The VoicePrivacy 2026 Challenge Evaluation Plan

Version **0.0**

Xiaoxiao Miao[1], Natalia Tomashenko[2], Ridwan Arefeen[3], Sarina Meyer[4], Michele Panariello[6], Xin Wang[5], Emmanuel Vincent[2], Nicholas Evans[6], Junichi Yamagishi[5], and Massimiliano Todisco[6]

[1]Duke Kunshan University, China
[2]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
[3]Singapore Institute of Technology, Singapore
[4]Institute for Natural Language Processing, University of Stuttgart, Germany
[5]National Institute of Informatics, Tokyo, Japan
[6]Audio Security and Privacy Group, EURECOM, France

https://www.voiceprivacychallenge.org/

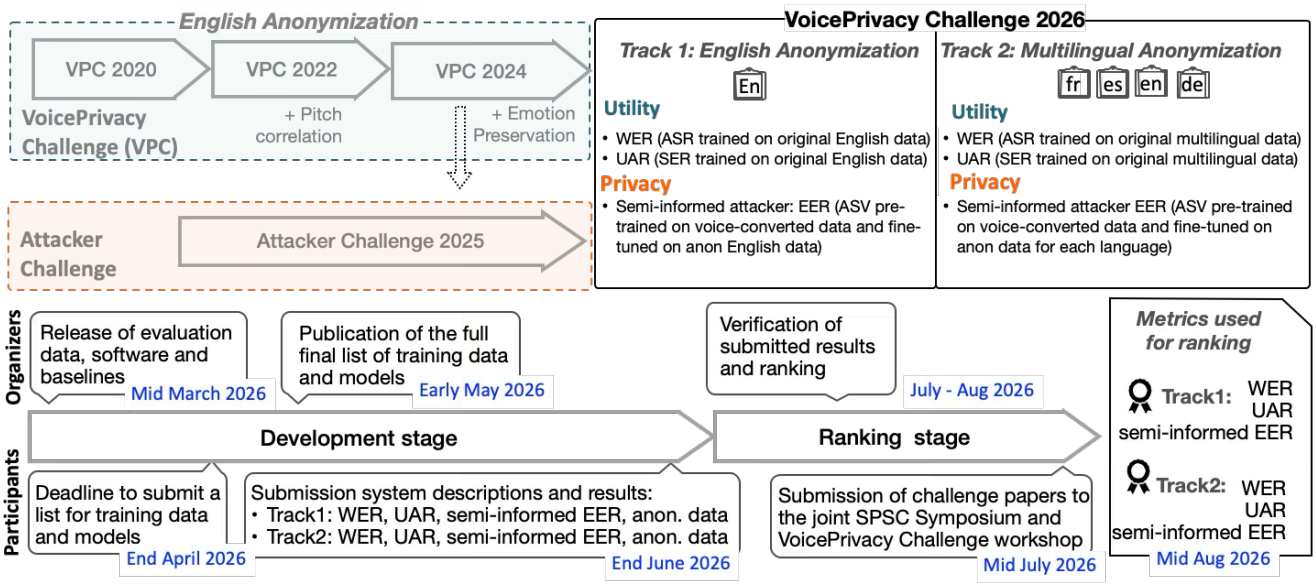


Figure 1: Illustration of past VPC challenges and current edition

For new participants — Executive summary

- This challenge runs in two tracks:
  - Track 1 (English Anonymization): Develop a voice anonymization system for English speech.
  - Track 2 (Multilingual Anonymization): Develop a voice anonymization system for speech in multiple languages: French, Spanish, English, and German.

  Objective for both tracks: Conceal speaker identity while preserving linguistic content and emotional states.
- The organizers provide development and evaluation datasets and evaluation scripts, as well as baseline anonymization systems and a list of training resources. Participants choose to complete either one or both of the two tracks, apply their developed anonymization systems, run evaluation scripts, and submit evaluation results and anonymized speech data to the organizers.
- Results will be presented at a workshop held in conjunction with Interspeech 2026, to which all participants are invited to present their challenge systems and to submit additional workshop papers.

For readers familiar with the VoicePrivacy Challenge — Changes w.r.t. 2024

- Track 1 follows the previous challenge setting, but in addition to same-gender trials, cross-gender trials are also considered, and a stronger attacker ASV model, pretrained on a massive amount of voice-converted data and finetuned on anonymized data, is used for privacy evaluation.
- Track 2 requires a single anonymization system that anonymizes speech data in multiple languages. The privacy metric adopts the equal error rate (EER) calculated by an attacker ASV model pretrained on a massive amount of voice-converted data and finetuned on language-specific anonymized data; the utility metrics are the word error rate (WER) for automatic speech recognition (ASR) and the unweighted average recall (UAR) for speech emotion recognition (SER). All metrics are averaged across languages.

Note: Both tracks share the same pretrained attacker ASV model but use track-specific anonymized data for finetuning.

1 Challenge objectives

Speech data fall within the scope of major privacy regulations, such as the European General Data Protection Regulation (GDPR). Indeed, speech signals encapsulate a wealth of personal (i.e., personally identifiable) information, including the speaker's identity, age, gender, health status, personality, racial or ethnic origin, geographical background, social identity, and socio-economic status [1].

Formed in 2020, the VoicePrivacy initiative [2] has spearheaded efforts to develop privacy-preserving solutions for speech technologies. To date, it has primarily focused on *voice anonymization*, i.e., transforming speech signals to conceal speaker identity while preserving speech utility. This objective has been pursued through a series of competitive benchmarking challenges, providing common datasets, standardized evaluation protocols, and meaningful metrics for fair comparison of anonymization systems. The first three editions of the VoicePrivacy Challenge (VPC) were held in 2020, 2022, and 2024 [2–9]. As illustrated in Figure 1, the scope of the VoicePrivacy Challenge has progressively evolved. While VPC 2020 established a foundational evaluation framework for English voice anonymization, VPC 2022 extended this framework to assess prosody preservation, and VPC 2024 further introduced explicit requirements on preserving the speaker's emotional state. Following VPC 2024, the *Attacker Challenge* [10–12] was introduced to foster the development of stronger attacker models, evaluated against a selection of top-performing anonymization systems submitted to VPC 2024, as well as strong baseline systems.

VoicePrivacy 2026, the fourth edition of the challenge, starts in March 2026 and culminates in the VoicePrivacy Challenge workshop held in conjunction with the 6th Symposium on Security and Privacy in Speech Communication (SPSC)[1], co-located with Interspeech 2026[2] in Sydney, Australia. In keeping with prior editions, the challenge focuses on the subtask of *voice anonymization*[3], i.e., altering the speaker's voice to conceal identity as effectively as possible while preserving linguistic content and relevant paralinguistic attributes. In VPC 2026, particular emphasis is placed on two key aspects. First, the challenge introduces *stronger, domain-aware attackers* optimized using domain-related data. Specifically, since most state-of-the-art anonymization approaches are based on neural voice conversion (VC) techniques, attacker models are correspondingly trained on diverse VC data to achieve stronger speaker re-identification performance. Second, beyond English anonymization, VPC 2026 explicitly extends the evaluation to a multilingual setting. The challenge is organised with two independent tracks.

- **Track 1: English anonymization**. This track largely follows the VPC 2024 setup and continues the evaluation of voice anonymization systems in English, with the objective of preserving linguistic content and emotional information while concealing the original speaker identity. Utility is assessed using the word error rate (WER) from an automatic speech recognition (ASR) model and the unweighted average recall (UAR) from a speech emotion recognition (SER) model. The key difference from previous editions lies in the privacy evaluation: 1) in addition to same-gender trials, cross-gender trials are also considered because the attacker does not know whether the original speaker's gender has been preserved or changed, 2) the speaker verification (ASV) system is pre-trained on large-scale voice-converted data and subsequently fine-tuned on anonymized English speech.

- **Track 2: Multilingual anonymization**. This track extends the challenge to a multilingual setting, covering multiple languages: French, Spanish, English, and German. Utility is measured using the word error rate (WER) from a multilingual ASR system and unweighted average recall (UAR) from a multilingual SER system, while privacy is evaluated using speaker verification systems trained on voice-converted data and finetuned on language-specific anonymized data.

This document details the challenge tasks, datasets, pretrained models, and baseline systems provided to participants, as well as the evaluation metrics, rules, and submission guidelines that will be used for the assessment of submitted systems.

2 Task

Privacy protection is formulated as a game between a *user* who shares data for a desired downstream task and an *attacker* who accesses this data or data derived from it and uses it to infer information about the data subjects [2, 13, 14]. Here, we consider the scenario where the user shares anonymized utterances for downstream ASR and SER tasks, and the attacker attempts to identify the speakers from their anonymized utterances.

[1] 6th Symposium on Security and Privacy in Speech Communication: https://spsc-symposium.de/
[2] https://interspeech2026.org/en-AU
[3] For brevity, we henceforth use the term "anonymization" to refer specifically to voice anonymization.

2.1 Voice anonymization task

Common to both tracks, the utterances shared by the user are referred to as *trial* utterances. In order to hide the identity of the speaker within each utterance, the user passes the utterance through a voice anonymization system prior to sharing. The resulting utterance sounds as if it was uttered by another speaker, which we refer to as a *pseudo-speaker*. The pseudo-speaker might, for instance, be an artificial voice not corresponding to any real speaker.

The task of challenge participants is to develop this voice anonymization system for each track. It should:

(a) output a speech waveform;

(b) conceal the speaker identity on the *utterance level*;

(c) not distort the linguistic contents and emotional states.

Additionally, Track 2 has an extra requirement: anonymize input waveforms in languages other than English.

The utterance-level anonymization requirement (b) means that the voice anonymization system must assign a pseudo-speaker to each utterance independently of the other utterances. The pseudo-speaker assignment process (or algorithm) must be identical across all utterances and not rely on speaker labels. When this process involves a random number generator, the random number(s) generated must be different for each utterance, typically resulting in a different pseudo-speaker for each utterance. Voice anonymization systems that assign a single pseudo-speaker to all utterances also satisfy this requirement.

The achievement of requirement (c) is assessed via *utility* metrics. For Track 1, we measure the WER and UAR obtained from ASR and SER systems trained on original (unprocessed) English data. For Track 2, multilingual ASR and multilingual SER systems pre-trained on original (unprocessed) data are used to measure WER and UAR, respectively. Details of the evaluation models are provided in §4 and Table 4.

2.2 Attack model

For each speaker of interest, the attacker is assumed to have access to utterances spoken by that speaker, which are referred to as *enrollment* utterances. He then uses an automatic speaker verification (ASV) system to re-identify the speaker corresponding to each anonymized trial utterance.

In this work, we assume that the attacker has access to:

(a) several enrollment utterances for each speaker;

(b) the voice anonymization system employed by the user;

(c) multiple training utterances that can be anonymized using a voice anonymization system and subsequently used to train a stronger attacker model.

Using this information, the attacker anonymizes the enrollment utterances to reduce the mismatch with the trial utterances, and trains an ASV system on the training lists described in (c), adapted to the anonymization system.

For Track 1, the training list is sampled from English data, whereas for Track 2, the training lists are sampled from a multilingual dataset. Rather than training the ASV system from scratch on anonymized data, the organizers provide a pretrained model trained on large-scale voice-converted speech. The *semi-informed* attacker further fine-tunes this model on anonymized training data, and is used for the final ranking of both tracks. This attacker represents the strongest threat model considered to date, and is therefore regarded as the most reliable reference for privacy assessment. Identity protection is assessed via a *privacy* metric, specifically the EER obtained by the attacker ASV system.

3 Data and pretrained models

Publicly available resources will be used for the training, development and evaluation of voice anonymization systems. The development and evaluation data are fixed, while the choice of training resources is open to the participants.

3.1 Training sources for both tracks

For training anonymization system, we propose to participants to choose the data and models that they wish to use to train their anonymization systems.

Table 1: Statistics of the LibriSpeech development and evaluation sets for ASV and ASR evaluation in Track 1. F–F and M–M denote gender-dependent trials, while Mixed (F–F, M–M, F–M and M–F) denotes gender-independent trials. The EER is calculated using the Mixed trials.

Subset			# Speakers			#Utt.	# ASV trials			
			F	M	Sum		Label	F–F	M–M	Mixed
LibriSpeech	Dev	Enrollment	15	14	29	343	Same-speaker	704	644	1,348
		Trial	20	20	40	1,978	Different-speaker	14,566	12,796	54,094
	Test	Enrollment	16	13	29	438	Same-speaker	548	449	997
		Trial	20	20	40	1,496	Different-speaker	11,196	9,457	40,807

Table 2: Statistics of the MLS development and evaluation sets for ASV and ASR evaluation in Track 2. F–F and M–M denote gender-dependent trials, while Mixed (F–F, M–M, F–M and M–F) denotes gender-independent trials. The EER is calculated using the Mixed trials.

Subset			# Speakers			#Utt.	# ASV trials			
			F	M	Sum		Label	F–F	M–M	Mixed
French (fr)	Dev	Enrollment	9	9	18	371	Same-speaker	1,043	1,002	2,045
		Trial	9	9	18	2,045	Different-speaker	8,344	8,016	34,765
	Test	Enrollment	9	9	18	372	Same-speaker	1,026	1,028	2,054
		Trial	9	9	18	2,054	Different-speaker	8,208	8,224	34,918
Spanish (es)	Dev	Enrollment	10	10	20	368	Same-speaker	1,000	1,040	2,040
		Trial	10	10	20	2,040	Different-speaker	9,000	9,360	38,760
	Test	Enrollment	10	10	20	368	Same-speaker	946	1,071	2,017
		Trial	10	10	20	2,017	Different-speaker	8,514	9,639	38,323
English (en)	Dev	Enrollment	21	21	42	587	Same-speaker	1,588	1,632	3,220
		Trial	21	21	42	3,220	Different-speaker	31,760	32,640	132,020
	Test	Enrollment	21	21	42	582	Same-speaker	1,574	1,613	3,187
		Trial	21	21	42	3,187	Different-speaker	31,480	32,260	130,667
German (de)	Dev	Enrollment	15	15	30	534	Same-speaker	1,451	1,484	2,935
		Trial	15	15	30	2,935	Different-speaker	20,314	20,776	85,115
	Test	Enrollment	15	15	30	525	Same-speaker	1,429	1,440	2,869
		Trial	15	15	30	2,869	Different-speaker	20,006	20,160	83,201

Requirements for training data and models

- All the proposed data and models shared by participants in VPC2024 are allowed to be used, see Appendix A.
- Any additional data and models not included in the above list (together with the corresponding links) should be reported to the organizers: organisers@lists.voiceprivacychallenge.org before **30 April**. All data and models must be publicly accessible. Commercial APIs are discouraged due to their lack of reproducibility.
- The organizers will verify all the submitted requests and make the final list of the corpora and models acceptable for training anonymization system. It will be shared with all the challenge participants and included in this section in the next (v1) version of the updated evaluation plan on **7 May**.

Table 3: Construction and statistics of the emotional development and evaluation sets for SER evaluation in Track 2.

Corpus	Language	Development					Evaluation				
		ang	neu	hap	sad	Total	ang	neu	hap	sad	Total
Oreau[1]	French (fr)	74	79	74	77	304	66	62	60	57	245
MESD[2]	Spanish (es)	71	71	72	72	286	72	72	72	72	288
EMNS[3]	English (en)	66	84	66	74	290	60	65	91	73	289
EmoDB[4]	German (de)	60	43	34	30	167	67	36	37	32	172
Total		**271**	**277**	**246**	**253**	**1047**	**265**	**235**	**260**	**234**	**994**

[1] https://zenodo.org/records/4405783
[2] https://data.mendeley.com/datasets/cy34mh68j9/5
[3] https://www.openslr.org/136
[4] https://www.kaggle.com/datasets/piyushagni5/berlin-database-of-emotional-speech-emodb/data

Table 4: Models and their training data used for objective privacy and utility evaluation in Track 1 and Track 2.

Task	Track 1 (English anonymization)	Track 2 (Multilingual anonymization)
ASV (EER)	**Track 1:** $ASV^{\text{anon}}_{\text{en}}$; **Track 2:** $ASV^{\text{anon}}_{\text{mls-fr}}$, $ASV^{\text{anon}}_{\text{mls-es}}$, $ASV^{\text{anon}}_{\text{mls-en}}$, $ASV^{\text{anon}}_{\text{mls-de}}$: WavLM–ECAPA (WavLM-large weighted features + Fbank) **Training data:** WavLM[1]is pre-trained on 94k hours of speech from LibriLight, VoxPopuli, and GigaSpeech, and further trained on approximately 9k hours of voice-converted data from the Source Speaker Tracing Challenge (SSTC)[2] Finally, the model is fine-tuned on the corresponding anonymized speech.	
ASR (WER)	$ASR_{\textbf{en}}$**:** wav2vec2-based **Training data:** LibriSpeech-train-960	$ASR_{\textbf{mls}}$**:** Whisper-large-v3[3] **Training data:** Trained on 1M hours of weakly labeled audio and 4M hours of pseudo-labeled audio
SER (UAR)	$SER_{\textbf{en}}$**:** wav2vec2-based **Training data:** IEMOCAP	$SER_{\textbf{en}}$**:** emotion2vec-large[4] **Training data:** over 40K hours

[1] https://huggingface.co/microsoft/wavlm-large
[2] https://sstc-challenge.github.io/download
[3] https://huggingface.co/openai/whisper-large-v3
[4] https://huggingface.co/emotion2vec/emotion2vec_plus_large

3.2 Track 1: English anonymization

3.2.1 Development and evaluation data

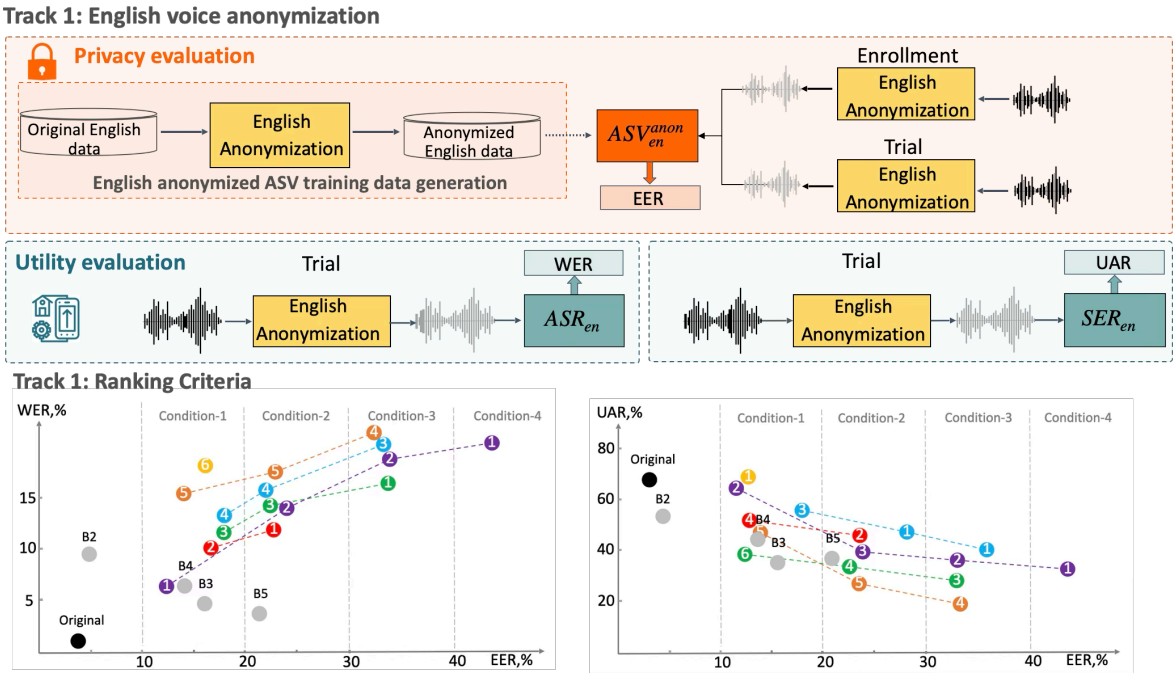
The development and evaluation data used in Track 1 are identical to those employed in VPC 2024, comprising subsets of ***LibriSpeech***[4] [15] and ***IEMOCAP*** [16]. In addition to same-gender trials, we also include cross-gender trials. Together, these are referred to as *mixed*. Table 1 summarizes the corresponding statistics.

3.3 Track 2: Multilingual anonymization

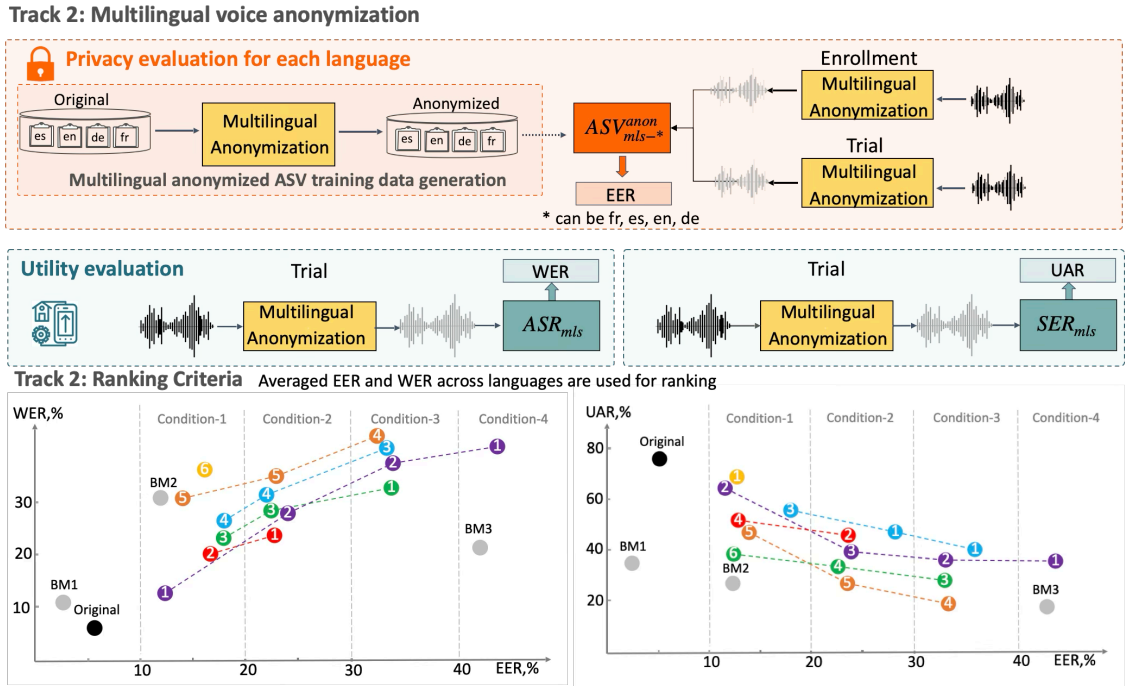
3.3.1 Development and evaluation data

Table 2 summarizes the speaker and utterance statistics for multilingual development and evaluation datasets used in Track 2 for EER and WER calculation. The datasets are derived from *Multilingual LibriSpeech (MLS)* [17], a large multilingual corpus of read speech derived from LibriVox audiobooks. For this challenge, we select four high-resource languages from the MLS development and test sets: French (18 speakers), Spanish (20 speakers), English (42 speakers), and German (30 speakers), where each language contains at least 9 female and 9 male speakers. The original development and test splits provided by MLS are preserved. Table 3 shows emotion distribution in Track 2 development and evaluation sets. The data is sampled from four corpora: Oreau (French) [18], MESD (Spanish) [19], EMNS (English) [20], and EmoDB (German) [21]. Each subset contains four emotion categories: angry (ang), neutral (neu), happy (hap), and sad (sad). The development set consists of 1,047 utterances, while the evaluation set contains 994 utterances, with a relatively balanced distribution of emotions across languages and splits.

[4]LibriSpeech: http://www.openslr.org/12



(a) Track 1



(b) Track 2

Figure 2: Illustration of evaluation and ranking schemes.

4 Privacy and utility evaluation

Figures 2a and 2b illustrate the evaluation pipelines and ranking criteria for Track 1 (English anonymization) and Track 2 (multilingual anonymization), respectively. In both tracks, privacy and utility are evaluated using objective metrics derived from ASV, ASR and SER systems.

4.1 Track 1: English anonymization

Track 1 follows the evaluation protocol of the previous challenge edition. Privacy evaluation is conducted using anonymized enrollment and trial utterances, while utility evaluation for both ASR and SER is performed only on anonymized trial utterances, as illustrated in Figure 2a. The evaluation models and their details are summarized in the middle column of Table 4.

Attacker ASV model. Privacy is evaluated using an ECAPA-TDNN speaker verification system with WavLM-large weighted features combined with filter-bank (Fbank) features. Instead of concatenating the

two features before feeding them into ECAPA-TDNN [22], the current attacker model performs mid-level feature fusion before the attentive pooling layer. The WavLM-large front end is initialized using a pre-trained checkpoint.[5] It is then jointly updated with the random initialized ECAPA-TDNN using 9k hours of voice-converted data from the Source Speaker Tracing Challenge (SSTC) [23].[6] This serves as the base attacker model. It is further fine-tuned on anonymized *LibriSpeech-train-clean-360* data, i.e. *semi-informed* attacker (denoted $ASV_{\text{en}}^{\text{anon}}$). All reported EER results are computed using this final fine-tuned model. The higher the EER, the greater the privacy.

Utility models. Speech content preservation is evaluated using a wav2vec2-based ASR model trained on *LibriSpeech-train-960* (denoted ASR_{en}). Emotion preservation is evaluated using the unweighted average recall (UAR). A wav2vec2-based SER model trained on *IEMOCAP* is employed (denoted SER_{en}), with performance averaged across five cross-validation folds. The lower the WER and higher the UAR, the greater the utility.

4.2 Track 2: Multilingual anonymization

Track 2 considers the same objective privacy metric and utility metric as Track 1. Privacy is evaluated using the ASV EER, and utility is evaluated using the ASR WER. Both metrics are computed separately for each language and then averaged to obtain the final EER and WER scores [7]. The evaluation models and their details are summarized to the right side of Table 4.

Privacy evaluation follows the ASV protocol described in Table 2. Similar to Track 1, the same pretrained ASV system is used, but fine-tuned on language-specific anonymized training data, resulting in four language-specific evaluation models $ASV_{\text{mls-fr}}^{\text{anon}}$, $ASV_{\text{mls-es}}^{\text{anon}}$, $ASV_{\text{mls-en}}^{\text{anon}}$, $ASV_{\text{mls-de}}^{\text{anon}}$ separately.

For utility evaluation, Track 2 uses the Whisper-large-v3 ASR model [32][8] (denoted ASR_{mls}) and the emotion2vec SER model [33] (denoted SER_{mls}).

4.3 Objective assessment of the privacy–utility tradeoff

As in the 2024 edition, multiple evaluation conditions are defined using a set of minimum target privacy requirements specified by N target EER values: $\{\text{EER}_1, \ldots, \text{EER}_N\}$. Each target EER corresponds to a separate evaluation condition.

Submissions that satisfy a given privacy requirement are ranked according to their utility performance. For both tracks, rankings are produced separately based on WER and UAR. In VoicePrivacy 2026, $N = 4$ evaluation conditions are considered, with minimum target EERs of 10%, 20%, 30%, and 40%.

Lower WER and higher UAR indicate better utility at a given privacy level. Example system rankings under this evaluation framework are illustrated at the bottom of Figures 2a and 2b, respectively. Note that the averaged EER and WER across all languages are used for ranking in Track 2.

5 Baseline voice anonymization systems

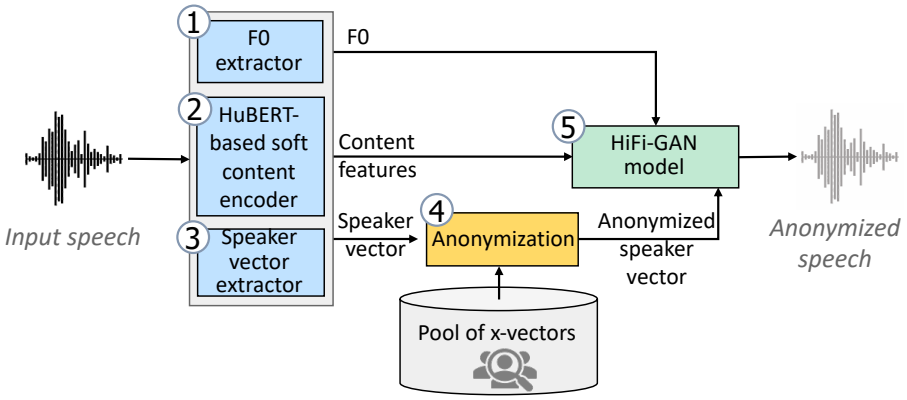


Figure 3: Baseline anonymization system **BM1**.

.

[5] https://huggingface.co/microsoft/wavlm-large

[6] The voice-converted data were created using multiple voice conversion (VC) systems, including AGAIN-VC [24], FreeVC [25], MediumVC [26], StyleTTS [27], TriAAN-VC [28], VQMIVC [29], and KNN-VC [30]. The input speech to the VC is from the LibriSpeech (1,172 speakers) dataset, and the VC target speakers are from VoxCeleb [31].

[7] We intentionally average the EERs across languages without any weight, which treats all the involved languges equally.

[8] During decoding, the ground-truth language label is explicitly provided to the Whisper.

Table 5: Summary of baseline systems for Track 1 and Track 2. **B2**–**B5** are the same as the VPC 2024 baselines, while **BM1-BM3** are newly introduced multilingual anonymization systems used for Track 2.

Track	ID	Prosody extractor	Content encoder	Speaker encoder	Synthesis model	Speaker anon.
1	**B2**	McAdams coefficients-based (DSP-based anonymization)				
	B3	Phone aligner + Praat	E2E ASR	GST	FastSpeech2 + HiFi-GAN	GAN
	B4	HuBERT Base (quantized semantic enc.) + EnCodec				Select
	B5	YAAPT	wav2vec2 + TDNN-F + VQ	ECAPA	HiFi-GAN	Select
2	**BM1**	YAAPT	SSL	ECAPA	HiFi-GAN	Select
	BM2	Phone aligner + Praat	Whisper	GST	IMS Toucan + HiFi-GAN	GAN
	BM3	—				

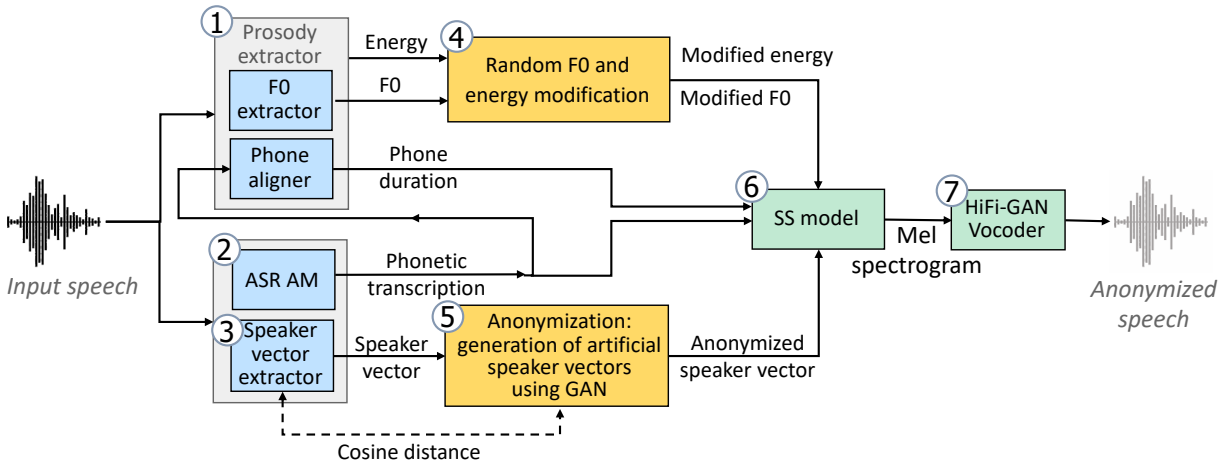


Figure 4: Baseline anonymization systems **BM2** and **BM3**. **BM2** follows the complete pipeline (steps 1–7). **BM3** excludes steps 1 and 4, i.e., no prosody extraction or F0/energy modification is applied.

Table 6: Track1: Privacy (semi-informed EER,%) and Utility (WER,%, UAR,%) on anonymized data vs. original (Orig.). The values highlighted in grey will be used for ranking.

Metric	Development					Evaluation				
	Orig.	**B2**	**B3**	**B4**	**B5**	**Orig.**	**B2**	**B3**	**B4**	**B5**
EER	7.34	6.97	19.14	17.96	24.20	3.91	4.71	16.85	14.33	21.28
WER	1.80	10.44	4.31	6.16	4.91	1.84	9.96	4.31	5.90	4.44
UAR	69.08	55.66	38.08	41.97	40.08	71.06	53.49	35.24	42.78	38.25

5.1 Track 1 English anonymization

The baselines, presented in the upper part of Table 5, are inherited from the 2024 challenge edition [8], but we remove the legacy system **B1**. Table 6 lists the corresponding privacy and utility results on the development and evaluation sets. More specifically, **B2** is a purely signal-processing approach based on McAdams coefficients, **B3** is a TTS-based anonymization system that extracts speaker embeddings, phonetic transcriptions, F0, energy and phone durations, replaces the original embedding with an artificial one generated by a Wasserstein GAN, and then synthesizes anonymized speech with a FastSpeech2 + HiFi-GAN pipeline, while **B4** is a neural audio codec (NAC) language modeling system that uses HuBERT-based semantic tokens and EnCodec acoustic tokens to render the input content with the voice of a pseudo-speaker from a predefined pool, and **B5** is an ASR-BN-based anonymization system that uses a wav2vec 2.0 + TDNN-F acoustic model with a vector-quantized bottleneck (VQ-BN) to extract linguistic features, which are combined with F0 and

Table 7: Track2: Privacy (Semi-informed EER,%) and Utility (WER,% and UAR,%) on anonymized data vs. original (Orig.) for different languages. The values highlighted in grey will be used for ranking.

Language	Metric	Development				Evaluation			
		Orig.	BM1	BM2	BM3	Orig.	BM1	BM2	BM3
fr	EER	6.15	1.86	16.68	46.65	7.51	2.49	16.64	46.53
	WER	6.40	12.71	63.21	24.82	5.67	10.07	44.22	22.47
es	EER	8.97	1.77	12.35	46.12	4.86	1.74	10.02	45.66
	WER	4.44	6.39	24.17	16.89	4.14	7.18	34.12	16.34
en	EER	2.24	1.37	6.58	23.69	4.75	3.67	8.46	22.69
	WER	5.30	8.05	16.48	12.65	6.28	9.34	17.11	13.61
de	EER	3.00	1.47	11.41	47.19	11.39	1.57	12.10	47.26
	WER	5.72	10.00	28.05	30.89	5.88	11.62	32.38	31.34
Avg.	EER	5.09	1.62	11.76	40.91	5.54	2.37	11.81	40.54
	WER	5.68	9.56	30.06	22.33	5.71	10.00	30.99	21.97
MLS	UAR	72.09	41.04	26.87	26.24	78.90	32.40	29.04	24.84

a target speaker one-hot vector and passed to a HiFi-GAN vocoder to generate anonymized speech. These baselines span a range of architectures and design strategies, from lightweight DSP methods to more complex neural pipelines, and are intended to provide representative operating points along the privacy-utility trade-off curve for English speech.

The results of Track 1 baselines are listed in Table 6.

5.2 Track 2 Multilingual anonymization

Track 2 baselines are listed in the lower part of Table 5.

- **BM1** is an HuBERT-based system designed for language independent anonymization [34]. It is similar to the legacy **B1** system from the previous challenge editions but replaces the English-oriented ASR model with the pre-trained HuBERT [35] for extracting the content vectors.
- **BM2** and **BM3** are multilingual extensions of the **B3** in Track 1. They replace the ASR model and the FastSpeech2-based speech synthesis model with the pre-trained Whisper [32] and IMS Toucan multilingual synthesis model [36], respectively. Compared with **BM2**, **BM3** removes the prosody extractor and does not fed the F0 into the synthesis model (See Fig. 4).

The results of Track 2 baselines are listed in Table 7.

6 Challenge Rules

- Participants are free to develop their own anonymization systems, using components of the baselines or not. These systems must operate on the utterance level (§ 2.1) and language labels are permitted at both training and inference time for anonymization.
- Participants are strongly encouraged to make multiple submissions corresponding to different privacy-utility tradeoffs.
- Participants can use only the training and development datasets and models specified in Section 3 (the full list of the allowed training data will be updated on **7 May 2026** in the version v1 of the evaluation plan) in order to train their system and tune hyperparameters. The use of any additional speech data is strictly prohibited.
- Participants must anonymize all the required datasets using the same anonymization system, i.e., the development, evaluation, and training data that will be used to fine-tune the attacker ASV evaluation model. See the bottom of the Table 8 and Table 9.
  - For both tracks, they must fine-tune the attacker ASV evaluation model on the anonymized training data and compute the evaluation metrics (EER, WER, UAR) on the development and evaluation sets using the provided scripts. Modifications to the training or evaluation recipes (e.g.,

changing the ASV model architecture or hyperparameters, retraining the ASR and SER models, etc.) are prohibited.

7 Registration and submission of results

7.1 Registration

Participants/teams are requested to register for the evaluation. Registration should be performed **once only** for each participating entity using the registration form. Participants will receive a confirmation email within ∼24 hours after successful registration, otherwise or in case of any questions they should contact the organizers:

organisers@lists.voiceprivacychallenge.org.

Also, for the updates, all participants and everyone interested the VoicePrivacy Challenge are encouraged to subscribe to the group:

https://groups.google.com/g/voiceprivacy.

Table 8: Required submission files for Track 1.

Category	Item	Path / Description
Result files	Ranking file	exp/results_summary/track1/result_for_rank<suffix>
	Submission archive	exp/results_summary/track1/result_for_submission<suffix>.zip
CSV files	ASR	exp/asr/results*<suffix>.csv
	SER	exp/ser/results*<suffix>.csv
	ASV (lazy-informed)	exp/asv_ssl/results*<suffix>.csv
	ASV (semi-informed)	exp/asv_anon<suffix>/ (all files at maxdepth 1)
Anonymized speech	Dev & Test	LibriSpeech dev & test (en)
	Emotion data	IEMOCAP dev & test
	Training data	train-clean-360

Table 9: Required submission files for Track 2.

Category	Item	Path / Description
Result files	Ranking file	exp/results_summary/track2/result_for_rank<suffix>
	Submission archive	exp/results_summary/track2/result_for_submission<suffix>.zip
CSV files	ASR	exp/openai/whisper-large-v3/results*<suffix>.csv
	SER	exp/ser_emotion2vec/results*<suffix>.csv
	ASV (lazy-informed)	exp/asv_ssl/results*<suffix>.csv
	ASV (semi-informed)	exp/asv_anon_track2*/results*<suffix>.csv (all files at maxdepth 1)
Anonymized speech	Dev & Test	Multilingual dev & test (fr, en, es, de)
	Emotion data	emodata_track2_dev, emodata_track2_test
	Training data	train_english, train_french, train_german, train_spanish
	Additional data	cn, ja†

†cn and ls evaluation data are downloaded and anonymized together with other languages using the default scripts. They will be used for post-evaluation analysis and future VoicePrivacy Attacker Challenge. **and are not used for official ranking**.

Note: All wav files should be 16 kHz, 16-bit signed integer PCM format. These data will be used by the challenge organizers to verify the submitted scores, perform post-evaluation analysis with other metrics and subjective listening tests. All anonymized speech data should be submitted in the form of a single compressed archive.

A summary of the WER and UAR results of Track 1 on the development and evaluation sets is saved in exp/results_summary/track1[9] and Track 2 in exp/results_summary/track2[10].

[9]Example *results* files for the baseline systems in Track 1:

- **B2**: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2026/blob/main/results/track1/result_for_rank_mcadams
- **B3**: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2026/blob/main/results/track1/result_for_rank_sttts
- **B4**: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2026/blob/main/results/track1/result_for_rank_nac
- **B5**: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2026/blob/main/results/track1/result_for_rank_asrbn_hifigan_bn_tdnnf_wav2vec2_vq_48_v1

[10]Example *results* files for the baseline systems in track 2:

Each participant should also submit a single, detailed system description. All submissions should be made according to the schedule below. Submissions received after the deadline will be marked as 'late' submissions, without exception. System descriptions will be made publicly available on the Challenge website. Further details concerning the submission procedure will be published via https://groups.google.com/g/voiceprivacy, by email, or via the VoicePrivacy Challenge website.

8 VoicePrivacy Challenge workshop at Interspeech 2026

The VoicePrivacy 2026 Challenge will culminate in a joint workshop held in Sydney, Australia, in conjunction with **Interspeech 2026** and in cooperation with the ISCA SPSC Symposium.[1] VoicePrivacy 2026 Challenge participants are encouraged to submit papers describing their challenge entry according to the paper submission schedule (see Section 9). Paper submissions must conform to the format of the ISCA SPSC Symposium proceedings, detailed in the author's kit[11], and be 4 to 6 pages long excluding references. Papers must be submitted via the online paper submission system. Submitted papers will undergo peer review via the regular ISCA SPSC Symposium review process, though the review criteria applied to regular papers will be adapted for VoicePrivacy Challenge papers to be more in keeping with systems descriptions and results. Nonetheless, the submission of regular scientific papers related to voice privacy and anonymization are also invited and will be subject to the usual review criteria. The same paper template should be used for system descriptions but may be 2 to 6 pages in length.

Accepted papers will be presented at the joint ISCA SPSC Symposium and VoicePrivacy Challenge Workshop and will be published as other symposium proceedings in the ISCA Archive. Challenge participants without accepted papers are also invited to participate in the workshop and present their challenge contributions reported in system descriptions.

More details will be announced in due course.

9 Schedule

The result submission deadline is **30th June 2026**. The paper submission deadline is to be confirmed and will be updated in the next (v1) version of the evaluation plan. All participants are invited to present their work at the joint SPSC Symposium and VoicePrivacy Challenge workshop that will be organized in conjunction with Interspeech 2026.

Table 10: Important dates

Deadline for participants to submit a list for training data and models	30th April 2026
Publication of the full final list of training data and models	7th May 2026
Deadline for participants to submit objective evaluation results, anonymized data, and system descriptions	30th June 2026
Submission of challenge papers to the joint SPSC Symposium and VoicePrivacy Challenge workshop	TBC
Author notification for challenge papers	TBC
Joint SPSC Symposium and VoicePrivacy Challenge workshop	TBC

10 Acknowledgement

Xin Wang is partially supported by JST, PRESTO Grant Number JPMJPR23P9, Japan. Part of the baseline experiment was conducted using the TSUBAME4.0 supercomputer of Institute of Science Tokyo.

This work was conducted in the context of the Inria–NII TrustedSpeech Associate Team and was partially supported by the French National Research Agency (ANR) under the Speech Privacy project and the IPoP project of the Cybersecurity PEPR. The experiments were partially carried out using the Grid'5000 testbed.

- **BM1**: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2026/blob/main/results/track2/result_for_rank_BM1
- **BM2**: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2026/blob/main/results/track2/result_for_rank_BM2
- **BM3**: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2026/blob/main/results/track2/result_for_rank_BM3

[11] https://interspeech2026.org/en-AU/pages/author-resources/resources

References

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

[2] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "Introducing the VoicePrivacy Initiative," in *Interspeech*, 2020, pp. 1693–1697.

[3] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O'Brien, A. Chanclu, J.-F. Bonastre, M. Todisco, and M. Maouche, "The VoicePrivacy 2020 Challenge: Results and findings," *Computer Speech and Language*, vol. 74, p. 101362, 2022.

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

[5] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "The VoicePrivacy 2020 Challenge evaluation plan," https://www.voiceprivacychallenge.org/vp2020/docs/VoicePrivacy_2020_Eval_Plan_v1_4.pdf, 2020.

[6] N. Tomashenko, X. Wang, X. Miao, H. Nourtel, P. Champion, M. Todisco, E. Vincent, N. Evans, J. Yamagishi, and J.-F. Bonastre, "The VoicePrivacy 2022 Challenge evaluation plan," *arXiv preprint arXiv:2203.12468*, 2022.

[7] N. Tomashenko, X. Wang, X. Miao, H. Nourtel, P. Champion, M. Todisco, E. Vincent, N. Evans, J. Yamagishi, J.-F. Bonastre, and M. Panariello, "The VoicePrivacy 2022 Challenge," 2022. [Online]. Available: https://www.voiceprivacychallenge.org/vp2022/docs/VoicePrivacy_2022_Challenge___Natalia_Tomashenko.pdf

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

[9] N. Tomashenko, X. Miao, P. Champion, S. Meyer, M. Panariello, X. Wang, N. Evans, E. Vincent, J. Yamagishi, and M. Todisco, "The Third VoicePrivacy Challenge: Preserving emotional expressiveness and linguistic content in voice anonymization," *arXiv preprint arXiv:2601.11846*, 2026.

[10] N. Tomashenko, X. Miao, E. Vincent, and J. Yamagishi, "The First VoicePrivacy Attacker Challenge evaluation plan," *arXiv preprint arXiv:2410.07428*, 2024.

[11] ——, "The First VoicePrivacy Attacker Challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–2.

[12] ——, "Privacy attacks on voice anonymization systems: Overview and key findings from the first VoicePrivacy Attacker Challenge," 2026.

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

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

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

[16] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[17] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A large-scale multilingual dataset for speech research," *arXiv preprint arXiv:2012.03411*, 2020.

[18] L. Kerkeni, C. Cleder, Y. Serrestou, and Y. Raood, "French emotional speech database - Oreau," Online, 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4405783

[19] M. M. Duville, L. M. Alonso-Valerdi, and D. I. Ibarra-Zarate, "Mexican emotional speech database (MESD)," Mendeley Data, V5, 2022, doi: 10.17632/cy34mh68j9.5.

[20] N. Kari, Y. Xiaosong, and Z. Jian, "EMNS /Imz/ Corpus: An emotive single-speaker dataset for narrative storytelling in games, television and graphic novels," march 2023.

[21] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Interspeech*, 2005, pp. 1517–1520.

[22] Z. Chen, S. Chen, Y. Wu, Y. Qian, C. Wang, S. Liu, Y. Qian, and M. Zeng, "Large-scale self-supervised speech representation learning for automatic speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6147–6151.

[23] M. Li, P. Zhang, Y. Ren, Z. Cai, and H. Nishizaki, "The SSTC 2024 Challenge evaluation plan," 2024. [Online]. Available: https://sstc-challenge.github.io/file/Evaluation_Plan.pdf

[24] Y.-H. Chen, D.-Y. Wu, T.-H. Wu, and H. yi Lee, "Again-VC: A one-shot voice conversion using activation guidance and adaptive instance normalization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5954–5958.

[25] J. Li, W. Tu, and L. Xiao, "FreeVC: Towards high-quality text-free one-shot voice conversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[26] Y. Gu, Z. Zhang, X. Yi, and X. Zhao, "MediumVC: Any-to-any voice conversion using synthetic specific-speaker speeches as intermedium features," *arXiv preprint arXiv:2110.02500*, 2021.

[27] Y. A. Li, C. Han, and N. Mesgarani, "StyleTTS: A style-based generative model for natural and diverse text-to-speech synthesis," *arXiv preprint arXiv:2205.15439*, 2022.

[28] H. J. Park, S. W. Yang, J. S. Kim, W. Shin, and S. W. Han, "TriAAN-VC: Triple adaptive attention normalization for any-to-any voice conversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[29] D. Wang, L. Deng, Y. T. Yeung, X. Chen, X. Liu, and H. Meng, "VQMIVC: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion," in *Interspeech*, 2021, pp. 1344–1348.

[30] M. Baas, B. van Niekerk, and H. Kamper, "Voice conversion with just nearest neighbors," in *Interspeech*, 2023, pp. 2053–2057.

[31] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Interspeech*, 2018, pp. 1086–1090.

[32] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*, 2023, pp. 28 492–28 518.

[33] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen, "emotion2vec: Self-supervised pre-training for speech emotion representation," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 15 747–15 760.

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

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

[36] F. Lux, S. Meyer, L. Behringer, F. Zalkow, P. Do, M. Coler, E. A. P. Habets, and N. T. Vu, "Meta learning text-to-speech synthesis in over 7000 languages," in *Interspeech*, 2024, pp. 4958–4962.

[37] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

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

[39] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[40] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 745–10 759, 2023.

[41] K. Qian, Y. Zhang, H. Gao, J. Ni, C.-I. Lai, D. Cox, M. Hasegawa-Johnson, and S. Chang, "Contentvec: An improved self-supervised speech representation by disentangling speakers," in *International Conference on Machine Learning*, 2022, pp. 18 003–18 017.

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

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

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

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

[46] J. Kong, J. Kim, and J. Bae, "Hifi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[47] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Transactions on Machine Learning Research*, pp. 1–20, 2023.

[48] K. Zhou, B. Sisman, R. Liu, and H. Li, "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)*, 2021, pp. 920–924.

[49] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[50] S. R. Livingstone and F. A. Russo, "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*, vol. 13, no. 5, p. e0196391, 2018.

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

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

[53] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[54] K. Ito and L. Johnson, "The LJ Speech Dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

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

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

[57] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "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)*, 2018, pp. 2236–2246.

[58] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

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

[60] G. Sharma, A. Dhall, and J. Cai, "Audio-visual automatic group affect analysis," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1056–1069, 2021.

[61] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.

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

[63] G. Maimon and Y. Adi, "Speaking style conversion in the waveform domain using discrete self-supervised units," *arXiv preprint arXiv:2212.09730*, 2022.

Table 11: List of models and data for training anonymization systems in VPC2024

#	Model	Link
1	WavLM Base and Large [37]	https://github.com/microsoft/unilm/tree/master/wavlm
2	Whisper [32]	https://github.com/openai/whisper
3	HuBERT [35]	https://github.com/facebookresearch/fairseq/blob/main/examples/hubert
4	XLS-R [38]	https://github.com/facebookresearch/fairseq/blob/main/examples/wav2vec/xlsr
5	wav2vec 2.0 [39]	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 [40]	https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim
7	ContentVec [41]	https://github.com/auspicious3000/contentvec
8	w2v-BERT [42]	https://github.com/facebookresearch/fairseq/tree/ust/examples/w2vbert
9	ECAPA2 [43]	https://huggingface.co/Jenthe/ECAPA2
10	ECAPA-TDNN [44]	https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb
11	NaturalSpeech 3 [45]	https://huggingface.co/amphion/naturalspeech3_facodec
12	NVIDIA Hifi-GAN Vocoder (en-US) [46]	https://huggingface.co/nvidia/tts_hifigan
13	CRDNN on CommonVoice 14.0 English	https://huggingface.co/speechbrain/asr-crdnn-commonvoice-14-en
14	Encodec [47]	https://huggingface.co/facebook/encodec_24khz
15	Bark	https://huggingface.co/suno/bark https://huggingface.co/erogol/bark/tree/main

#	Dataset	Link
16	ESD [48]	https://hltsingapore.github.io/ESD/download.html
17	LibriSpeech [15]: train-clean-100, train-clean-360, train-other-500	https://www.openslr.org/12
18	CREMA-D [49]	https://github.com/CheyneyComputerScience/CREMA-D
19	RAVDESS [50]	https://datasets.activeloop.ai/docs/ml/datasets/ravdess-dataset/ https://zenodo.org/records/1188976
20	VCTK [51]	https://datashare.ed.ac.uk/handle/10283/2651 https://huggingface.co/datasets/vctk
21	SAVEE [52]	http://kahlan.eps.surrey.ac.uk/savee/ https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee
22	EMO-DB [53]	http://emodb.bilderbar.info/download/
23	LJSpeech [54]	https://keithito.com/LJ-Speech-Dataset/
24	Libri-light [55] (only train part)	https://github.com/facebookresearch/libri-light/blob/main/data_preparation/README.md
25	VoxCeleb-1,2 [31]	https://www.robots.ox.ac.uk/~vgg/data/voxceleb/index.html#about
26	LibriTTS [56]: train-clean-100, train-clean-360, train-other-500	https://openslr.org/60/
27	CMU-MOSEI [57]	http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/
28	MUSAN [58]	https://www.openslr.org/17/
29	RIR [59]	https://www.openslr.org/28/
30	VGAF [60] (from EmotiW challenge)	https://sites.google.com/view/emotiw2023 https://www.kaggle.com/datasets/amirabdrahimov/vgaf-dataset
31	MSP-Podcast [61]	https://ecs.utdallas.edu/research/researchlabs/msp-lab/MSP-Podcast.html

#	**Software with pre-trained models**	**Link**
32	Resemblyzer	https://github.com/resemble-ai/Resemblyzer Model: https://github.com/resemble-ai/Resemblyzer/blob/master/resemblyzer/pretrained.pt
33	VITS [62]	https://github.com/jaywalnut310/vits/ Models: https://drive.google.com/drive/folders/1ksarh-cJf3F5eKJjLVWYOX1j1qsQqiS2
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 [63]	https://github.com/gallilmaimon/DISSC

A Data and model list