The First VoicePrivacy Attacker Challenge Evaluation Plan

Version 2.2

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

[1]Inria, France
[2]Singapore Institute of Technology, Singapore
[3]National Institute of Informatics, Tokyo, Japan

https://www.voiceprivacychallenge.org/attacker/
attacker.challenge@inria.fr

Abstract

The First VoicePrivacy Attacker Challenge is a new kind of challenge organized as part of the VoicePrivacy initiative [1] and supported by **ICASSP 2025** as the **SP Grand Challenge**.[1] It focuses on developing **attacker systems against voice anonymization**, which will be evaluated against a set of anonymization systems submitted to the VoicePrivacy 2024 Challenge[2]. Training, development, and evaluation datasets are provided along with a baseline attacker system. Participants shall develop their attacker systems in the form of automatic speaker verification systems and submit their scores on the development and evaluation data to the organizers. To do so, they can use any additional training data and models, provided that they are openly available and declared before the specified deadline. The metric for evaluation is equal error rate (EER). Results will be presented at the ICASSP 2025 special session to which 5 selected top-ranked participants will be invited to submit and present their challenge systems.

Changes in version 2.2 w.r.t. 2.0

Updated list of data and models to train attacker systems (Table 2).

1 Context

Speech encapsulates a wealth of personal, private data, e.g., age and gender, health and emotional state, racial or ethnic origin, geographical background, social identity, and socio-economic status [2]. Formed in 2020, the VoicePrivacy initiative [1] is promoting the development of privacy preservation solutions for speech technology via a series of competitive benchmarking challenges, with common datasets, protocols and metrics. In this context, privacy preservation is classically formulated as a game between *users* who process their utterances (referred to as *trial* utterances) with a privacy preservation system prior to sharing with others, and *attackers* who access these processed utterances or data derived from them and wish to infer information about the users. The level of privacy offered by a given solution is measured as the lowest error rate among all attackers.

The first three VoicePrivacy Challenge editions [3–8] focused on the development of voice anonymization systems. In particular, the systems submitted to the VoicePrivacy 2024 Challenge had to meet the following requirements: (a) output a speech waveform; (b) conceal the speaker identity at the *utterance level*; (c) not distort the linguistic and emotional content. The processed utterances sound as if they were uttered by another speaker, which we refer to as a *pseudo-speaker*. The pseudo-speaker is selected independently for every utterance, and it can be an artificial voice not corresponding to any real speaker. In practice, many voice anonymization systems select the pseudo-speaker or modify prosody in a random or semi-random way using a random number generator. A *semi-informed attack model* [9] was assumed, whereby attackers have access to the voice anonymization system (but not to the random numbers drawn by that system for each utterance, if any), and they seek to re-identify the original speaker behind each anonymized trial utterance. Specifically, an ECAPA-TDNN [10] automatic speaker

[1] https://2025.ieeeicassp.org/sp-grand-challenges/#gc7

[2] The VoicePrivacy Challenge focuses on strengthening voice anonymization systems from the user's perspective, often assuming fixed attack scenarios, which may not fully reflect practical use cases, as real-world attacks can exploit any available clues and resources. In contrast, the Attacker Challenge aims to develop more robust and practical attacker systems capable of challenging various advanced anonymization systems from the attacker's point of view.

verification (ASV) system was provided by the organizers and trained by the participants on data anonymized using their anonymization system. While this attack model is undeniably the most realistic to date, the provided attacker system is not its strongest possible implementation as it does not exploit spoken content similarities, specific pseudo-speaker selection strategies [11], or stronger ASV architectures [12], among others.

To ensure a fair and reliable privacy assessment, it is essential to find the strongest possible attacker against every anonymization system. Hence, the current challenge edition takes the attacker's perspective and focuses on the development of attacker systems against voice anonymization systems.

2 Task

Participants are required to develop one or more attacker systems against one or more voice anonymization systems selected among three VoicePrivacy 2024 Challenge baselines [4] and four systems developed by the VoicePrivacy 2024 Challenge participants. For each speaker of interest, the attacker is assumed to have access to one or more utterances spoken by that speaker, which are referred to as *enrollment* utterances. The attacker system shall output an ASV score for every given pair of trial utterance and enrollment speaker, where higher (resp., lower) scores correspond to same-speaker (resp., different-speaker) pairs.

To develop and evaluate their attacker system against a given voice anonymization system, in line with the assumed semi-informed attack model, participants have access to:

(a) anonymized trial utterances;
(b) original and anonymized enrollment utterances;
(c) original and anonymized training data (as well as other publicly available training resources that will be specified in Section 3) for the ASV system;
(d) a written description of the voice anonymization system;
(e) the software implementation of that voice anonymization system when available.

3 Data

For each voice anonymization system, participants are provided with training, development and evaluation data anonymized using that system. A detailed description of these datasets is presented below and in Table 1.

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

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

Training resources. The training set is the *train-clean-360* subset of the *LibriSpeech* [13] corpus. Besides the provided anonymized training data, participants are allowed to use the original *train-clean-360* data. In addition, participants were allowed to propose other resources such as speech corpora and pretrained models before the deadline (13th October). Based on the suggestions received from the challenge participants, in this version of the evaluation plan, we publish the final list of training data and pretrained models allowed for training attacker systems. All the allowed resources are listed in Table 2.

For some models, the provided link is a webpage listing multiple versions of the model. In this case, unless otherwise stated, all model versions available on that page before 15th October 2024 can be used by participants in training their attacker systems. Participants are allowed to use any existing software in the development and training of their attacker systems. If the software uses pretrained models, these models should be explicitly listed in this table or on the main page (readme) of the corresponding repository before 15th October 2024.

Table 2: Final list of models and data for training attacker systems.

#	Model	Link
1	WavLM Base and Large [14]	https://github.com/microsoft/unilm/tree/master/wavlm https://huggingface.co/microsoft/wavlm-large
2	Whisper [15]	https://github.com/openai/whisper https://huggingface.co/openai/whisper-large

3	HuBERT [16]	https://github.com/facebookresearch/fairseq/blob/main/examples/hubert https://huggingface.co/facebook/hubert-large-ls960-ft
4	XLS-R [17]	https://github.com/facebookresearch/fairseq/blob/main/examples/wav2vec/xlsr https://huggingface.co/facebook/wav2vec2-large-xlsr-53
5	wav2vec 2.0 [18]	https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec https://dl.fbaipublicfiles.com/voxpopuli/models/wav2vec2_large_west_germanic_v2.pt https://huggingface.co/facebook/wav2vec2-large-960h-lv60
6	ECAPA-TDNN [10]	https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb
7	NaturalSpeech 3 [19]	https://huggingface.co/amphion/naturalspeech3_facodec
8	Encodec [20]	https://huggingface.co/facebook/encodec_24khz
9	Bark	https://huggingface.co/suno/bark https://huggingface.co/erogol/bark/tree/main
10	Resnet34	https://wenet.org.cn/downloads?models=wespeaker&version=voxceleb_resnet34.zip
11	Resnet34_lm	https://wenet.org.cn/downloads?models=wespeaker&version=voxceleb_resnet34_LM.zip
12	VGGVox	https://github.com/a-nagrani/VGGVox
13	Kaldi models	http://kaldi-asr.org/models.html
14	Conformer models	https://huggingface.co/models?search=conformer
15	NVIDIA TitaNet-Large (en-US) [21]	https://huggingface.co/nvidia/speakerverification_en_titanet_large

#	**Dataset**	**Link**
16	LibriSpeech [13]: train-clean-100, train-clean-360, train-other-500	https://www.openslr.org/12
17	RAVDESS [22]	https://datasets.activeloop.ai/docs/ml/datasets/ravdess-dataset/ https://zenodo.org/records/1188976
18	SAVEE [23]	http://kahlan.eps.surrey.ac.uk/savee/ https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee
19	EMO-DB [24]	http://emodb.bilderbar.info/download/
20	VoxCeleb-1,2 [25]	https://www.robots.ox.ac.uk/~vgg/data/voxceleb/index.html#about
21	CMU-MOSEI [26]	http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/
22	MUSAN [27]	https://www.openslr.org/17/
23	RIR [28]	https://www.openslr.org/28/
24	CN-Celeb [29]	https://cnceleb.org/
25	Common Voice [30]	https://commonvoice.mozilla.org/en/datasets
26	IEMOCAP [31]	https://sail.usc.edu/iemocap/iemocap_info.htm
27	Emo-DB [24]	http://emodb.bilderbar.info/index-1280.html
28	Toronto Emotional Speech Database	https://tspace.library.utoronto.ca/handle/1807/24487
29	LIRIS-ACCEDE [32]	https://liris-accede.ec-lyon.fr/database.php
30	AESDD [33]	http://m3c.web.auth.gr/research/aesdd-speech-emotion-recognition/
31	ANAD	https://www.kaggle.com/suso172/arabic-natural-audio-dataset
32	BAVED [34]	https://www.kaggle.com/a13x10/basic-arabic-vocal-emotions-dataset
33	VoxForge	http://www.voxforge.org/
34	TED-LIUM 3 [35]	https://www.openslr.org/51/
35	AISHELL-1 [36]	https://www.openslr.org/33/
36	AISHELL-WakeUp-1	https://www.aishelltech.com/wakeup_data
37	CMU-MOSEI [26]	http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/
38	VoxBlink2 [37]	https://voxblink2.github.io/

39	Emotional Voices Database [38]	https://github.com/numediart/EmoV-DB
40	DEMoS [39]	https://zenodo.org/record/2544829
41	Multilingual LibriSpeech (MLS) [40], English	https://www.openslr.org/94/

#	**Software with pre-trained models**	**Link**
42	VITS [41]	https://github.com/jaywalnut310/vits/ Models: https://drive.google.com/drive/folders/1ksarh-cJf3F5eKJjLVWY0X1j1qsQqiS2
43	SASV fusion-based baseline from the SASV 2022 challenge [42]	https://github.com/sasv-challenge/SASVC2022_Baseline
44	Baseline of the ASVspoof 5 challenge Track 2, SASV [43]	https://github.com/sasv-challenge/SASV2_Baseline/tree/asvspoof5
45	ASV-Subtools [44]	https://github.com/Snowdar/asv-subtools
46	WeSpeaker [45]	https://github.com/wenet-e2e/wespeaker Models: https://github.com/wenet-e2e/wespeaker/blob/master/docs/pretrained.md
47	3D-Speaker *ERes2Net* [46] *ERes2NetV2* [47] *CAM++* [48]	https://github.com/modelscope/3D-Speaker
48	SpeechBrain [49]	https://speechbrain.github.io/ Models: https://huggingface.co/speechbrain
49	ResNet [50]	https://github.com/ranchlai/speaker-verification
50	VQMIVC [51]	https://github.com/Wendison/VQMIVC
51	DeepSpeech [52]	https://github.com/mozilla/DeepSpeech Models: https://github.com/mozilla/DeepSpeech/releases
52	AutoVC [53]	https://github.com/auspicious3000/autovc

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

Voice anonymization systems. The voice anonymization systems to be attacked include three baseline systems (**B3**, **B4**, and **B5**) [4] and four selected systems developed by the VoicePrivacy 2024 Challenge participants (**T8-5**, **T10-2**, **T12-5**, and **T25-1**) [54]. The participants' systems were chosen according to their anonymization performance in the highest privacy category (EER $\geq$ 40%), excluding cascaded anonymization systems based on automatic speech recognition (ASR) followed by text-to-speech (TTS). Thus, a total of seven systems are to be attacked:

- **B3** – based on phonetic transcription, pitch and energy modification, and artificial pseudo-speaker embedding generation [4, 55].
- **B4** – based on neural audio codec language modeling [4, 56].
- **B5** – based on vector quantized bottleneck (VQ-BN) features extracted from an ASR model and on original pitch [4, 57].
- **T8-5** (team *JHU-CLSP*, system *"Admixture ($p = 0.4$)"* [58]) – random selection of one of two methods for each utterance (with probability p for the second method): (1) a cascaded ASR-TTS system with *Whisper* [15] for ASR and *VITS* [41] for TTS and (2) a k-nearest neighbor (kNN) voice conversion (VC) system operating on *WavLM* [14] features.
- **T10-2** (team *NPU-NTU*, system *"C4"* [59]) – neural audio codec, with a specific disentanglement strategy for linguistic content, speaker identity and emotional state.
- **T12-5** (team *NTU-NPU*, system *"3"* [60]) – based on **B5**, with additional pitch smoothing.

- **T25-1** (team *USTC-PolyU*, system *"large: ESD+LibriTTS"* [61]) – disentanglement of content (VQ-BN as in **B5**) and style (global style token (GST) [62]) features and emotion transfer from target speaker utterances.

The code of **B3**, **B4**, and **B5** is available on GitHub[3] and can be used to develop attacker systems by, e.g., generating different or additional training data to train those systems.

4 Evaluation metric

We use the equal error rate (EER) metric to evaluate the attacker's performance. This metric has been used in all VoicePrivacy Challenge editions. For every given pair of trial utterance and enrollment speaker, the attacker system outputs an ASV score from which a same-speaker vs. different-speaker decision is made by thresholding. Denoting by $P_{\text{fa}}(\theta)$ and $P_{\text{miss}}(\theta)$ the false alarm and miss rates at threshold θ, the EER metric corresponds to the threshold θ_{EER} at which the two detection error rates are equal, i.e., $\text{EER} = P_{\text{fa}}(\theta_{\text{EER}}) = P_{\text{miss}}(\theta_{\text{EER}})$. The lower this metric, the stronger the attacker. The number of same-speaker and different-speaker trials in the development and evaluation datasets is given in Table 3. The attackers will be ranked separately for each voice anonymization system.

Table 3: Number of speaker verification trials.

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

5 Baseline attacker system

As a baseline, we consider the attacker system used in the VoicePrivacy 2024 Challenge [4] (see Figure 1). The ASV system (denoted $ASV_{\text{eval}}^{\text{anon}}$) is an ECAPA-TDNN [10] with 512 channels in the convolution frame layers, implemented by adapting the *SpeechBrain* [49] *VoxCeleb* recipe to *LibriSpeech*, and it is trained on anonymized training data. For a given trial utterance and enrollment speaker, the attacker computes the average speaker embedding of all anonymized enrollment utterances from that speaker and compares it to the speaker embedding of the anonymized trial utterance. Results for this baseline attacker system are reported in Table 4. The code to train the baseline attacker systems for given anonymized data is available in the GitHub VoicePrivacy 2024 Challenge repository: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2024.[4]

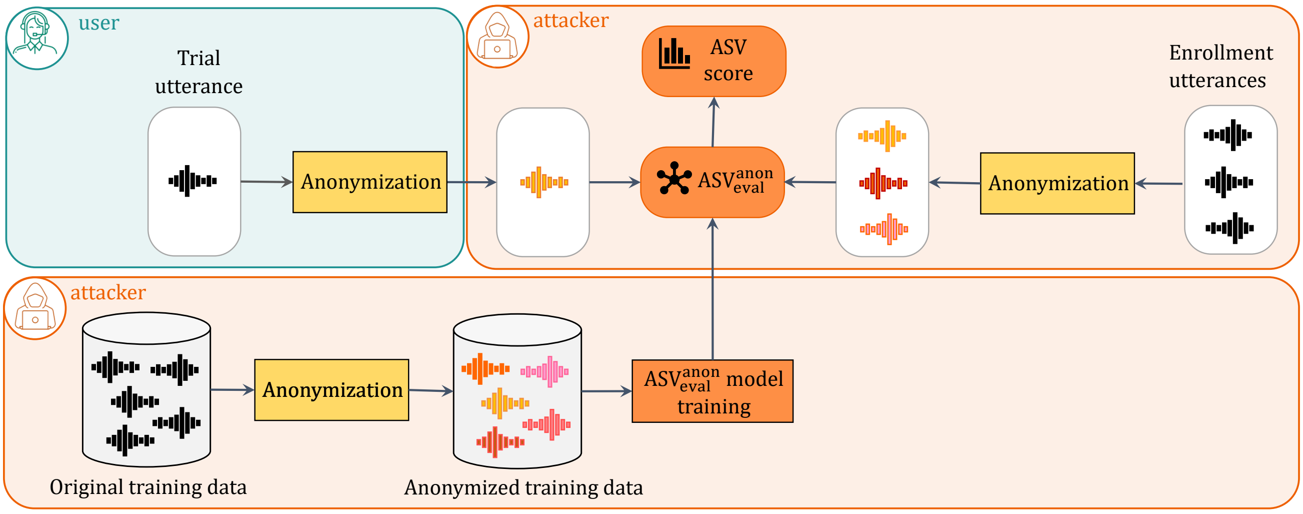

Figure 1: Baseline attacker: training $ASV_{\text{eval}}^{\text{anon}}$ on anonymized training data and using it to compare anonymized trial and enrollment data.

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

[4] See *Step 2: Evaluation*, *run_ evaluation.py*

Table 4: EER achieved by the baseline attacker system $ASV_{\text{eval}}^{\text{anon}}$ on data processed by different anonymization systems vs. EER achieved on the original (Orig.) unprocessed data by an ASV model trained on original data.

Dataset	**Gender**	**EER (%)**							
		Orig.	**B3**	**B4**	**B5**	**T8-5**	**T10-2**	**T12-5**	**T25-1**
LibriSpeech-dev	female	10.51	28.43	34.37	35.82	39.63	43.63	43.32	42.65
	male	0.93	22.04	31.06	32.92	40.84	40.04	44.10	40.06
Average dev		5.72	25.24	32.71	34.37	40.24	41.83	43.71	41.36
LibriSpeech-test	female	8.76	27.92	29.37	33.95	42.50	41.97	43.61	42.34
	male	0.42	26.72	31.16	34.73	40.05	38.75	41.88	41.92
Average eval		4.59	27.32	30.26	34.34	41.28	40.36	42.75	42.13

6 Evaluation rules

- Participants are free to develop their own attacker systems, using components of the provided baseline or not. They are encouraged (but not required) to submit results for each anonymization system and to design attacker systems that target the specific weaknesses of each anonymization system.
- Participants can use the training resources and development datasets specified in Section 3 in order to train their system and tune hyperparameters.
- To compute the score for the pair {set of enrollment utterances, trial utterance} only the utterances included in this pair can be used from the evaluation data.
- Anonymization system authors and members of their team are welcome to participate. Their results will be considered as official only when attacking other systems than their own. If the authors release the code for the corresponding anonymization system before the deadline for *"Declaration of additional training/development data and models"* (see Table 5), they can also attack their own anonymization system and participate in official ranking for this system.

7 Registration and submission of results

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 48 hours after successful registration, otherwise or in case of any questions they should contact the organizers:
attacker.challenge@inria.fr.

Submission of results. Each participant may submit scores/results for one or more attacker systems, each targeting all anonymization systems or only some of them. Each single submission should include the EER and corresponding ASV scores (for the development and evaluation data) obtained with the proposed attacker system for 4 trial lists (in the same format as generated by the baseline attacker system[5]):

- data/libri_dev_trials_f/trials (example: libri_dev_enrolls-libri_dev_trials_f scores)
- data/libri_dev_trials_m/trials (example: libri_dev_enrolls-libri_dev_trials_m scores)
- data/libri_test_trials_f/trials (example: libri_test_enrolls-libri_test_trials_f scores)
- data/libri_test_trials_m/trials (example: libri_test_enrolls-libri_test_trials_m scores)

All data should be submitted in the form of a single compressed archive.

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 the participants mailing list and via the VoicePrivacy Attacker Challenge website.

Special session at ICASSP 2025. Results will be presented at the ICASSP 2025 special session to which 5 selected top-ranked participants will be invited to submit and present their challenge systems. All participants will be invited to submit the extended versions of their papers to the SPSC 2025 Symposium. Accepted papers will be published in the ICASSP proceedings. According to `https://2025.ieeeicassp.org/call-for-gc-proposals/`: *"The review process is coordinated by the challenge organizers and the SPGC chairs. All 2-page papers should be covered by an ICASSP registration and should be presented in person at the conference."*

[5] Example: link

8 Schedule

The result submission deadline is **5th December 2024**.

Table 5: Important dates

Release of the training, development and evaluation data, baselines and evaluation software	September 2024
Declaration of additional training/development data and models	13th October 2024
Publication of the full final list of training data and models	15th October 2024
Deadline for participants to submit scores, evaluation results and system descriptions	5th December 2024
Deadline for participants to submit 2-page papers to ICASSP-2015 (by invitation only)	9th December 2024
Paper Acceptance Notification	30th December 2024

9 Acknowledgement

This work was supported by the French National Research Agency under project Speech Privacy and project IPoP of the Cybersecurity PEPR.

References

[1] 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 *Proc. Interspeech 2020*, 2020, pp. 1693–1697. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1333

[2] A. Nautsch, A. Jimenez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado *et al.*, "Preserving privacy in speaker and speech characterisation," *Computer Speech and Language*, vol. 58, pp. 441–480, 2019.

[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, 2022, https://arxiv.org/pdf/2109.00648.pdf. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230822000080

[4] 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.

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

[6] M. Panariello, N. Tomashenko, X. Wang, X. Miao, P. Champion, H. Nourtel, M. Todisco, N. Evans, E. Vincent, and J. Yamagishi, "The VoicePrivacy 2022 Challenge: Progress and perspectives in voice anonymisation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[7] 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.

[8] J.-F. Bonastre, H. Delgado, N. Evans, T. Kinnunen, K. A. Lee, X. Liu, A. Nautsch, P.-G. Noe, J. Patino, M. Sahidullah, B. M. L. Srivastava, M. Todisco, N. Tomashenko, E. Vincent, X. Wang, and J. Yamagishi, "Benchmarking and challenges in security and privacy for voice biometrics," pp. 52–56, 2021.

[9] 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 *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 2802–2806.

[10] 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.

[11] P. Champion, T. Thebaud, G. Le Lan, A. Larcher, and D. Jouvet, "On the invertibility of a voice privacy system using embedding alignment," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 191–197.

[12] C. Zeng, X. Wang, E. Cooper, X. Miao, and J. Yamagishi, "Attention back-end for automatic speaker verification with multiple enrollment utterances," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6717–6721.

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

[14] 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.

[15] 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.* PMLR, 2023, pp. 28 492–28 518.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Transactions on Machine Learning Research*, 2023. [Online]. Available: https://openreview.net/forum?id=ivCd8z8zR2

[21] 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 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2022, pp. 8102–8106.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," 2015.

[28] 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.

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

[30] 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," *arXiv preprint arXiv:1912.06670*, 2019.

[31] 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," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[32] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "LIRIS-ACCEDE: A video database for affective content analysis," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 43–55, 2015.

[33] N. Vryzas, R. Kotsakis, A. Liatsou, C. A. Dimoulas, and G. Kalliris, "Speech emotion recognition for performance interaction," *Journal of the Audio Engineering Society*, vol. 66, no. 6, pp. 457–467, 2018.

[34] A. Aouf, "Basic arabic vocal emotions dataset," 2020. [Online]. Available: https://www.kaggle.com/ds/345828

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

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

[37] Y. Lin, M. Cheng, F. Zhang, Y. Gao, S. Zhang, and M. Li, "Voxblink2: A 100k+ speaker recognition corpus and the open-set speaker-identification benchmark," in *Interspeech 2024*, 2024, pp. 4263–4267.

[38] 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.

[39] E. Parada-Cabaleiro, G. Costantini, A. Batliner, M. Schmitt, and B. W. Schuller, "Demos: An italian emotional speech corpus: Elicitation methods, machine learning, and perception," *Language Resources and Evaluation*, vol. 54, no. 2, pp. 341–383, 2020.

[40] 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.

[41] 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.

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

[43] H. Delgado, N. Evans, J.-w. Jung, T. Kinnunen, I. Kukanov, K. A. Lee, X. Liu, H.-j. Shim, M. Sahidullah, H. Tak *et al.*, "Asvspoof 5 evaluation plan," Tech. Rep., 2024. [Online]. Available: https://www.asvspoof.org/file/ASVspoof5___Evaluation_Plan_Phase2.pdf

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

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

[46] Y. Chen, S. Zheng, H. Wang, L. Cheng, Q. Chen, and J. Qi, "An enhanced Res2Net with local and global feature fusion for speaker verification," in *INTERSPEECH*, 2023.

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

[48] H. Wang, S. Zheng, Y. Chen, L. Cheng, and Q. Chen, "CAM++: A fast and efficient network for speaker verification using context-aware masking," in *INTERSPEECH*, 2023.

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

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

[51] 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 *Proc. Interspeech 2021*, 2021, pp. 1344–1348.

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

[53] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "AutoVC: Zero-shot voice style transfer with only autoencoder loss," ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. Long Beach, California, USA: PMLR, 09–15 Jun 2019, pp. 5210–5219. [Online]. Available: http://proceedings.mlr.press/v97/qian19c.html

[54] 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," 2024. [Online]. Available: https://www.voiceprivacychallenge.org/vp2024/docs/VPC-2024-.pdf

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

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

[57] P. Champion, "Anonymizing speech: Evaluating and designing speaker anonymization techniques," Ph.D. dissertation, Université de Lorraine, 2023.

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

[59] J. Yao, N. Kuzmin, Q. Wang, P. Guo, Z. Ning, D. Guo, K. A. Lee, E.-S. Chng, and L. Xie, "NPU-NTU System for Voice Privacy 2024 Challenge," *arXiv preprint arXiv:2409.04173*, 2024.

[60] N. Kuzmin, H.-T. Luong, J. Yao, L. Xie, and K. A. Lee, "NTU-NPU System for Voice Privacy 2024 Challenge," *SPSC 2024*, 2024. [Online]. Available: https://www.voiceprivacychallenge.org/vp2024/docs/T12_____NTU-NPU_System_for_Voice_Privacy_2024_Challenge.pdf

[61] W. Gu, Z. Liu, L. Chen, R. Wang, C. Guo, W. Guo, K. A. Lee, and Z.-H. Ling, "USTC-PolyU system for the VoicePrivacy 2024 Challenge," *SPSC 2024*, 2024. [Online]. Available: https://www.voiceprivacychallenge.org/vp2024/docs/T25_____USTC-PolyU_system_for_the_VoicePrivacy_2024_Challenge.pdf

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