The VoicePrivacy 2026 Challenge Evaluation Plan

Version **1.2**

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/

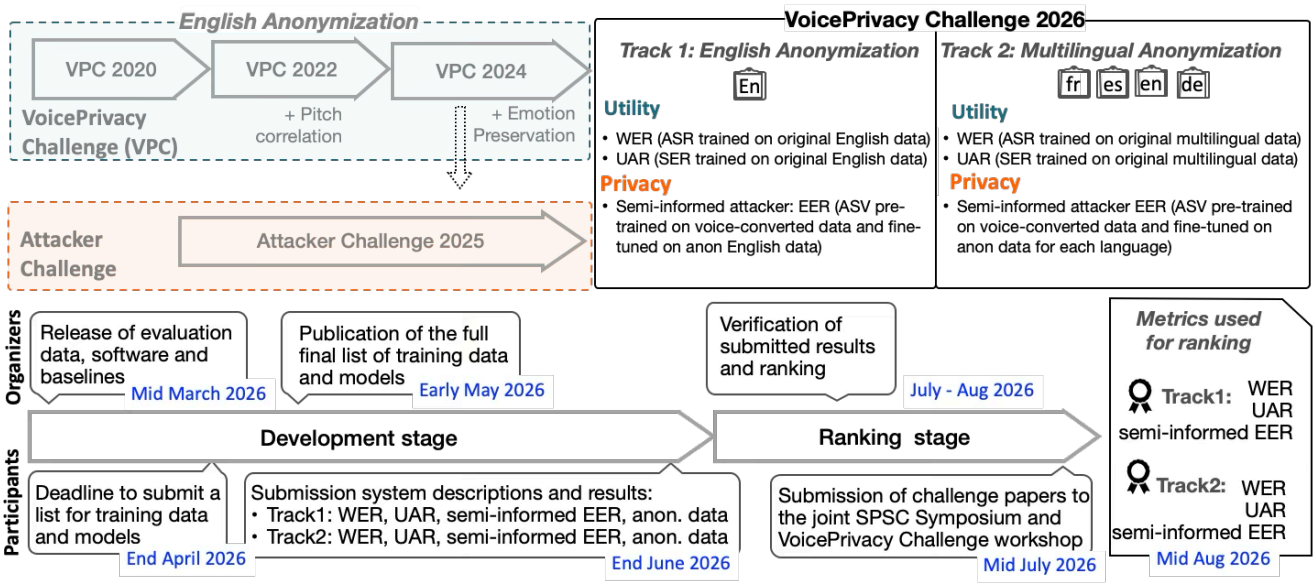


Figure 1: Illustration of past VPC challenges and current edition

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 the SPSC Symposium 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

New in 2026: stronger ASV attacker; cross-gender trials in Track 1; multilingual Track 2 with language-wise ASV attacker finetuning.

- Track 1 follows the previous challenge setting but, in addition to same-gender trials, cross-gender trials are also considered, and a stronger ASV attack model, pretrained on a massive amount of voice-converted data and finetuned on anonymized data, is used for privacy evaluation.
- Track 2 requires anonymization in multiple languages.The privacy metric is the equal error rate (EER) calculated using the ASV attacker model is identical to that for track 1 except for fine-tuning using language-specific anonymized data. All metrics are averaged over the four evaluation languages.

Changes in version 1.2 w.r.t. 1.1

- Applied text normalization and chunked transcription (for utterances longer than 30s) to the Track 2 Whisper WER evaluation and updated the results in Table 8.

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., the transformation of speech signals to conceal voice identity while preserving speech utility. This objective has been pursued through a series of competitive benchmarking challenges, providing common datasets, common evaluation protocols and metrics for the fair comparison of competing anonymization solutions. 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 VPC has progressively evolved. While VPC 2020 established a foundational evaluation framework for English voice anonymization, VPC 2022 extended this framework to assess prosody preservation, while VPC 2024 further introduced explicit requirements to preserve 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 VPC 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 *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, VPC 2026 extends evaluation to a multilingual setting beyond the anonyization of only English-language data. The challenge is organised as two independent tracks.

- **Track 1: English anonymization**. This track largely follows the VPC 2024 setup and continues the evaluation of voice anonymization systems for English-language data, 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) for an automatic speech recognition (ASR) model and the unweighted average recall (UAR) for 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. The attacker is assumed not to 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) for a multilingual ASR system and the unweighted average recall (UAR) for 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] 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 at the *utterance level*;

(c) distort neither the linguistic contents nor emotional states.

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

The utterance-level anonymization requirement (b) means that the voice anonymization system must assign the voice of a pseudo-speaker to each utterance independently of any 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. Specifically, we measure the WER and UAR obtained from ASR and SER systems trained on original (unprocessed) English speech for Track 1 and multilingual speech for Track 2. Details of the evaluation models are provided in §4 and Table 5.

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. The attacker 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-langauge data, whereas for Track 2, the training lists are sampled from a multilingual dataset. Instead of an ASV system trained from scratch on anonymized data, the attacker uses an ASV model pretrained on large-scale voice-converted speech. Under the *semi-informed* attack, the ASV model is further fine-tuned on anonymized training data, and is used in deriving the final submission rankings for both tracks. The semi-informed attack is 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

Table 1: The list of models, datasets, and software allowed for training anonymization systems in VPC2026. Entries 1–15 (models), 76–91 (datasets) and 111–115 (softwares) are inherited from VPC2024; the remaining entries are new in VPC2026.

#	Model	Link
1	WavLM Base and Large [15]	https://github.com/microsoft/unilm/tree/master/wavlm
2	Whisper [16]	https://github.com/openai/whisper
3	HuBERT [17]	https://github.com/facebookresearch/fairseq/blob/main/examples/hubert
4	XLS-R [18]	https://github.com/facebookresearch/fairseq/blob/main/examples/wav2vec/xlsr
5	wav2vec 2.0 [19]	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 [20]	https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim
7	ContentVec [21]	https://github.com/auspicious3000/contentvec
8	w2v-BERT [22]	https://github.com/facebookresearch/fairseq/tree/ust/examples/w2vbert
9	ECAPA2 [23]	https://huggingface.co/Jenthe/ECAPA2
10	ECAPA-TDNN [24]	https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb
11	NaturalSpeech 3 [25]	https://huggingface.co/amphion/naturalspeech3_facodec https://github.com/plachtaa/facodec
12	Hifi-GAN [26]	Official Implementation: https://github.com/jik876/hifi-gan NVIDIA Hifi-GAN Vocoder (en-US): https://huggingface.co/nvidia/tts_hifigan
13	CRDNN on CommonVoice 14.0 English	https://huggingface.co/speechbrain/asr-crdnn-commonvoice-14-en
14	Encodec [27]	https://huggingface.co/facebook/encodec_24khz
15	Bark	https://huggingface.co/suno/bark https://huggingface.co/erogol/bark/tree/main
16	BigVGAN [28]	https://github.com/NVIDIA/BigVGAN
17	Vocos [29]	https://github.com/charactr-platform/vocos https://github.com/gemelo-ai/vocos
18	kNN-VC [30]	https://github.com/bshall/knn-vc
19	FreeVC [31]	https://github.com/OlaWod/FreeVC
20	SpeechT5 [32]	https://github.com/microsoft/SpeechT5
21	seed-vc [33]	https://github.com/Plachtaa/seed-vc
22	OpenVoice [34]	https://github.com/myshell-ai/OpenVoice
23	ReDimNet [35]	https://github.com/IDRnD/redimnet
24	TitaNet-Large [36]	https://huggingface.co/nvidia/speakerverification_en_titanet_large
25	MFA-Conformer SV [37]	https://github.com/zyzisyz/mfa_conformer
26	ResNet-SE-34 (voxceleb_trainer) [38]	https://github.com/clovaai/voxceleb_trainer
27	Canary-1B-Flash	https://huggingface.co/nvidia/canary-1b-flash
28	SeamlessM4T v2 [39]	https://github.com/facebookresearch/seamless_communication
29	emotion2vec [40]	https://github.com/ddlBoJack/emotion2vec
30	IMS Toucan [41]	https://github.com/DigitalPhonetics/IMS-Toucan

31	XTTS-v2 [42]	https://huggingface.co/coqui/XTTS-v2
32	Parler-TTS [43]	https://github.com/huggingface/parler-tts
33	VibeVoice-1.5B	https://huggingface.co/microsoft/VibeVoice-1.5B
34	F5-TTS [44]	https://github.com/SWivid/F5-TTS
35	AudioLDM 2 [45]	https://github.com/haoheliu/AudioLDM2
36	DiffWave [46]	https://github.com/lmnt-com/diffwave
37	Grad-TTS / DiffVC [47, 48]	https://github.com/huawei-noah/Speech-Backbones/tree/main/Grad-TTS https://github.com/huawei-noah/Speech-Backbones/tree/main/DiffVC
38	NaturalSpeech 2 [49]	https://speechresearch.github.io/naturalspeech2
39	Voicebox-style flow matching [50]	https://github.com/lucidrains/voicebox-pytorch
40	DAC (Descript Audio Codec) [51]	https://github.com/descriptinc/descript-audio-codec https://huggingface.co/descript/dac_16khz
41	SNAC [52]	https://github.com/hubertsiuzdak/snac
42	Moshi [53]	https://github.com/kyutai-labs/moshi
43	PESQ [54]	https://github.com/ludlows/PESQ
44	STOI [55]	https://github.com/mpariente/pystoi
45	ViSQOL [56]	https://github.com/google/visqol
46	CREPE [57]	https://github.com/maxrmorrison/torchcrepe
47	FCPE [58]	https://github.com/CNChTu/FCPE
48	AutoVC [59]	https://github.com/auspicious3000/autovc
49	Qwen2-Audio [60]	https://huggingface.co/Qwen/Qwen2-Audio
50	ChatTTS	https://github.com/2noise/ChatTTS
51	StarGANv2-VC [61]	https://github.com/yl4579/StarGANv2-VC
52	CosyVoice 2.0 [62]	https://huggingface.co/FunAudioLLM/CosyVoice2-0.5B
53	StyleTTS 2 [63]	https://github.com/yl4579/StyleTTS2
54	Matcha-TTS [64]	https://github.com/shivammehta25/Matcha-TTS
55	VoxCPM2 [65]	https://github.com/OpenBMB/VoxCPM
56	UUVC [66]	https://github.com/b04901014/UUVC
57	Vevo [67]	https://github.com/open-mmlab/Amphion/blob/main/models/vc/vevo
58	Granite 4.0 1B Speech [68]	https://huggingface.co/ibm-granite/granite-4.0-1b-speech
59	Qwen3-TTS	https://qwen.ai/blog?id=qwen3tts-0115 https://huggingface.co/collections/Qwen/qwen3-tts
60	CapSpeech [69]	https://huggingface.co/OpenSound/CapSpeech-models
61	Fish-Audio [70]	https://github.com/fishaudio/fish-speech https://huggingface.co/fishaudio/s2-pro
62	MeanVC [71]	https://github.com/ASLP-lab/MeanVC
63	WeNet [72]	https://github.com/wenet-e2e/wenet
64	Fast-U2++ [73]	https://github.com/cdliang11/fast-u2pp
65	SyllableLM	https://github.com/AlanBaade/SyllableLM
66	JHCodec [74]	https://huggingface.co/jhcodec/jhcodec
67	SW2V	https://huggingface.co/jhcodec/sw2v_120k

68	StreamVoice (unofficial impl.)	https://github.com/hrnoh24/stream-vc
69	StreamVoiceAnon	https://github.com/Plachtaa/StreamVoiceAnon
70	WavLM–ECAPA (SSTC)[4]	https://duke.app.box.com/shared/static/na6grb7akap4ze66stiazp2azw4zb1f1
71	w2v-BERT 2.0 [75]	https://huggingface.co/facebook/w2v-bert-2.0
72	wav2vec2-large-960h [19]	https://huggingface.co/facebook/wav2vec2-large-960h
73	Whisper-large-v3 [16]	https://huggingface.co/openai/whisper-large-v3
74	WavLM-ECAPA-joint [76]	https://huggingface.co/espnet/voxcelebs12_ecapa_wavlm_joint
75	mHuBERT-147 [77]	https://huggingface.co/utter-project/mHuBERT-147

#	**Dataset**	**Link**
76	ESD [78]	https://hltsingapore.github.io/ESD/download.html
77	LibriSpeech [79]: train-clean-100, train-clean-360, train-other-500	https://www.openslr.org/12
78	CREMA-D [80]	https://github.com/CheyneyComputerScience/CREMA-D
79	RAVDESS [81]	https://datasets.activeloop.ai/docs/ml/datasets/ravdess-dataset/ https://zenodo.org/records/1188976
80	VCTK [82]	https://datashare.ed.ac.uk/handle/10283/3443
81	SAVEE [83]	http://kahlan.eps.surrey.ac.uk/savee/ https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee
82	EMO-DB [84]	http://emodb.bilderbar.info/download/
83	LJSpeech [85]	https://keithito.com/LJ-Speech-Dataset/
84	Libri-light [86] (only train part)	https://github.com/facebookresearch/libri-light/blob/main/data_preparation/README.md
85	VoxCeleb-1,2 [87]	https://www.robots.ox.ac.uk/~vgg/data/voxceleb/index.html#about
86	LibriTTS [88]: train-clean-100, train-clean-360, train-other-500	https://openslr.org/60/
87	CMU-MOSEI [89]	http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/
88	MUSAN [90]	https://www.openslr.org/17/
89	RIR [91]	https://www.openslr.org/28/
90	VGAF [92] (from EmotiW challenge)	https://sites.google.com/view/emotiw2023 https://www.kaggle.com/datasets/amirabdrahimov/vgaf-dataset
91	MSP-Podcast [93]	https://ecs.utdallas.edu/research/researchlabs/msp-lab/MSP-Podcast.html
92	LibriTTS-R [94]	https://www.openslr.org/141/
93	LibriHeavy [95]	https://huggingface.co/datasets/pkufool/libriheavy https://github.com/k2-fsa/libriheavy
94	MLS: Multilingual LibriSpeech (only train part) [96]	https://www.openslr.org/94

[4]Same model as the lazy-informed ASV attacker; permitted for anonymization development but risks overfitting to this attacker.

95	VoxPopuli [97]	https://github.com/facebookresearch/voxpopuli
96	GigaSpeech [98]	https://github.com/SpeechColab/GigaSpeech
97	Hi-Fi TTS [99]	https://www.openslr.org/109/
98	CaFE [100]	https://zenodo.org/record/1478765
99	PAVOQUE [101]	https://github.com/marytts/pavoque-data
100	EmoV-DB [102]	https://github.com/numediart/EmoV-DB
101	MEAD [103]	https://github.com/uniBruce/Mead
102	TESS	https://doi.org/10.5683/SP2/E8H2MF
103	JL-Corpus [104]	https://github.com/tli725/JL-Corpus
104	EmoBox (exclude IEMOCAP) [105]	https://github.com/emo-box/EmoBox
105	Emilia [106]	https://huggingface.co/datasets/amphion/Emilia-Dataset
106	WenetSpeech [107]	https://github.com/wenet-e2e/WenetSpeech
107	Common Voice (English v23.0) [108]	https://datacollective.mozillafoundation.org/organization/cmfh0j9o10006ns07jq45h7xk https://commonvoice.mozilla.org/
108	AISHELL-3 [109]	https://www.openslr.org/93
109	TidyVoice [110]	https://mozilladatacollective.com/datasets/cmihtsewu023so207xot1iqqw
110	CAMEO [111]	https://huggingface.co/datasets/amu-cai/CAMEO

#	**Software with pretrained models**	**Link**
111	Resemblyzer	https://github.com/resemble-ai/Resemblyzer Model: https://github.com/resemble-ai/Resemblyzer/blob/master/resemblyzer/pretrained.pt
112	VITS [112]	https://github.com/jaywalnut310/vits/ Models: https://drive.google.com/drive/folders/1ksarh-cJf3F5eKJjLVWYOX1j1qsQqiS2
113	PIPER pretrained on VITS	https://github.com/rhasspy/piper/?tab=readme-ov-file Models: https://huggingface.co/datasets/rhasspy/piper-checkpoints/tree/main
114	RVC-Project	https://github.com/RVC-Project Models: https://huggingface.co/lj1995/VoiceConversionWebUI/tree/main
115	DISSC [113]	https://github.com/gallilmaimon/DISSC
116	3D-Speaker [114]	https://github.com/modelscope/3D-Speaker
117	HiFTNet [115]	https://github.com/yl4579/HiFTNet
118	Masked prosody model [116]	https://huggingface.co/cdminix/masked_prosody_model

In addition to the training data used in previous challenge editions and the baseline anonymization systems, participants were allowed to propose additional datasets and pretrained models for anonymization system development before the deadline (30 April). Based on participant submissions, this version of the evaluation plan publishes the final list of permitted training data and pretrained models for anonymization systems, shown in Table 1.

Unless otherwise stated, all model versions available on the corresponding webpages before 7th May 2026 can be used for anonymization system development and training. If any software relies on pretrained models, those models must be explicitly listed in the table.

Table 2: 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 3: 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

3.2 Development and evaluation data

Track 1: The development and evaluation data used in Track 1 are identical to those employed in VPC 2024, comprising subsets of *LibriSpeech* [79] and *IEMOCAP* [117]. In addition to same-gender trials, we also include cross-gender trials. Together, these are referred to as *mixed*. Table 2 summarizes the corresponding statistics.

Track 2: Table 3 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)* [96], 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 4 shows emotion distribution in Track 2 development and evaluation sets (denoted as *EmoTrack2*). The data is sampled from four corpora: *Oreau* (French) [118], *MESD* (Spanish) [119], *EMNS* (English) [120], and *EmoDB* (German) [121]. 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.

Table 4: Construction and statistics of EmoTrack2 development and evaluation set 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 5: Models and training data used for objective privacy and utility evaluation in both tracks.

Task	Track 1 (English anonymization)	Track 2 (Multilingual anonymization)
ASV (EER)	WavLM–ECAPA (WavLM-large weighted features + Fbank). WavLM[1] is pre-trained on 94k h of speech from *LibriLight*, *VoxPopuli*, and *GigaSpeech*, and further trained on ∼9k h of voice-converted data from the Source Speaker Tracing Challenge (SSTC)[2]. It is then fine-tuned on language-specific anonymized speech.	
	ASV^{anon}_{en} finetuned on anonymized *LibriSpeech-train-clean-360*	$ASV^{anon}_{mls\text{-}en}$, $ASV^{anon}_{mls\text{-}fr}$, $ASV^{anon}_{mls\text{-}de}$, $ASV^{anon}_{mls\text{-}es}$ finetuned on per-language anonymized *MLS* [96] subsets[5] **Lang.** **Spk** **Utt/Spk** **Total** mls-en 3,930 5 18,986 mls-fr 142 50 4,658 mls-de 176 50 6,915 mls-es 86 50 3,587
ASR (WER)	ASR_{en}: wav2vec2-based **Training data:** *LibriSpeech-train-960*	ASR_{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_{en}: wav2vec2-based **Training data:** *IEMOCAP*	SER_{mls}: emotion2vec-large[4] **Training data:** over 40K hours

[1] https://huggingface.co/microsoft/wavlm-large
[2] https://sstc-challenge.github.io
[3] https://huggingface.co/openai/whisper-large-v3
[4] https://huggingface.co/emotion2vec/emotion2vec_plus_large
[5] The speakers in the MLS English subset were selected based on the list used in the ASVspoof 5 dataset [122], which maintains a suitable balance between the influence of speakers for which data is abundant and those for which data is sparse. We then randomly selected 50 utterances from each speaker. For the remaining languages, we kept all speakers and randomly selected 5 utterances per speaker.

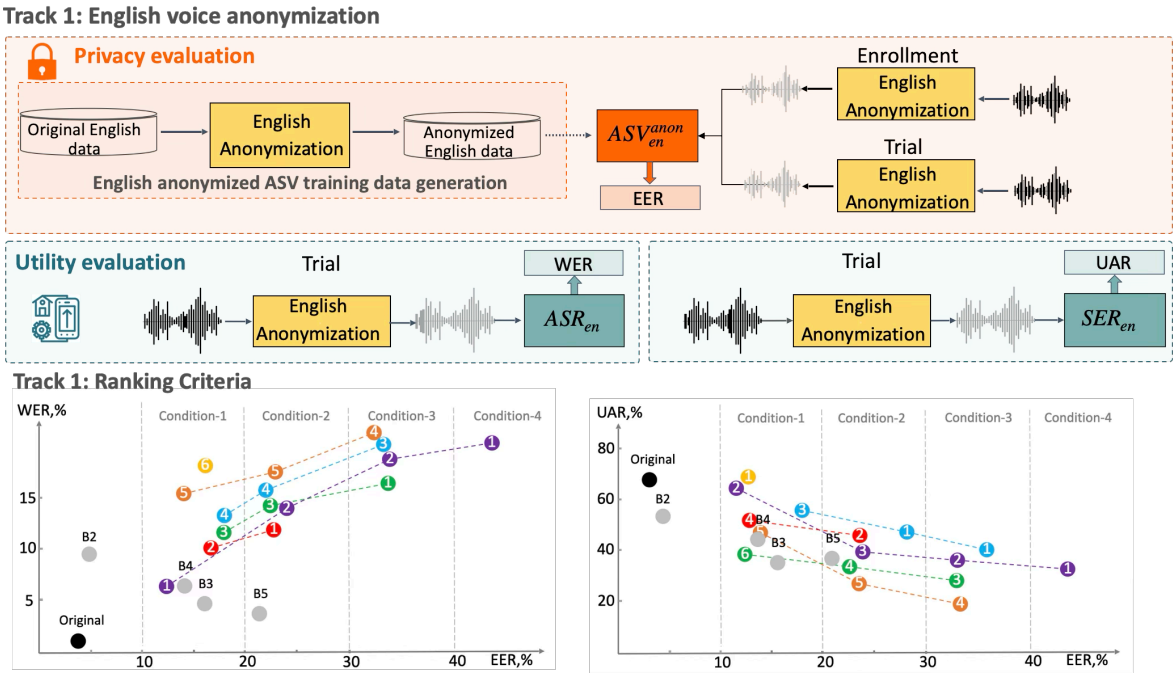
4 Privacy and utility evaluation

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.

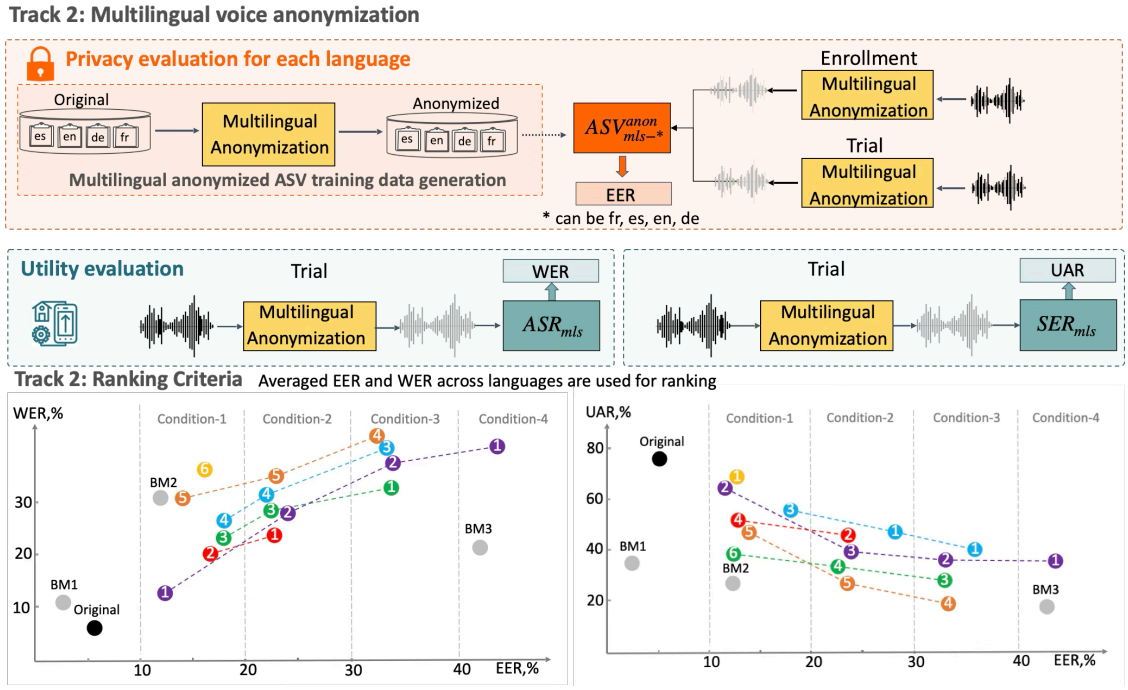
4.1 Track 1: English anonymization

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 5.

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 [123], 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. It is then jointly updated with the random initialized ECAPA-TDNN using 9k hours



(a) Track 1



(b) Track 2

Figure 2: Illustration of evaluation and ranking schemes.

of voice-converted data from the Source Speaker Tracing Challenge (SSTC) [124][5]. This serves as the base attacker model (training details in [132]). 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.

[5]The voice-converted data were created using multiple voice conversion (VC) systems, including AGAIN-VC [125], FreeVC [126], MediumVC [127], StyleTTS [128], TriAAN-VC [129], VQMIVC [130], and KNN-VC [131]. The input speech to the VC is from the LibriSpeech (1,172 speakers) dataset, and the VC target speakers are from VoxCeleb [87].

Table 6: 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	one-hot vector	HiFi-GAN	Select
2	**BM1**	YAAPT	SSL	ECAPA	HiFi-GAN	Select
	BM2	Phone aligner + Praat	Whisper	ECAPA	IMS Toucan + HiFi-GAN	GAN
	BM3	—				

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. The evaluation models and their details are summarized to the right side of Table 5.

Both metrics are computed separately for each language and then averaged to obtain the final EER and WER scores[6].

Privacy evaluation follows the ASV protocol described in Table 3. 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 $ASV^{\text{anon}}_{\text{mls-fr}}$, $ASV^{\text{anon}}_{\text{mls-es}}$, $ASV^{\text{anon}}_{\text{mls-en}}$, $ASV^{\text{anon}}_{\text{mls-de}}$ separately.

For utility evaluation, Track 2 uses the Whisper-large-v3 ASR model [16][7] (denoted ASR_{mls}) and the emotion2vec SER model [40] (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

5.1 Track 1 English anonymization

The baselines, presented in the upper part of Table 6, are inherited from the 2024 challenge edition [8], but we remove the legacy system **B1** and excluded **B6**, as it is similar to **B5** but not better. Table 7 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, randomly modifies F0 and energy per phone, 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 a target speaker one-hot vector and passed to a HiFi-GAN vocoder to generate anonymized speech. These

[6] We intentionally average the EERs across languages without any weight, which treats all the involved languages equally.
[7] During decoding, the ground-truth language label is explicitly provided to the Whisper.

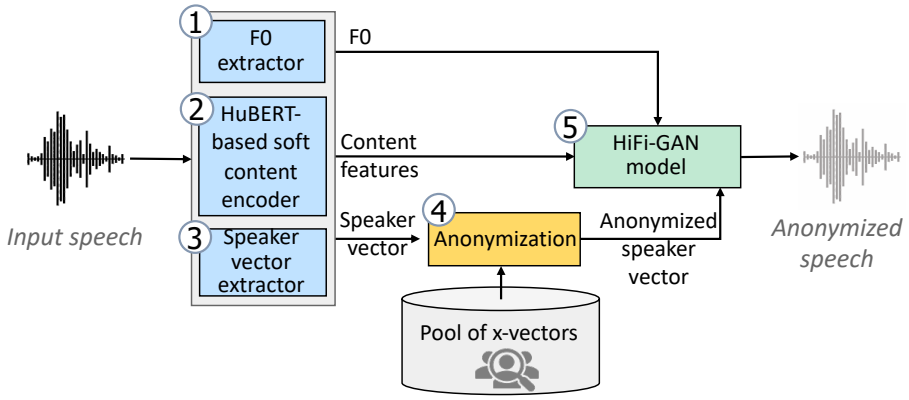


Figure 3: Baseline anonymization system **BM1**.

.

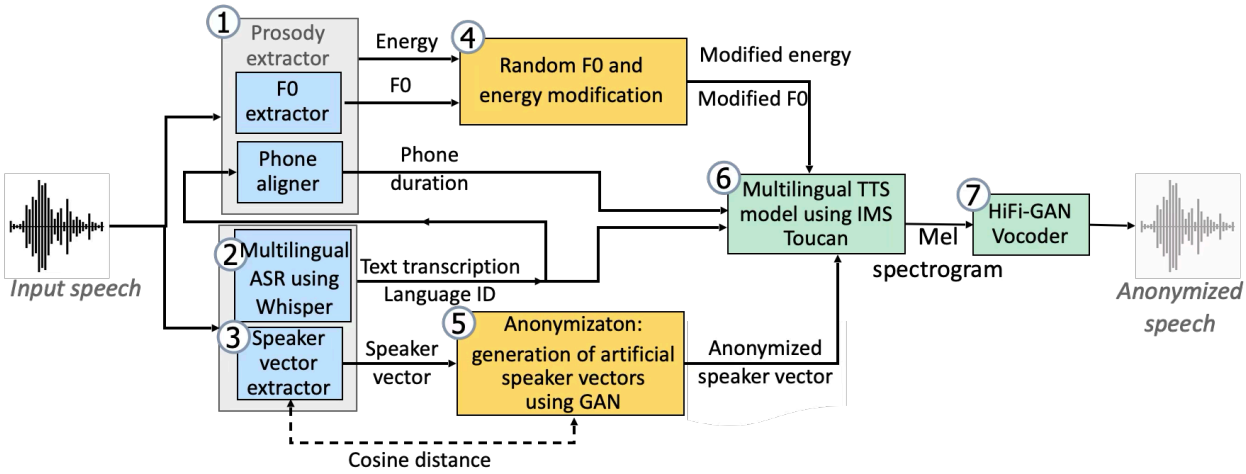


Figure 4: Baseline anonymization systems **BM2** and **BM3**. **BM2** follows the complete pipeline (steps 1–7). **BM3** excludes steps 1 and 4, i.e., no prosody extraction or F0/energy modification is applied.

Table 7: Track1: Privacy (semi-informed EER,%) and Utility (WER,% and 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

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 7.

5.2 Track 2 Multilingual anonymization

Track 2 baselines are listed in the lower part of Table 6.

- **BM1** is an HuBERT-based system designed for language independent anonymization [133]. 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 [17] for extracting the content vectors (see Fig. 3).
- **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 [16] and IMS Toucan multilingual synthesis model [134], respectively, and are similar to the anonymization system proposed

Table 8: 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
MLS-fr	EER	6.15	2.10	16.93	47.09	7.51	2.63	16.11	46.54
	WER	5.80	12.42	52.17	13.77	5.13	9.93	42.24	11.99
MLS-es	EER	8.97	1.63	11.91	46.23	4.86	1.83	10.23	46.46
	WER	3.79	6.15	20.37	10.56	3.75	6.21	28.92	10.12
MLS-en	EER	2.24	1.40	6.34	23.49	4.75	3.73	8.82	22.25
	WER	4.57	7.40	14.29	6.71	5.46	8.54	15.41	7.73
MLS-de	EER	3.00	1.43	11.38	46.84	11.39	1.75	11.88	47.02
	WER	5.25	7.81	19.28	9.74	5.38	9.38	24.40	10.10
MLS-Avg.	EER	5.09	1.64	11.64	40.91	7.13	2.49	11.76	40.57
	WER	4.85	8.45	26.52	10.20	4.93	8.52	27.74	9.98
EmoTrack2	UAR	72.09	41.30	27.13	26.24	78.90	42.63	29.04	24.84

in [135]. Instead of GST-based embeddings, ECAPA-TDNN is used to encode the speaker. Compared with **BM2**, **BM3** removes the prosody extractor and does not feed the original F0, energy and phone durations into the synthesis model, instead they are estimated based on the transcription (see Fig. 4).

The results of Track 2 baselines are listed in Table 8.

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 the models and data listed in Table 1. The use of any other data or models not included in this table 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 9 and Table 10.
  - 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 in the VoicePrivacy Challenge are encouraged to subscribe to the group:

Table 9: 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

*Semi-informed EER is used for official ranking.
‡Lazy-informed EER is collected for post-evaluation analysis only and is not used for ranking.

Table 10: 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†

*Semi-informed EER is used for official ranking.
‡Lazy-informed EER is collected for post-evaluation analysis only and is not used for ranking.
†Mandarin (cn) data sampled from *AISHELL-3* [136] and Japanese (ja) data sampled from [137] 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. **They are not used for official ranking**.

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

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[8] and Track 2 in exp/results_summary/track2[9].

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]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

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

- **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

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[10], 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.

In addition to workshop paper submissions, the organizers plan to propose a special journal issue on voice privacy. If approved, participants will be invited to submit extended versions of their work. Confirmation will be announced in a later update.

9 Schedule

The results and paper submission deadline is **30th June 2026**. 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 11: 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	30th June 2026
Author notification for challenge papers	18th July 2026
Joint SPSC Symposium and VoicePrivacy Challenge workshop	26th September 2026

10 Acknowledgement

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. Some experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see `https://www.grid5000.fr`). Parts of this work were also funded by the German Research Foundation (DFG), Project: Multilingual Controllable Voice Privacy (VoiPy), Project number 533241795. 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.

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.

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

[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," *Computer Speech & Language*, vol. 100, p. 101988, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230826000513

[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," *https://hal.science/hal-05543730*, 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] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

[28] S. gil Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "BigVGAN: A universal neural vocoder with large-scale training," in *ICLR*, 2023.

[29] H. Siuzdak, "Vocos: Closing the gap between time-domain and Fourier-Based neural vocoders for high-quality audio synthesis," in *ICLR*, 2024.

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

[31] J. Li, W. Tu, and L. Xiao, "FreeVC: Towards high-quality text-free one-shot voice conversion," in *ICASSP*, 2023.

[32] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang, Z. Wei, Y. Qian, J. Li, and F. Wei, "SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing," in *ACL*, 2022.

[33] S. Liu, "Zero-shot voice conversion with diffusion transformers," *arXiv preprint arXiv:2411.09943*, 2024.

[34] Z. Qin, W. Zhao, X. Yu, and X. Sun, "OpenVoice: Versatile instant voice cloning," *arXiv preprint arXiv:2312.01479*, 2024.

[35] I. Yakovlev, R. Makarov, A. Balykin, P. Malov, A. Okhotnikov, and N. Torgashov, "Reshape dimensions network for speaker recognition," in *Interspeech*, 2024, pp. 3235–3239.

[36] N. R. Koluguri, T. Park, and B. Ginsburg, "Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context," in *ICASSP*. IEEE, 2022, pp. 8102–8106.

[37] Y. Zhang, Z. Lv, H. Wu, S. Zhang, P. Hu, Z. Wu, H. yi Lee, and H. Meng, "MFA-Conformer: Multi-scale Feature Aggregation Conformer for Automatic Speaker Verification," in *Interspeech*, 2022, pp. 306–310.

[38] H. S. Heo, B.-J. Lee, J. Huh, and J. S. Chung, "Clova baseline system for the voxceleb speaker recognition challenge 2020," *arXiv preprint arXiv:2009.14153*, 2020.

[39] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elsahar, H. Gong, K. Heffernan, J. Hoffman *et al.*, "Seamlessm4t: Massively multilingual & multimodal machine translation," *arXiv preprint arXiv:2308.11596*, 2023.

[40] 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.

[41] F. Lux, J. Koch, S. Meyer, T. Bott, N. Schauffler, P. Denisov, A. Schweitzer, and N. T. Vu, "The ims toucan system for the blizzard challenge 2023," *arXiv preprint arXiv:2310.17499*, 2023.

[42] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi *et al.*, "Xtts: a massively multilingual zero-shot text-to-speech model," *arXiv preprint arXiv:2406.04904*, 2024.

[43] D. Lyth and S. King, "Natural language guidance of high-fidelity text-to-speech with synthetic annotations," *arXiv preprint arXiv:2402.01912*, 2024.

[44] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. JianZhao, K. Yu, and X. Chen, "F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 6255–6271.

[45] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, "Audioldm 2: Learning holistic audio generation with self-supervised pretraining," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2871–2883, 2024.

[46] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," in *ICLR*, 2021.

[47] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-TTS: A diffusion probabilistic model for text-to-speech," in *ICML*, 2021, pp. 8599–8608.

[48] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, and J. Wei, "Diffusion-based voice conversion with fast maximum likelihood sampling scheme," in *ICLR*, 2022.

[49] K. Shen, Z. Ju, X. Tan, Y. Liu, Y. Leng, L. He, T. Qin, S. Zhao, and J. Bian, "NaturalSpeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers," in *ICLR*, 2024, pp. 698–722.

[50] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar *et al.*, "Voicebox: Text-guided multilingual universal speech generation at scale," in *NeurIPS*, vol. 36, 2023, pp. 14 005–14 034.

[51] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved RVQGAN," in *NeurIPS*, vol. 36, 2024, pp. 27 980–27 993.

[52] H. Siuzdak, F. Grötschla, and L. A. Lanzendörfer, "SNAC: Multi-scale neural audio codec," *arXiv preprint arXiv:2410.14411*, 2024.

[53] A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave, and N. Zeghidour, "Moshi: A speech-text foundation model for real-time dialogue," *arXiv preprint arXiv:2410.00037*, 2024.

[54] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, vol. 2, 2001, pp. 749–752.

[55] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *ICASSP*. IEEE, 2010, pp. 4214–4217.

[56] M. Chinen, F. S. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "Visqol v3: An open source production ready objective speech and audio metric," in *2020 twelfth international conference on quality of multimedia experience (QoMEX)*. IEEE, 2020, pp. 1–6.

[57] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "CREPE: A convolutional representation for pitch estimation," in *ICASSP*, 2018, pp. 161–165.

[58] Y. Luo, R. Zhang, L.-C. Liu, T. Li, and H. Liu, "Fcpe: A fast context-based pitch estimation model," 2025. [Online]. Available: https://arxiv.org/abs/2509.15140

[59] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "AutoVC: Zero-shot voice style transfer with only autoencoder loss," in *ICML*, 2019, pp. 5210–5219.

[60] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin *et al.*, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[61] Y. A. Li, A. Zare, and N. Mesgarani, "StarGANv2-VC: A Diverse, Unsupervised, Non-Parallel Framework for Natural-Sounding Voice Conversion," in *Interspeech*, 2021, pp. 1349–1353.

[62] Z. Du, Y. Wang, Q. Chen, X. Shi, X. Lv, T. Zhao, Z. Gao, Y. Yang, C. Gao, H. Wang *et al.*, "CosyVoice 2: Scalable streaming speech synthesis with large language models," *arXiv preprint arXiv:2412.10117*, 2024.

[63] Y. A. Li, C. Han, V. Raghavan, G. Mischler, and N. Mesgarani, "StyleTTS 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," in *NeurIPS*, vol. 36, 2023, pp. 19 594–19 621.

[64] S. Mehta, R. Tu, J. Beskow, É. Székely, and G. E. Henter, "Matcha-TTS: A fast TTS architecture with conditional flow matching," in *ICASSP*, 2024, pp. 11 341–11 345.

[65] V. Team, "Voxcpm2: Tokenizer-free tts for multilingual speech generation, creative voice design, and true-to-life cloning," *GitHub*, 2026.

[66] L.-W. Chen, S. Watanabe, and A. Rudnicky, "A unified one-shot prosody and speaker conversion system with self-supervised discrete speech units," in *ICASSP*. IEEE, 2023, pp. 1–5.

[67] X. Zhang, X. Zhang, K. Peng, Z. Tang, V. Manohar, Y. Liu, J. Hwang, D. Li, Y. Wang, J. Chan, Y. Huang, Z. Wu, and M. Ma, "Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement," in *ICLR*. OpenReview.net, 2025.

[68] I. G. S. Team, "Granite 4.0 speech," 2026. [Online]. Available: https://huggingface.co/ibm-granite/granite-4.0-1b-speech

[69] H. Wang, J. Hai, D. Chong, K. Thakkar, T. Feng, D. Yang, J. Lee, L. M. Velazquez, J. Villalba, Z. Qin, S. Narayanan, M. Elhiali, and N. Dehak, "Capspeech: Enabling downstream applications in style-captioned text-to-speech," 2025. [Online]. Available: https://arxiv.org/abs/2506.02863

[70] S. Liao, Y. Wang, S. Liu, Y. Cheng, R. Zhang, T. Li, S. Li, Y. Zheng, X. Liu, Q. Wang, Z. Zhou, J. Liu, X. Chen, and D. Han, "Fish audio s2 technical report," 2026. [Online]. Available: https://arxiv.org/abs/2603.08823

[71] G. Ma, J. Yao, Z. Ning, Y. Jiang, L. Xiong, L. Xie, and P. Zhu, "Meanvc: Lightweight and streaming zero-shot voice conversion via mean flows," *arXiv preprint arXiv:2510.08392*, 2025.

[72] B. Zhang, D. Wu, Z. Peng, X. Song, Z. Yao, H. Lv, L. Xie, C. Yang, F. Pan, and J. Niu, "WeNet 2.0: More Productive End-to-End Speech Recognition Toolkit," in *Interspeech 2022*, 2022, pp. 1661–1665.

[73] C. Liang, X.-L. Zhang, B. Zhang, D. Wu, S. Li, X. Song, Z. Peng, and F. Pan, "Fast-u2++: Fast and accurate end-to-end speech recognition in joint ctc/attention frames," *arXiv preprint arXiv:2211.00941*, 2022.

[74] J. Lee, X. He, J. Lee, H. Wang, S. Narayanan, T. Thebaud, L. Moro-Velazquez, J. Villalba, and N. Dehak, "Reconstruct! don't encode: Self-supervised representation reconstruction loss for high-intelligibility and low-latency streaming neural audio codec," *arXiv preprint arXiv:2603.05887*, 2026.

[75] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, M. Duppenthaler, P.-A. Duquenne, B. Ellis, H. Elsahar, J. Haaheim *et al.*, "Seamless: Multilingual expressive and streaming speech translation," *arXiv preprint arXiv:2312.05187*, 2023.

[76] J.-w. Jung, W. Zhang, J. Shi, Z. Aldeneh, T. Higuchi, B.-J. Theobald, A. H. Abdelaziz, and S. Watanabe, "Espnet-spk: full pipeline speaker embedding toolkit with reproducible recipes, self-supervised front-ends, and off-the-shelf models," *arXiv preprint arXiv:2401.17230*, 2024.

[77] M. Z. Boito, V. Iyer, N. Lagos, L. Besacier, and I. Calapodescu, "mHuBERT-147: A Compact Multilingual HuBERT Model," in *Interspeech 2024*, 2024.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

[84] 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.

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

[86] 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.

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

[88] 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.

[89] 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.

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

[91] 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.

[92] 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.

[93] 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.

[94] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, M. Bacchiani, Y. Zhang, W. Han, and A. Bapna, "LibriTTS-R: A Restored Multi-Speaker Text-to-Speech Corpus," in *Interspeech*, 2023, pp. 5496–5500.

[95] W. Kang, X. Yang, Z. Yao, F. Kuang, Y. Yang, L. Guo, L. Lin, and D. Povey, "Libriheavy: A 50,000 hours asr corpus with punctuation casing and context," in *ICASSP*. IEEE, 2024, pp. 10 991–10 995.

[96] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A Large-Scale Multilingual Dataset for Speech Research," in *Interspeech 2020*, 2020, pp. 2757–2761.

[97] C. Wang, M. Rivière, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *ACL*, 2021, pp. 993–1003.

[98] G. Chen, S. Chai, G.-B. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, M. Jin, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Z. You, and Z. Yan, "GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio," in *Interspeech*, 2021, pp. 3670–3674.

[99] E. Bakhturina, V. Lavrukhin, B. Ginsburg, and Y. Zhang, "Hi-Fi multi-speaker English TTS dataset," in *Interspeech*, 2021, pp. 2776–2780.

[100] P. Gournay, O. Lahaie, and R. Lefebvre, "A canadian french emotional speech dataset," in *ACM multimedia systems conference*, 06 2018, pp. 399–402.

[101] I. Steiner, M. Schröder, and A. Klepp, "The PAVOQUE corpus as a resource for analysis and synthesis of expressive speech," in *Phonetik & Phonologie*, 2013, pp. 83–84.

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

[103] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, "MEAD: A large-scale audio-visual dataset for emotional talking-face generation," in *ECCV*, 2020, pp. 700–717.

[104] J. James, L. Tian, and C. Inez Watson, "An Open Source Emotional Speech Corpus for Human Robot Interaction Applications," in *Interspeech*, 2018, pp. 2768–2772.

[105] Z. Ma, M. Chen, H. Zhang, Z. Zheng, W. Chen, X. Li, J. Ye, X. Chen, and T. Hain, "EmoBox: Multilingual Multi-corpus Speech Emotion Recognition Toolkit and Benchmark," in *Interspeech*, 2024, pp. 1580–1584.

[106] H. He, Z. Shang, C. Wang, X. Li, Y. Gu, H. Hua, L. Liu, C. Yang, J. Li, P. Shi, Y. Wang, K. Chen, P. Zhang, and Z. Wu, "Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation," in *SLT*, 2024.

[107] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. Zeng, D. Wu, and Z. Peng, "WenetSpeech: A 10000+ hours multi-domain Mandarin corpus for speech recognition," in *ICASSP*, 2022.

[108] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *LREC*, 2020.

[109] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "AISHELL-3: A multi-speaker Mandarin TTS corpus and the baselines," in *Interspeech*, 2021, pp. 2756–2760.

[110] A. Farhadipour, J. Marquenie, S. Madikeri, T. Vukovic, V. Dellwo, K. Reid, F. M. Tyers, I. Siegert, and E. Chodroff, "Tidyvoice challenge: Cross-lingual speaker verification," in *Interspeech 2026*, 2026.

[111] I. Christop and M. Czajka, "Cameo: Collection of multilingual emotional speech corpora," 2025. [Online]. Available: https://arxiv.org/abs/2505.11051

[112] 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.

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

[114] Y. Chen, S. Zheng, H. Wang, L. Cheng *et al.*, "3d-speaker-toolkit: An open source toolkit for multi-modal speaker verification and diarization," 2025.

[115] Y. A. Li, C. Han, X. Jiang, and N. Mesgarani, "Hiftnet: A fast high-quality neural vocoder with harmonic-plus-noise filter and inverse short time fourier transform," *arXiv preprint arXiv:2309.09493*, 2023.

[116] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[117] 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.

[118] 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

[119] 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.

[120] N. Kari, Y. Xiaosong, and Z. Jian, "EMNS /Imz/ Corpus: An emotive single-speaker dataset for narrative storytelling in games, television and graphic novels," 2023. [Online]. Available: https://www.openslr.org/136/

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

[122] X. Wang, H. Delgado, H. Tak, J.-w. Jung, H.-j. Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T. Kinnunen, N. Evans, K. A. Lee, J. Yamagishi, M. Jeong, G. Zhu, Y. Zang, Y. Zhang, S. Maiti, F. Lux, N. Müller, W. Zhang, C. Sun, S. Hou, S. Lyu, S. Le Maguer, C. Gong, H. Guo, L. Chen, and V. Singh, "ASVspoof 5: Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech," *Computer Speech & Language*, vol. 95, p. 101825, Jan. 2026. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0885230825000506

[123] 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.

[124] 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

[125] Y.-H. Chen, D.-Y. Wu, T.-H. Wu, and H.-Y. 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.

[126] 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.

[127] 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.

[128] 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.

[129] 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.

[130] 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.

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

[132] R. Arefeen, X. Miao, R. Tong, A. B. Ng, S. See, and T. Liu, "Dast: A dual-stream voice anonymization attacker with staged training," *arXiv preprint arXiv:2603.12840*, 2026.

[133] 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.

[134] 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.

[135] S. Meyer, F. Lux, and N. T. Vu, "Probing the Feasibility of Multilingual Speaker Anonymization," in *Interspeech 2024*, 2024, pp. 4448–4452.

[136] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "Aishell-3: A multi-speaker mandarin tts corpus and the baselines," *arXiv preprint arXiv:2010.11567*, 2015.

[137] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "Jvs corpus: free japanese multi-speaker voice corpus," *arXiv preprint arXiv:1908.06248*, 2019.