., Chiu, C.C., Qin, J., Pang, R., Wu, Y., 2021. W2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training, in: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE. pp. 244–250.

Das, A., Franzreb, C., Herzig, T., Pirlet, P., Polzehl, T., 2024. System description for Voice Privacy challenge 2024, in: The VoicePrivacy 2024 Challenge.

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.

Deng, J., Teng, F., Chen, Y., Chen, X., Wang, Z., Xu, W., 2023. {V-Cloak}: Intelligibility-, naturalness-& {Timbre-Preserving}{Real-Time} voice anonymization, in: 32nd USENIX Security Symposium (USENIX Security 23), pp. 5181–5198.

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.

Dubey, H., Gopal, V., Cutler, R., Aazami, A., Matusevych, S., Braun, S., Eskimez, S.E., Thakker, M., Yoshioka, T., Gamper, H., Aichner, R., 2022. ICASSP 2022 Deep Noise Suppression Challenge, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 9271–9275.

Eyben, F., Wöllmer, M., Schuller, B., 2010. Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM International Conference on Multimedia, Association for Computing Machinery, New York, NY, USA. p. 1459–1462. URL: https://doi.org/10.1145/1873951.1874246, doi:10.1145/1873951.1874246.

Fonseca, E., Favory, X., Pons, J., Font, F., Serra, X., 2022. Fsd50k: An open dataset of human-labeled sound events. IEEE/ACM Transactions on Audio, Speech, and Language Processing 30, 829–852. doi:10.1109/TASLP.2021.3133208.

Franzreb, C., Das, A., Polzehl, T., Möller, S., 2025. Private knn-vc: Interpretable anonymization of converted speech, in: Proc. Interspeech.

Freitas Pereira, T., Marcel, S., 2021. Fairness in biometrics: a figure of merit to assess biometric verification systems. IEEE Transactions on Biometrics, Behavior, and Identity Science 4, 19–29.

Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M., 2017. Audio set: An ontology and human-labeled dataset for audio events, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 776–780. doi:10.1109/ICASSP.2017.7952261.

Ghosh, S., Tyagi, U., Ramaneswaran, S., Srivastava, H., Manocha, D., 2023. Mmer: Multimodal multi-task learning for speech emotion recognition, in: Interspeech 2023, pp. 1209–1213. doi:10.21437/Interspeech.2023-2271.

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.

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.

He, J., Zhou, J., Sun, H., Wang, H., Qin, Y., 2024. FaCodec-based anonymization solution enhanced with prosody anonymization. The VoicePrivacy 2024 Challenge .

Howard, J.J., Laird, E.J., Rubin, R.E., Sirotin, Y.B., Tipton, J.L., Vemury, A.R., 2022. Evaluating proposed fairness models for face recognition algorithms, in: International Conference on Pattern Recognition, Springer. pp. 431–447.

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.

Hua, H., Shang, Z., Li, X., Shi, P., Yang, C., Wang, L., Zhang, P., 2024. Emotional speech anonymization: Preserving emotion characteristics in pseudo-speaker speech generation, in: Proc. SPSC 2024, pp. 55–60.

Hui, B.S.H., Miao, X., Wang, X., 2025. Securespeech: Prompt-based speaker and content protection. arXiv preprint arXiv:2507.07799 .

Ito, K., Johnson, L., 2017. The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/.

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 .

Kahn, J., Rivière, M., Zheng, W., Kharitonov, E., Xu, Q., Mazaré, P.E., Karadayi, J., Liptchinsky, V., Collobert, R., Fuegen, C., Likhomanenko, T., Synnaeve, G., Joulin, A., Mohamed, A., Dupoux, E., 2020. Libri-light: A benchmark for ASR with limited or no supervision, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7669–7673.

Kasi, K., Zahorian, S.A., 2002. Yet another algorithm for pitch tracking, in: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. I–361–I–364. doi:10.1109/ICASSP.2002.5743729.

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.

Kong, J., Kim, J., Bae, J., 2020. Hifi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. Advances in Neural Information Processing Systems 33, 17022–17033.

Kulkarni, A., Teixeira, F., Hermann, E., Rolland, T., Trancoso, I., Doss, M.M., 2025. Children's Voice Privacy: First Steps and Emerging Challenges, in: Interspeech 2025, pp. 2810–2814. doi:10.21437/Interspeech.2025-2117.

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 .

Le Blouch, O., Bakari, R., Gengembre, N., 2024a. Orange SHIVA system description for the Voice Privacy challenge 2024. The VoicePrivacy 2024 Challenge .

Le Blouch, O., Bakari, R., Gengembre, N., 2024b. Tuning DISSC for voice privacy challenge 2024, in: Proc. SPSC 2024.

Leang, S., Augusma, A., Vaufreydaz, D., Castelli, E., Sam, S., Letué, F., 2024. Exploring vector-quantized variational auto-encoder with prosody parameters for speaker anonymization. The VoicePrivacy 2024 Challenge .

Lee, J., Park, T., You, Y., 2024a. System description for Voice Privacy Challenge 2024. The VoicePrivacy 2024 Challenge .

Lee, J., Park, T., You, Y., 2024b. Voice anonymization using emotion-enriched feature integration with STT and TTS models, in: Proc. SPSC 2024, pp. 50–54.

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

Li, Y., Zheng, Y., Fang, J., Chen, J., 2024. Emotion-enhanced speaker anonymisation using the FreeVC framework. The VoicePrivacy 2024 Challenge .

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.

Lotfian, R., Busso, C., 2019. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. IEEE Transactions on Affective Computing 10, 471–483. doi:10.1109/TAFFC.2017.2736999.

Lux, F., Koch, J., Schweitzer, A., Vu, N.T., 2021. The IMS Toucan system for the Blizzard Challenge 2021, in: Blizzard Challenge Workshop.

Maimon, G., Adi, Y., 2023a. Speaking style conversion in the waveform domain using discrete self-supervised units, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore. pp. 8048–8061. URL: https://aclanthology.org/2023.findings-emnlp.541/, doi:10.18653/v1/2023.findings-emnlp.541.

Maimon, G., Adi, Y., 2023b. Speaking style conversion in the waveform domain using discrete self-supervised units. URL: https://arxiv.org/abs/2212.09730, arXiv:2212.09730.

McAdams, S., 1984. Spectral fusion, spectral parsing and the formation of the auditory image. Ph.D. thesis. Stanford University.

Meyer, S., Kolos, E., Vu, N.T., 2025. First steps towards voice anonymization for code-switching speech. Proc. Interspeech .

Meyer, S., Lux, F., Koch, J., Denisov, P., Tilli, P., Vu, N.T., 2023a. 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.

Meyer, S., Lux, F., Vu, N.T., 2024. Probing the feasibility of multilingual speaker anonymization. Accepted by INTERSPEECH 2024 .

Meyer, S., Miao, X., Vu, N.T., 2023b. VoicePAT: An efficient open-source evaluation toolkit for voice privacy research. IEEE Open Journal of Signal Processing .

Meyer, S., Tilli, P., Denisov, P., Lux, F., Koch, J., Vu, N.T., 2023c. Anonymizing speech with generative adversarial networks to preserve speaker privacy, in: IEEE Spoken Language Technology Workshop (SLT), pp. 912–919.

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

Miao, X., Wang, X., Cooper, E., Yamagishi, J., Evans, N., Todisco, M., Bonastre, J.F., Rouvier, M., 2024. Synvox2: Towards a privacy-friendly voxceleb2 dataset, in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 11421–11425.

Miao, X., Wang, X., Cooper, E., Yamagishi, J., Tomashenko, N., 2022a. Analyzing language-independent speaker anonymization framework under unseen conditions, in: Proc. Interspeech 2022, pp. 4426–4430.

Miao, X., Wang, X., Cooper, E., Yamagishi, J., Tomashenko, N., 2022b. Language-independent speaker anonymization approach using self-supervised pre-trained models. arXiv preprint arXiv:2202.13097 .

Miao, X., Wang, X., Cooper, E., Yamagishi, J., Tomashenko, N., 2023. Speaker anonymization using orthogonal Householder neural network. IEEE/ACM Trans. Audio, Speech, and Language Processing 31, 3681–3695. doi:10.1109/TASLP.2023.3313429.

Miao, X., Zhang, Y., Wang, X., Tomashenko, N., Soh, D.C.L., Mcloughlin, I., 2025b. Adapting general disentanglement-based speaker anonymization for enhanced emotion preservation. Computer Speech & Language , 101810.

Nautsch, A., Jimenez, A., Treiber, A., Kolberg, J., Jasserand, C., Kindt, E., Delgado, H., Todisco, M., Hmani, M.A., Mtibaa, A., Abdelraheem, M.A., Abad, A., Teixeira, F., Matrouf, D., Gomez-Barrero, M., Petrovska-Delacrétaz, D., Chollet, G., Evans, N., Schneider, T., Bonastre, J.F., Busch, C., 2019. Preserving privacy in speaker and speech characterisation. Computer Speech and Language 58, 441–480.

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.

Noé, P.G., Bonastre, J.F., Matrouf, D., Tomashenko, N., Nautsch, A., Evans, N., 2020. Speech pseudonymisation assessment using voice similarity matrices, in: Interspeech, pp. 1718–1722.

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.

O'Brien, B., Tomashenko, N., Chanclu, A., Bonastre, J.F., 2021. Anonymous speaker clusters: Making distinctions between anonymised speech recordings with clustering interface, in: Interspeech, pp. 3580–3584.

Panariello, M., Meyer, S., Champion, P., Miao, X., Todisco, M., Vu, N.T., Evans, N., 2025. The risks and detection of overestimated privacy protection in voice anonymisation. SPSC 2025 .

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.

Pappagari, R., Wang, T., Villalba, J., Chen, N., Dehak, N., 2020. X-vectors meet emotions: A study on dependencies between emotion and speaker recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7169–7173.

Patino, J., Tomashenko, N., Todisco, M., Nautsch, A., Evans, N., 2021. Speaker anonymisation using the McAdams coefficient, in: Interspeech, pp. 1099–1103.

Peng, Y., Dalmia, S., Lane, I., Watanabe, S., 2022. Branchformer: Parallel MLP-attention architectures to capture local and global context for speech recognition and understanding, in: International Conference on Machine Learning (ICML), pp. 17627–17643.

Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., Khudanpur, S., 2018. Semi-orthogonal low-rank matrix factorization for deep neural networks., in: Interspeech, pp. 3743–3747.

Qian, J., Han, F., Hou, J., Zhang, C., Wang, Y., Li, X.Y., 2018. Towards privacy-preserving speech data publishing, in: IEEE Conference on Computer Communications (INFOCOM), pp. 1079–1087.

Qian, K., Zhang, Y., Gao, H., Ni, J., Lai, C.I., Cox, D., Hasegawa-Johnson, M., Chang, S., 2022. Contentvec: An improved self-supervised speech representation by disentangling speakers, in: International Conference on Machine Learning, PMLR. pp. 18003–18017.

Quamer, W., Gutierrez-Osuna, R., 2024. End-to-end streaming model for low-latency speech anonymization, in: 2024 IEEE Spoken Language Technology Workshop (SLT), IEEE. pp. 727–734.

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 .

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.Y., 2020. FastSpeech 2: Fast and high-quality end-to-end text to speech, in: International Conference on Learning Representations (ICLR).

Sharma, G., Dhall, A., Cai, J., 2021. Audio-visual automatic group affect analysis. IEEE Transactions on Affective Computing 14, 1056–1069.

Sinha, Y., Raivakhovskyi, M., Schubert, M., Siegert, I., 2024. Safeguarding speech content style: Enhancing privacy beyond speaker identity. Proc. SPSC 2024 , 92–101.

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

Srivastava, B.M.L., Maouche, M., Sahidullah, M., Vincent, E., Bellet, A., Tommasi, M., Tomashenko, N., Wang, X., Yamagishi, J., 2022. Privacy and utility of x-vector based speaker anonymization. IEEE/ACM Transactions on Audio, Speech and Language Processing 30, 2383–2395.

Srivastava, B.M.L., Tomashenko, N., Wang, X., Vincent, E., Yamagishi, J., Maouche, M., Bellet, A., Tommasi, M., 2020a. Design choices for x-vector based speaker anonymization, in: Interspeech, pp. 1713–1717.

Srivastava, B.M.L., Vauquier, N., Sahidullah, M., Bellet, A., Tommasi, M., Vincent, E., 2020b. Evaluating voice conversion-based privacy protection against informed attackers, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2802–2806.

Tan, T., Liu, S., Duan, Y., Zhao, S., Shao, X., 2024. System description: Speaker anonymization system with sentiment transfer and feature interpolation. The VoicePrivacy 2024 Challenge .

Tang, B., Miao, X., Wang, X., Li, M., 2025. Sef-mk: Speaker-embedding-free voice anonymization through multi-k-means quantization. Proc. ASRU .

Tayebi Arasteh, S., Arias-Vergara, T., Pérez-Toro, P.A., Weise, T., Packhäuser, K., Schuster, M., Noeth, E., Maier, A., Yang, S.H., 2024. Addressing challenges in speaker anonymization to maintain utility while ensuring privacy of pathological speech. Communications Medicine 4, 182.

Thienpondt, J., Demuynck, K., 2023. ECAPA2: A hybrid neural network architecture and training strategy for robust speaker embeddings, in: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 1–8.

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., 2020a. The VoicePrivacy 2020 Challenge evaluation plan. https://www.voiceprivacychallenge.org/vp2020/docs/VoicePrivacy_2020_Eval_Plan_v1_4.pdf.

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., 2020b. 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.

Tomashenko, N., Wang, X., Miao, X., Nourtel, H., Champion, P., Todisco, M., Vincent, E., Evans, N., Yamagishi, J., Bonastre, J.F., 2022a. The VoicePrivacy 2022 Challenge evaluation plan. arXiv preprint arXiv:2203.12468 .

Tomashenko, N., Wang, X., Miao, X., Nourtel, H., Champion, P., Todisco, M., Vincent, E., Evans, N., Yamagishi, J., Bonastre, J.F., Panariello, M., 2022b. The VoicePrivacy 2022 Challenge URL: https://www.voiceprivacychallenge.org/vp2022/docs/VoicePrivacy_2022_Challenge___Natalia_Tomashenko.pdf.

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., 2021. Supplementary material to the paper. The VoicePrivacy 2020 Challenge: Results and findings. https://hal.archives-ouvertes.fr/hal-03335126.

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., 2022c. 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., 2025d. Target speaker anonymization in multi-speaker recordings. arXiv preprint arXiv:2510.09307 .

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.

Veaux, C., Yamagishi, J., MacDonald, K., 2019. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92). https://datashare.is.ed.ac.uk/handle/10283/3443.

Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., Schuller, B.W., 2023. Dawn of the transformer era in speech emotion recognition: closing the valence gap. IEEE Transactions on Pattern Analysis and Machine Intelligence .

Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S., Wei, F., 2023. Neural codec language models are zero-shot text to speech synthesizers. arXiv preprint arXiv:2301.02111 .

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, pp. 1344–1348.

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.

Williams, J., Pizzi, K., Das, S., Noé, P.G., 2022. New challenges for content privacy in speech and audio, in: Proc. SPSC 2022, pp. 1–6.

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.

Williams, J., Yamagishi, J., Valentini-Botinhao, C., Bonastre, J.F., et al., 2021. Revisiting speech content privacy, in: Proc. SPSC 2021, pp. 42–46.

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 .

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

Yao, J., Liu, H., Chng, E.S., Xie, L., 2025. EASY: Emotion-aware Speaker Anonymization via Factorized Distillation, in: Interspeech 2025, pp. 3219–3223. doi:10.21437/Interspeech.2025-194.

Yao, J., Wang, Q., Guo, P., Ning, Z., Yang, Y., Pan, Y., Xie, L., 2024b. MUSA: Multi-lingual speaker anonymization via serial disentanglement. arXiv preprint arXiv:2407.11629 .

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.

Zanon Boito, M., Besacier, L., Tomashenko, N., Estève, Y., 2022. A study of gender impact in self-supervised models for speech-to-text systems, in: Proc. Interspeech 2022, pp. 1278–1282.

Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R.J., Jia, Y., Chen, Z., Wu, Y., 2019. LibriTTS: A corpus derived from LibriSpeech for text-to-speech, in: Interspeech, pp. 1526–1530.

Zhang, J., Liu, S., Hou, J., Wang, Z., Yu, H., Li, X.Y., 2025. {SpeechGuard}: Recoverable and customizable speech privacy protection, in: 34th USENIX Security Symposium (USENIX Security 25), pp. 5931–5948.

Zhou, K., Sisman, B., Liu, R., Li, H., 2021. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 920–924.

A. Training resources

Table 8 presents the list of allowed training resources for anonymization systems.

Table 8: List of models and data for training anonymization systems. ⋆ denotes multiple requests from different teams for the corresponding model/dataset.

#	Model
1	WavLM Base and Large (Chen et al., 2022) ⋆ ⋆ ⋆⋆ https://github.com/microsoft/unilm/tree/master/wavlm
2	Whisper (Radford et al., 2023) ⋆⋆ https://github.com/openai/whisper
3	HuBERT (Hsu et al., 2021) ⋆ ⋆ ⋆ https://github.com/facebookresearch/fairseq/blob/main/examples/hubert
4	XLS-R (Babu et al., 2021) ⋆⋆ https://github.com/facebookresearch/fairseq/blob/main/examples/wav2vec/xlsr
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

6	wav2vec2-large-robust-12-ft-emotion-msp-dim (Wagner et al., 2023) https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim
7	ContentVec (Qian et al., 2022); https://github.com/auspicious3000/contentvec
8	w2v-BERT (Chung et al., 2021) https://github.com/facebookresearch/fairseq/tree/ust/examples/w2vbert
9	ECAPA2 (Thienpondt and Demuynck, 2023); https://huggingface.co/Jenthe/ECAPA2
10	ECAPA-TDNN (Desplanques et al., 2020); https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb
11	NaturalSpeech 3 (Ju et al., 2024); https://huggingface.co/amphion/naturalspeech3_facodec
12	NVIDIA Hifi-GAN Vocoder (en-US) (Kong et al., 2020) ; https://huggingface.co/nvidia/tts_hifigan
13	CRDNN on CommonVoice 14.0 English https://huggingface.co/speechbrain/asr-crdnn-commonvoice-14-en
14	Encodec (Défossez et al., 2023); https://huggingface.co/facebook/encodec_24khz
15	Bark; https://huggingface.co/suno/bark; https://huggingface.co/erogol/bark/tree/main

#	Dataset
16	ESD (Zhou et al., 2021) ★ ★ ★★ https://hltsingapore.github.io/ESD/download.html
17	LibriSpeech (Panayotov et al., 2015): train-clean-100, train-clean-360, train-other-500 ★ ★ ★ ★ ★ https://www.openslr.org/12
18	CREMA-D (Cao et al., 2014) ★★ https://github.com/CheyneyComputerScience/CREMA-D
19	RAVDESS (Livingstone and Russo, 2018) https://datasets.activeloop.ai/docs/ml/datasets/ravdess-dataset/ https://zenodo.org/records/1188976
20	VCTK (Veaux et al., 2019) ★ ★ ★ ★ ★★ https://datashare.ed.ac.uk/handle/10283/2651; https://huggingface.co/datasets/vctk
21	SAVEE (Haq et al., 2009); http://kahlan.eps.surrey.ac.uk/savee/ https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee
22	EMO-DB (Burkhardt et al., 2005); http://emodb.bilderbar.info/download/
23	LJSpeech (Ito and Johnson, 2017); https://keithito.com/LJ-Speech-Dataset/
24	Libri-light (Kahn et al., 2020) (only train part) ★★ https://github.com/facebookresearch/libri-light/blob/main/data_preparation/README.md
25	VoxCeleb-1,2 (Chung et al., 2018) ★ ★ ★ https://www.robots.ox.ac.uk/~vgg/data/voxceleb/index.html#about
26	LibriTTS (Zen et al., 2019): train-clean-100, train-clean-360, train-other-500 ★ ★ ★★ https://openslr.org/60/
27	CMU-MOSEI (Zadeh et al., 2018); http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/
28	MUSAN (Snyder et al., 2015); https://www.openslr.org/17/
29	RIR (Ko et al., 2017); https://www.openslr.org/28/
30	VGAF (Sharma et al., 2021) (from EmotiW challenge); https://sites.google.com/view/emotiw2023 https://www.kaggle.com/datasets/amirabdrahimov/vgaf-dataset
31	MSP-Podcast (Lotfian and Busso, 2019)

https://ecs.utdallas.edu/research/researchlabs/msp-lab/MSP-Podcast.html

#	Software with pretrained models
32	Resemblyzer: https://github.com/resemble-ai/Resemblyzer ⋆ Model: https://github.com/resemble-ai/Resemblyzer/blob/master/resemblyzer/pretrained.pt
33	VITS (Kim et al., 2021); https://github.com/jaywalnut310/vits/ Models: https://drive.google.com/drive/folders/1ksarh-cJf3F5eKJjLVWY0X1j1qsQqiS2
34	PIPER pretrained on VITS: https://github.com/rhasspy/piper/?tab=readme-ov-file Models: https://huggingface.co/datasets/rhasspy/piper-checkpoints/tree/main
35	RVC-Project: https://github.com/RVC-Project Models: https://huggingface.co/lj1995/VoiceConversionWebUI/tree/main
36	DISSC (Maimon and Adi, 2023b); https://github.com/gallilmaimon/DISSC

B. Results

Table 9 shows performance metrics by system on development and test data.

Figure 13 demonstrates normalized performance matrix showing all systems ranked by overall score (unofficial ranking). The heatmap presents a normalized performance comparison across multiple speech anonymization systems evaluated on three key metrics (relative change w.r.t to the results on the original data): ΔUAR, ΔWER, and ΔEER. Values are normalized on a scale from 0 to 1, where 1 corresponds to the best performance and 0 to the worst. The ΔUAR and ΔWER metrics are inverted during normalization to reflect that lower values indicate better utility (less utility loss), while ΔEER values are kept as is to indicate higher privacy. Systems are ranked by a combined score, averaging their normalized scores across all three metrics, and the heatmap is sorted accordingly. Color coding from red to green visually highlights system strengths and weaknesses, with green indicating excellent performance and red indicating poorer performance.

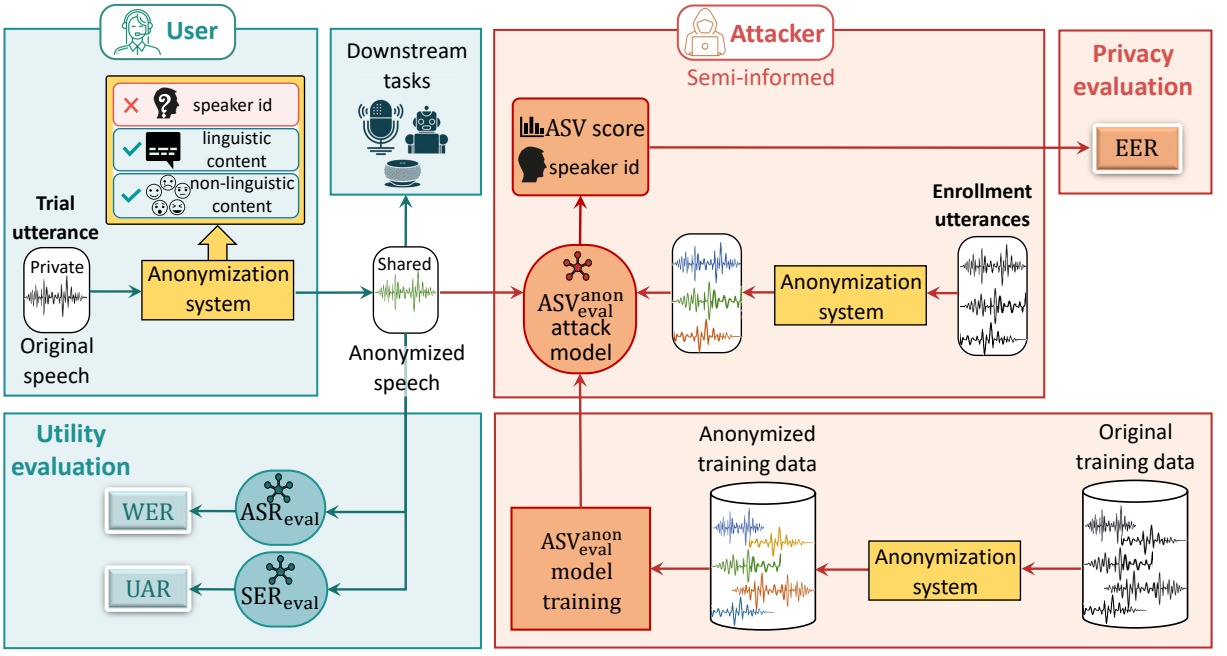

Figure 1: Privacy preservation scenario as a game between *users* and *attackers* in the case where speaker identity is considered as personal information to be protected using anonymization, while linguistic and emotional content should be preserved for utility downstream tasks. Privacy evaluation of anonymized data is performed using a *semi-informed* attack model ASV^{anon}_{eval} and equal error rate (EER) metric. Utility evaluation is performed using (1) automatic speech recognition (ASR) model ASR_{eval} and word error rate (WER) metric and (2) speech emotion recognition (SER) model SER_{eval} and unweighted average recall (UAR) metric.

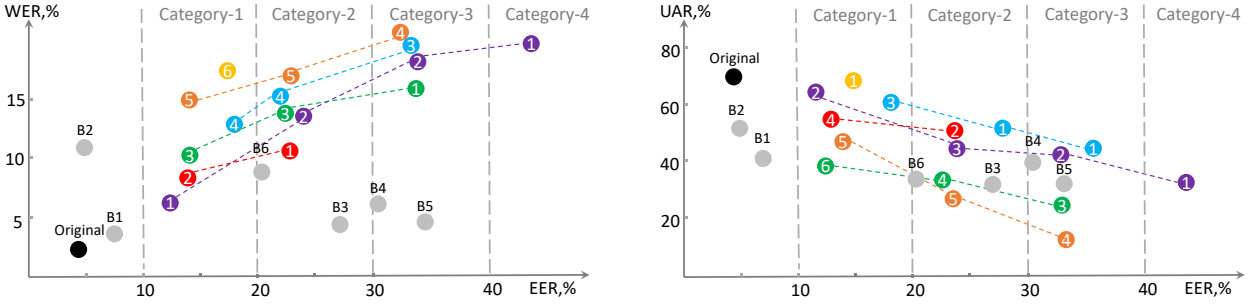

Figure 2: Example system rankings according to the privacy (EER) and utility (WER and UAR) results for 4 minimum target EERs. Different colors correspond to 6 different teams. Numbers within each circle show system ranks for a given category. Grey circles correspond to the baseline systems, and the black circle to the original (unprocessed) data.

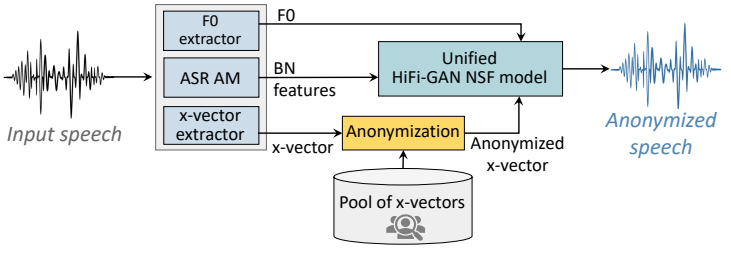

Figure 3: Baseline anonymization system **B1**.

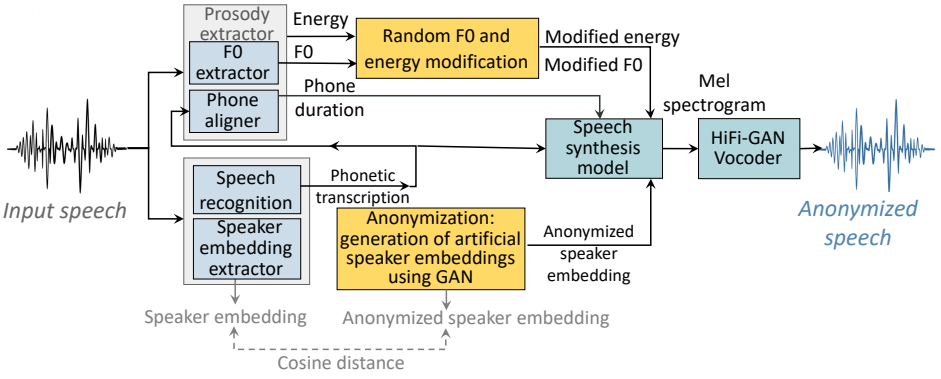

Figure 4: Baseline anonymization system **B3**.

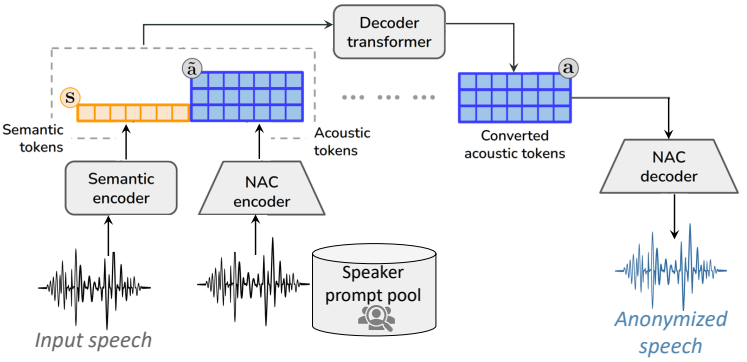

Figure 5: Baseline anonymization system **B4**.

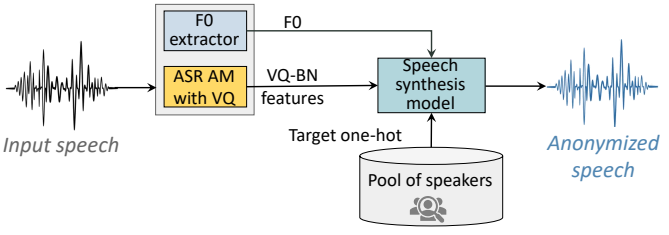

Figure 6: Baseline anonymization systems **B5** and **B6**.

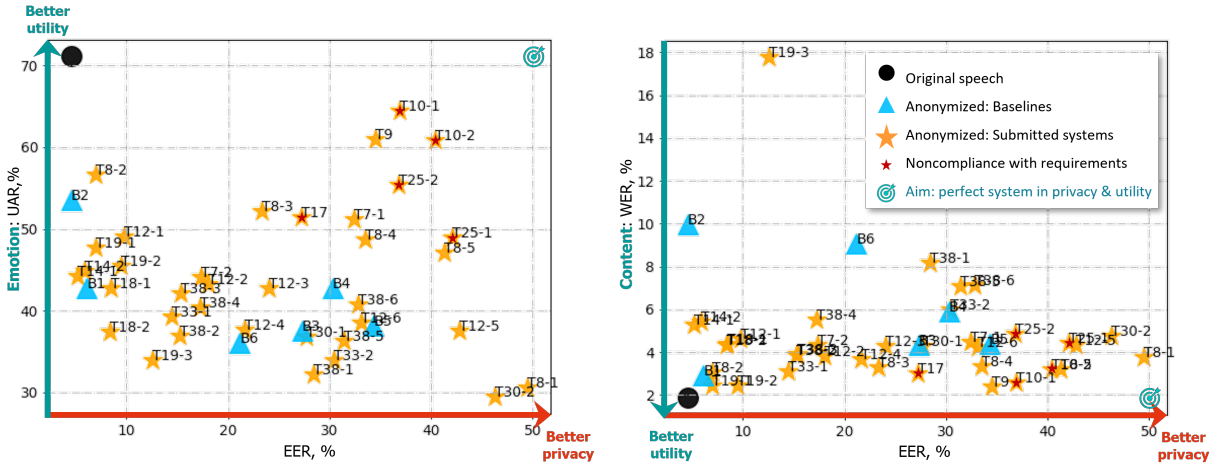

Figure 7: Challenge results on the test data

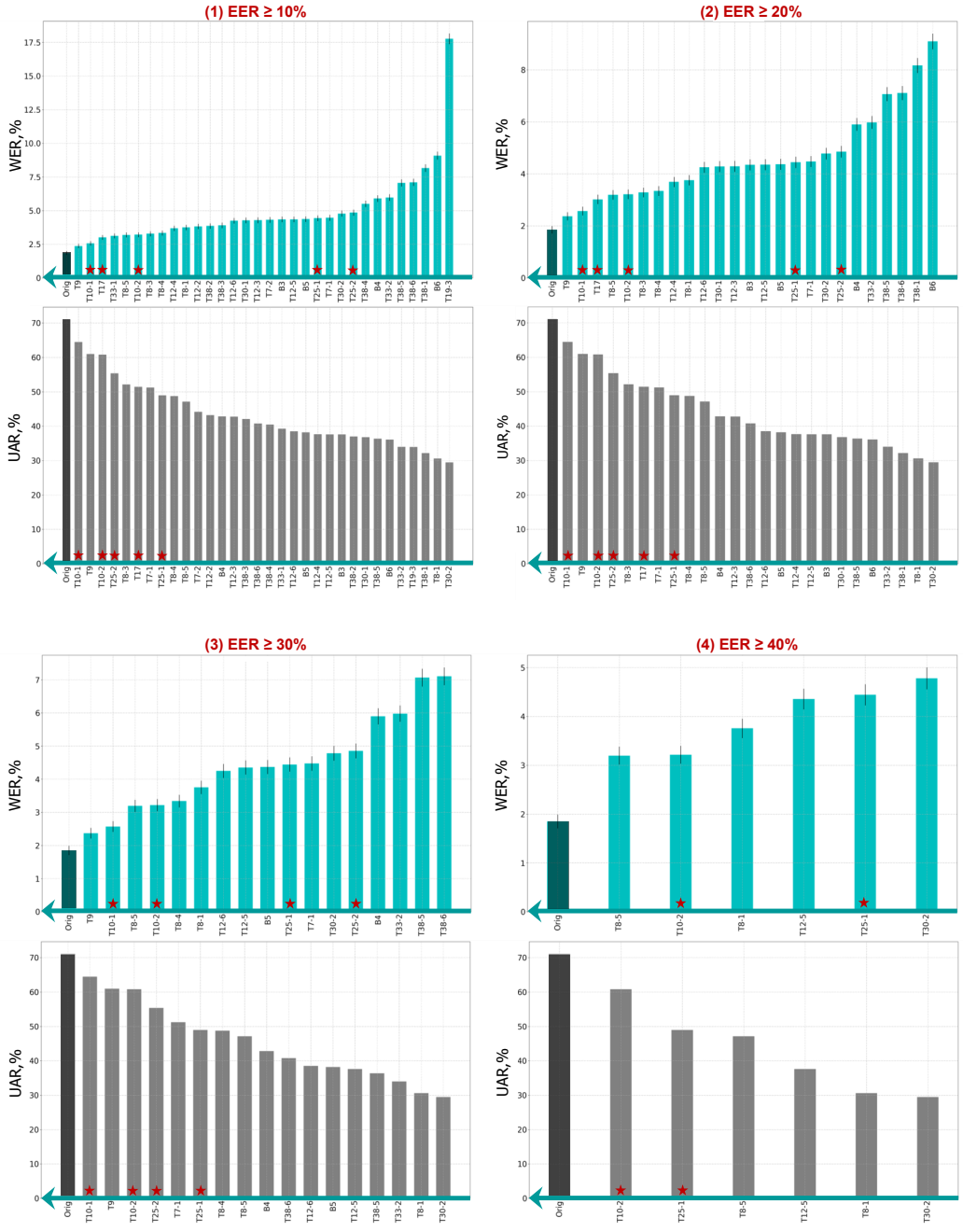

Figure 8: Anonymization system ranking according to the utility metrics (WER,% and UAR,%) for 4 privacy thresholds on the test data (official ranking).

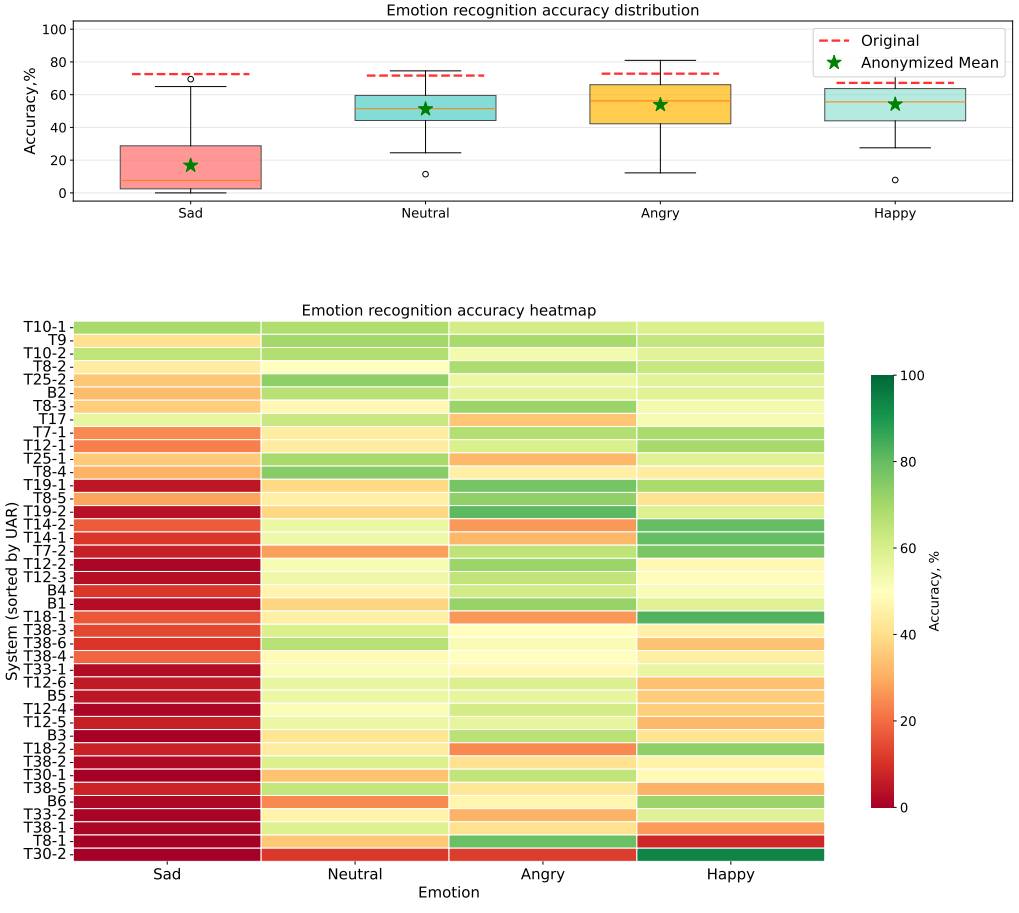

Figure 9: Emotion recognition accuracy by emotion class on the *IEMOCAP* test set

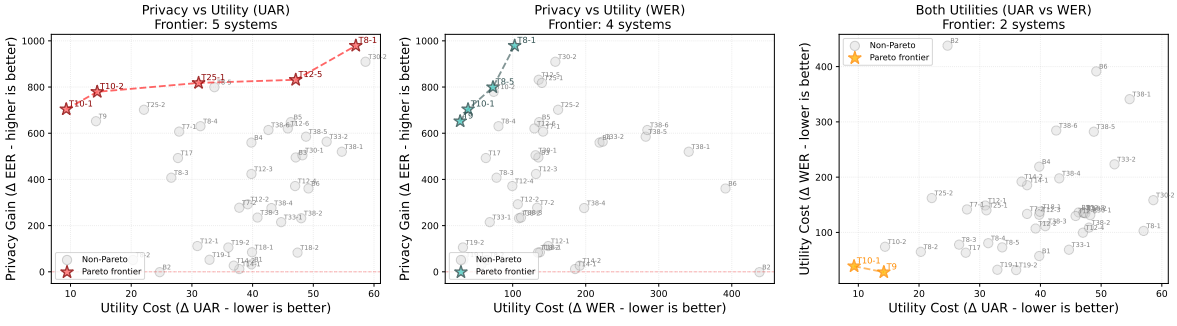

Figure 10: Pareto curve analysis on the test sets

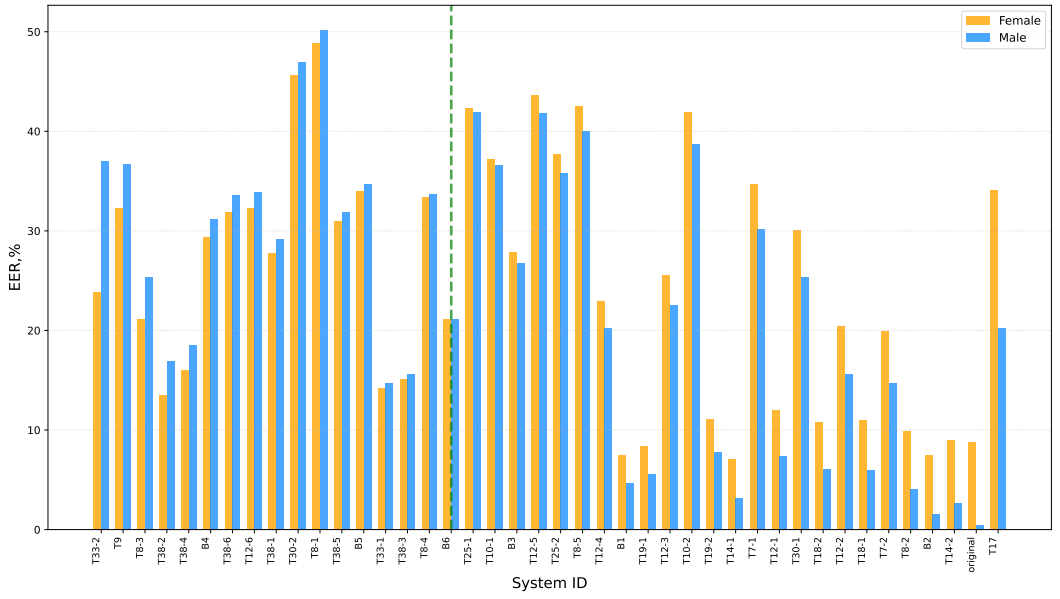

Figure 11: Gender bias in anonymization: systems ordered by male-female EER difference

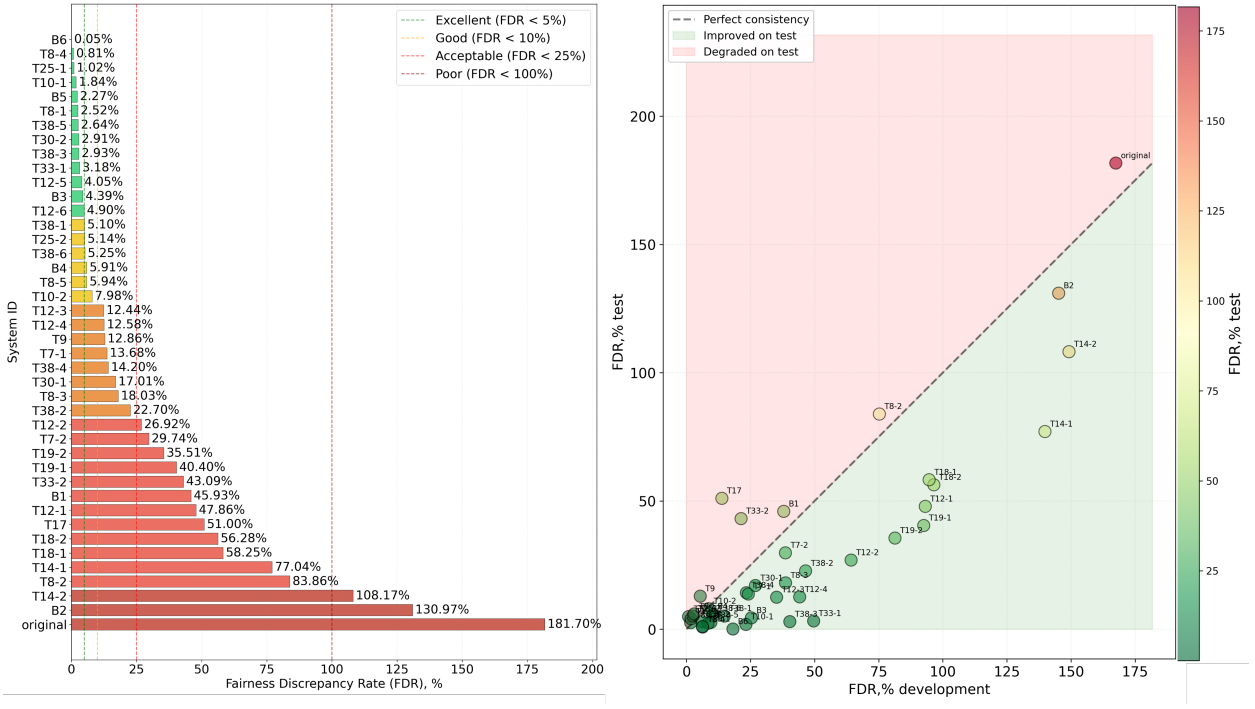

Figure 12: Gender fairness analysis in anonymization: (1) System ranking according mFDR results on the test set; (2) mFDR consistency across development and test sets.

Table 9
Performance metrics by system: EER,% and WER.% on *LibriSpeech* and UAR,% on *IEMOCAP* development and test sets.

System ID	development					test				
	EER, %			UAR, %	WER, %	EER, %			UAR, %	WER, %
	female	male	aver			female	male	aver		
orig	10.51	0.93	5.72	69.08	1.80	8.76	0.42	4.59	71.10	1.85
B1	10.94	7.45	9.20	42.71	3.07	7.47	4.68	6.08	42.80	2.91
B2	12.91	2.05	7.48	55.61	10.44	7.48	1.56	4.52	53.50	9.95
B3	28.43	22.04	25.24	38.09	4.29	27.92	26.72	27.32	37.60	4.35
B4	34.37	31.06	32.72	41.97	6.15	29.37	31.16	30.27	42.80	5.90
B5	35.82	32.92	34.37	38.08	4.73	33.95	34.73	34.34	38.20	4.37
B6	25.14	20.96	23.05	36.39	9.69	21.15	21.14	21.15	36.10	9.09
T7-1	35.23	27.64	31.43	48.70	4.86	34.67	30.23	32.45	51.22	4.47
T7-2	21.57	14.59	18.08	42.38	4.59	19.89	14.74	17.32	44.18	4.32
T8-1	47.33	48.12	47.72	30.08	3.74	48.90	50.15	49.53	30.59	3.75
T8-2	11.65	5.28	8.46	56.73	3.27	9.85	4.03	6.94	56.67	3.05
T8-3	25.00	16.90	20.95	51.28	3.30	21.19	25.39	23.29	52.13	3.29
T8-4	34.93	32.80	33.86	49.34	3.51	33.40	33.67	33.53	48.73	3.34
T8-5	39.63	40.84	40.24	47.08	3.45	42.50	40.05	41.28	47.10	3.20
T9	32.10	33.88	32.99	60.66	2.33	32.30	36.74	34.52	60.97	2.37
T10-1	43.33	34.32	38.82	65.98	2.57	37.24	36.56	36.90	64.45	2.57
T10-2	43.63	40.04	41.83	62.93	3.68	41.97	38.75	40.36	60.80	3.22
T12-1	19.18	6.99	13.08	49.20	4.97	12.04	7.39	9.71	49.12	4.60
T12-2	22.44	11.53	16.99	42.76	3.81	20.44	15.59	18.02	43.21	3.83
T12-3	26.14	18.32	22.23	44.67	4.21	25.53	22.54	24.03	42.78	4.29
T12-4	24.57	15.68	20.13	39.18	3.61	22.99	20.27	21.63	37.67	3.69
T12-5	43.32	44.10	43.71	38.06	4.76	43.61	41.88	42.75	37.60	4.36
T12-6	33.10	33.38	33.24	39.41	4.73	32.27	33.89	33.08	38.51	4.25
T14-1	19.32	3.42	11.37	45.45	6.14	7.12	3.16	5.14	44.23	5.28
T14-2	17.33	2.52	9.92	45.57	6.59	8.96	2.67	5.81	44.85	5.40
T17	36.08	41.44	38.76	51.71	2.95	34.13	20.26	27.20	51.41	3.02
T18-1	14.77	5.28	10.03	45.09	4.46	10.95	6.01	8.48	42.74	4.39
T18-2	13.78	4.81	9.29	43.52	4.79	10.77	6.04	8.40	37.37	4.34
T19-1	11.79	4.33	8.06	46.94	2.60	8.39	5.57	6.98	47.66	2.45
T19-2	16.62	7.01	11.82	43.20	2.48	11.11	7.76	9.43	45.50	2.44
T25-1	42.65	40.06	41.35	52.38	4.73	42.34	41.91	42.13	48.99	4.44
T25-2	36.65	35.72	36.18	55.31	5.19	37.74	35.85	36.79	55.39	4.85
T30-2	45.59	48.45	47.02	28.45	5.02	45.65	47.00	46.33	29.44	4.78
T30-1	29.10	22.20	25.65	38.48	4.13	30.11	25.39	27.75	36.78	4.28
T33-2	29.71	36.80	33.25	37.31	6.17	23.89	37.01	30.45	34.00	5.98
T33-1	22.16	13.35	17.76	38.69	3.16	14.23	14.69	14.46	39.27	3.12
T38-1	40.31	34.77	37.54	32.30	8.47	27.73	29.18	28.45	32.17	8.16
T38-2	25.43	15.84	20.63	37.86	4.18	13.51	16.97	15.24	36.98	3.86
T38-3	25.00	16.61	20.81	44.51	4.20	15.15	15.60	15.37	42.08	3.91
T38-4	27.70	21.90	24.80	42.51	5.85	16.03	18.48	17.26	40.45	5.51
T38-5	41.19	37.44	39.31	36.08	7.78	31.02	31.85	31.44	36.37	7.07
T38-6	42.58	38.35	40.47	41.19	7.80	31.91	33.63	32.77	40.79	7.11
T25-p1	45.12	41.22	43.17	53.84	4.72	44.19	41.40	42.80	50.07	4.45
T25-p2	38.29	34.11	36.20	57.86	4.98	37.94	35.51	36.73	55.79	5.12

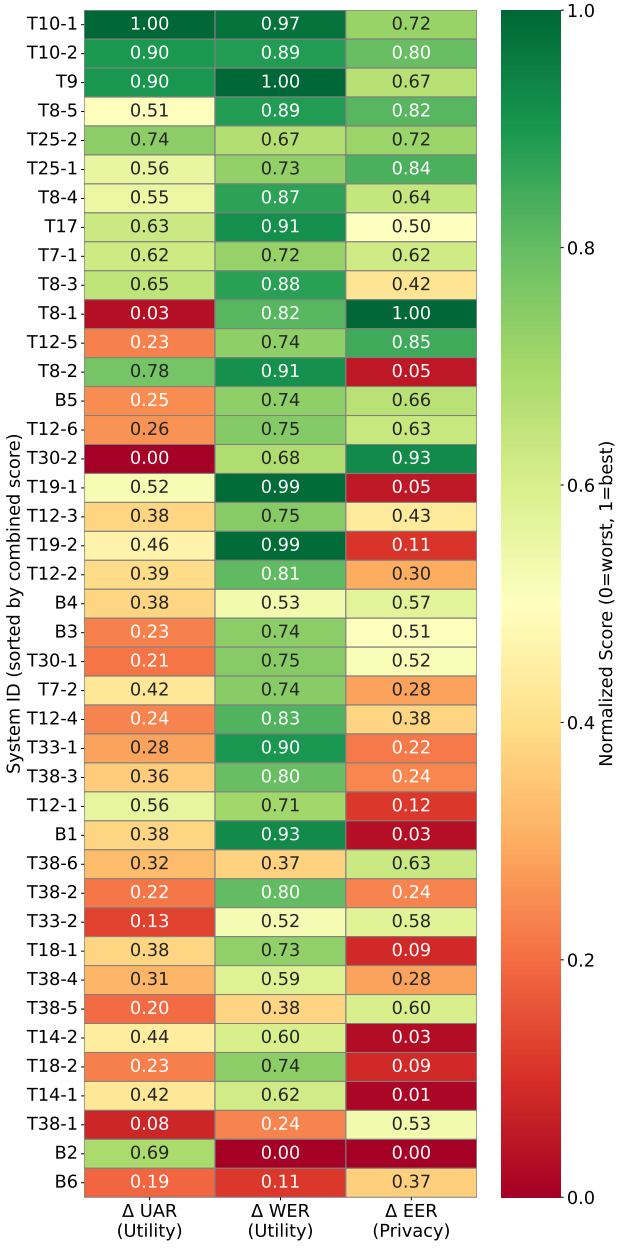

Figure 13: Normalized performance matrix showing all systems ranked by overall score (unofficial ranking).