The Voice Privacy 2024 Challenge Evaluation Plan

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

The VoicePrivacy 2024 Challenge Evaluation Plan

Version **2.0**

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

[1]Inria, France
[2]Singapore Institute of Technology, Singapore
[3]Institute for Natural Language Processing, University of Stuttgart, Germany
[4]National Institute of Informatics, Tokyo, Japan
[5]Audio Security and Privacy Group, EURECOM, France

https://voiceprivacychallenge.org

For new participants — Executive summary

- The task is to develop a voice anonymization system for speech data which conceals the speaker's voice identity while protecting 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 formed on the basis of the participants' requests. Participants apply their developed anonymization systems, run evaluation scripts and submit evaluation results and anonymized speech data to the organizers.
- Results will be presented at a workshop held in conjunction with Interspeech 2024 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. 2022

- In line with the considered application scenarios, the requirements that anonymization preserves voice distinctiveness and intonation are removed, hence the associated G_{VD} and ρ^{F_0} metrics are no longer used. All the data are anonymized on the *utterance level*.
- An extended list of datasets and pretrained models, formed on the basis of the participants' requests, will be provided for training anonymization systems.
- The complexity of the evaluation protocol and the running time of the evaluation scripts have been greatly reduced. The scripts are now primarily in Python, which makes it easy for participants who are new to the field to catch up.
- Only objective evaluation will be performed. Three complementary metrics will be used: the equal error rate (EER) as the privacy metric and two utility metrics, namely the word error rate (WER) for automatic speech recognition (ASR) and the unweighted average recall (UAR) for speech emotion recognition (SER).
- Models for utility evaluation (ASR and SER) are trained on original (unprocessed) data to ensure that linguistic and emotional content is undistorted. These models are provided with the evaluation scripts, hence utility evaluation is much faster.

Changes in version 2.0 w.r.t. 1.0

- The final list of data and models to build and train anonymization systems (Table 1).
- New anonymization baselines: **B3**, **B4**, **B5**, and **B6** (Sections 5.3, 5.4, 5.5) and results (Section 5.6).

1 Challenge objectives

Speech data fall within the scope of privacy regulations such as the European General Data Protection Regulation (GDPR). Indeed, they encapsulate a wealth of personal (a.k.a. personally identifiable) information such as the speaker's identity, age and gender, health status, personality, racial or ethnic origin, geographical background, social identity, and socio-economic status [1]. Formed in 2020, the VoicePrivacy initiative [2] is spearheading efforts to develop privacy preservation solutions for speech technology. So far, it has focused on promoting the development of *anonymization* solutions which conceal all personal information, facilitating their comparison using common datasets and protocols, and defining meaningful evaluation metrics through a series of competitive benchmarking challenges. The first two editions of VoicePrivacy were held in 2020 and 2022 [2–7]. VoicePrivacy 2024, the third edition, starts in March 2024 and culminates in the VoicePrivacy Challenge workshop held in conjunction with the 4nd Symposium on Security and Privacy in Speech Communication (SPSC)[1], a joint event co-located with Interspeech 2024[2] in Kos Island, Greece.

Anonymization requires a combination of solutions to alter not only the speaker's voice, but also linguistic content, extra-linguistic traits, and background sounds which might reveal the speaker's identity. In keeping with the previous VoicePrivacy Challenge editions, the current edition focuses on the subgoal of *voice anonymization*, that is the task of altering the speaker's voice to conceal their identity to the greatest possible extent, while leaving the linguistic content and paralinguistic attributes intact. Specifically, this edition focuses on preserving the emotional state, that is the key paralinguistic attribute in many real-world application scenarios of voice anonymization, e.g., in call centers to enable the use of third-party speech analytics. In the following, we often refer to "voice anonymization" as "anonymization" alone for the sake of conciseness.

This document describes the challenge task, the data, pretrained models and baseline systems that participants can use to build their own voice anonymization system, and the evaluation metrics and rules that will be used for assessment, in addition to guidelines for registration and submission.

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 [2,8,9]. Here, we consider the scenario where the user shares anonymized utterances for downstream automatic speech recognition (ASR) and speech emotion recognition (SER) tasks, and the attacker wants to identify the speakers from their anonymized utterances.

2.1 Voice anonymization task

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. It should:

(a) output a speech waveform;

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

(c) not distort the linguistic and emotional content.

The utterance-level anonymization requirement (b) means that the voice anonymization system must assign a pseudo-speaker to each utterance independently of the other utterances. The pseudo-speaker assignment process 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 will measure the WER and UAR obtained by ASR and SER systems trained on original (unprocessed) data.

[1] 4th Symposium on Security and Privacy in Speech Communication: http://www.spsc2024.mobileds.de/
[2] https://www.interspeech2024.org/

2.2 Attack model

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

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

(a) several enrollment utterances for each speaker;

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

Using this information, the attacker anonymizes the enrollment utterances to reduce the mistmatch with the trial utterances, and trains an ASV system adapted to that specific anonymization system. This attack model is the strongest known to date, hence we consider it as the most reliable for privacy assessment.

The protection of identity information is assessed via a *privacy* metric. Specifically, we will measure the EER obtained by the attacker.

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 resources

In addition to the training data used in the previous challenge editions and those used to train the baseline anonymization systems (see Section 5), the participants were allowed to propose additional resources to build and train anonymization systems before the deadline (20th March). These include both data and pretrained models. 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 anonymization systems. All the allowed resources are listed in Table 1.

For models # 1, 2, 3, 4, 5, 7, and 8, 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 21st March 2024 can be used by participants in the development and training of their anonymization systems. Participants are allowed to use any existing software in the development and training of their anonymization systems. If the software uses pretrained models, these models should be explicitly listed in this table. This includes models # 32-35 and the models listed on the main page (readme) of the repository # 36 before 21st March 2024.

For the purpose of the challenge, the *MSP-Podcast* [10] corpus providers can share the *MSP-Podcast* corpus for companies using the academic license. If a company wants to use the corpus beyond this challenge, it will have to obtain a commercial license by approaching the corpus providers.

Table 1: Final list of models and data for training anonymization systems.

#	**Model**	**Link**
1	WavLM Base and Large [11]	https://github.com/microsoft/unilm/tree/master/wavlm
2	Whisper [12]	https://github.com/openai/whisper
3	HuBERT [13]	https://github.com/facebookresearch/fairseq/blob/main/examples/hubert
4	XLS-R [14]	https://github.com/facebookresearch/fairseq/blob/main/examples/wav2vec/xlsr
5	wav2vec 2.0 [15]	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 [16]	https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim
7	ContentVec [17]	https://github.com/auspicious3000/contentvec
8	w2v-BERT [18]	https://github.com/facebookresearch/fairseq/tree/ust/examples/w2vbert
9	ECAPA2 [19]	https://huggingface.co/Jenthe/ECAPA2
10	ECAPA-TDNN [20]	https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

11	NaturalSpeech 3 [21]	https://huggingface.co/amphion/naturalspeech3_facodec
12	NVIDIA Hifi-GAN Vocoder (en-US) [22]	https://huggingface.co/nvidia/tts_hifigan
13	CRDNN on CommonVoice 14.0 English	https://huggingface.co/speechbrain/asr-crdnn-commonvoice-14-en
14	Encodec [23]	https://huggingface.co/facebook/encodec_24khz
15	Bark	https://huggingface.co/suno/bark https://huggingface.co/erogol/bark/tree/main

#	**Dataset**	**Link**
16	ESD [24]	https://hltsingapore.github.io/ESD/download.html
17	LibriSpeech [25]: train-clean-100, train-clean-360, train-other-500	https://www.openslr.org/12
18	CREMA-D [26]	https://github.com/CheyneyComputerScience/CREMA-D
19	RAVDESS [27]	https://datasets.activeloop.ai/docs/ml/datasets/ravdess-dataset/ https://zenodo.org/records/1188976
20	VCTK [28]	https://datashare.ed.ac.uk/handle/10283/2651 https://huggingface.co/datasets/vctk
21	SAVEE [29]	http://kahlan.eps.surrey.ac.uk/savee/ https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee
22	EMO-DB [30]	http://emodb.bilderbar.info/download/
23	LJSpeech [31]	https://keithito.com/LJ-Speech-Dataset/
24	Libri-light [32] (only train part)	https://github.com/facebookresearch/libri-light/blob/main/data_preparation/README.md
25	VoxCeleb-1,2 [33]	https://www.robots.ox.ac.uk/~vgg/data/voxceleb/index.html#about
26	LibriTTS [34]: train-clean-100, train-clean-360, train-other-500	https://openslr.org/60/
27	CMU-MOSEI [35]	http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/
28	MUSAN [36]	https://www.openslr.org/17/
29	RIR [37]	https://www.openslr.org/28/
30	VGAF [38] (from EmotiW challenge)	https://sites.google.com/view/emotiw2023 https://www.kaggle.com/datasets/amirabdrahimov/vgaf-dataset
31	MSP-Podcast [10]	https://ecs.utdallas.edu/research/researchlabs/msp-lab/MSP-Podcast.html

#	**Software with pre-trained models**	**Link**
32	Resemblyzer	https://github.com/resemble-ai/Resemblyzer Model: https://github.com/resemble-ai/Resemblyzer/blob/master/resemblyzer/pretrained.pt
33	VITS [39]	https://github.com/jaywalnut310/vits/ Models: https://drive.google.com/drive/folders/1ksarh-cJf3F5eKJjLVWYOX1j1qsQqiS2
34	PIPER pretrained on VITS	https://github.com/rhasspy/piper/?tab=readme-ov-file Models: https://huggingface.co/datasets/rhasspy/piper-checkpoints/tree/main
35	RVC-Project	https://github.com/RVC-Project Models: https://huggingface.co/lj1995/VoiceConversionWebUI/tree/main
36	DISSC [40]	https://github.com/gallilmaimon/DISSC

3.2 Development and evaluation data

Development and evaluation data comprise subsets of the following corpora:

- ***LibriSpeech***[3] [25] is a corpus of read English speech derived from audiobooks and designed for ASR research. It contains 960 hours of speech sampled at 16 kHz. This data will be used for ASV and ASR evaluation. The *LibriSpeech* evaluation and development sets are the same as in the previous challenge editions.
- ***IEMOCAP*** [41] is an emotional audio-visual dataset that will be used for SER evaluation. It contains 12 hours of speech sampled at 16 kHz corresponding to improvised and scripted two-speaker conversations between 5 female and 5 male English actors. We consider only 4 emotions out of the 9 annotated ones: *neutral*, *sadness*, *anger*, and *happiness*. Following [42–44], we merge the original happiness and excitement classes into the happiness class to balance the number of utterances in each class. To accommodate for the small number of speakers and the small amount of data, we adopt a leave-one-conversation out cross-validation protocol. In each cross-validation fold, four conversations (eight speakers) are used to train the SER evaluation system[4], while the two speakers from the remaining conversation form the development and evaluation sets, respectively.

A detailed description of the datasets provided for development and evaluation is presented in Tables 2 and 3 below.

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

Subset			**Female**	**Male**	**Total**	**#Utterances**
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

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

<table>
<tr><th colspan="2">Conversation</th><th>#Utterances</th><th>Fold 1</th><th>Fold 2</th><th>Fold 3</th><th>Fold 4</th><th>Fold 5</th></tr>
<tr><td rowspan="2">Session 1</td><td>Female</td><td>528</td><td>Dev</td><td rowspan="2">Train</td><td rowspan="4">Train</td><td rowspan="6">Train</td><td rowspan="8">Train</td></tr>
<tr><td>Male</td><td>557</td><td>Eval</td></tr>
<tr><td rowspan="2">Session 2</td><td>Female</td><td>481</td><td rowspan="6">Train</td><td>Eval</td></tr>
<tr><td>Male</td><td>542</td><td>Dev</td></tr>
<tr><td rowspan="2">Session 3</td><td>Female</td><td>522</td><td rowspan="4">Train</td><td>Dev</td></tr>
<tr><td>Male</td><td>629</td><td>Eval</td></tr>
<tr><td rowspan="2">Session 4</td><td>Female</td><td>528</td><td rowspan="4">Train</td><td>Eval</td></tr>
<tr><td>Male</td><td>503</td><td>Dev</td></tr>
<tr><td rowspan="2">Session 5</td><td>Female</td><td>590</td><td rowspan="2">Train</td><td>Eval</td></tr>
<tr><td>Male</td><td>651</td><td>Dev</td></tr>
</table>

4 Privacy and utility evaluation

We consider one objective privacy metric to assess the speaker re-identification risk and two objective utility metrics to assess the fulfillment of the downstream tasks specified in Section 2.

4.1 Objective assessment of the privacy-utility tradeoff

Three metrics will be used for the objective ranking of submitted systems: the equal error rate (EER) as the privacy metric and two utility metrics: word error rate (WER) and unweighted average recall (UAR). These metrics rely on automatic speaker verification (ASV), automatic speech recognition (ASR), and speech emotion recognition (SER) systems. The ASV system is trained on *LibriSpeech-train-clean-360* and the ASR system on the full *LibriSpeech-train-960* dataset, whose statistics are presented in Table 4. The SER system for each *IEMOCAP* cross-validation fold is trained on the corresponding *IEMOCAP* training subset, whose statistics are reported in Table 3. Training and evaluation will be performed with the provided recipes and models.[5] More specifically, models for privacy evaluation will be trained by participants on their anonymized training data with the provided training scripts, while the models for utility evaluation are provided by the organizers.

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

[4] Trained SER evaluation systems corresponding to the 5 folds are provided by the organizers. The participants should not use this data for their own training purposes.

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

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

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

As in the 2022 edition, multiple evaluation conditions specified with a set of minimum target privacy requirements will be considered. For each minimum target privacy requirement, submissions that meet this requirement will be ranked according to the resulting utility for each utility metric separately. The goal is to measure the privacy-utility trade-off at multiple operating points, e.g. when systems are configured to offer better privacy at the cost of utility and vice versa. This approach to assessment aligns better the VoicePrivacy Challenge with the user expectation of privacy and allows for a more comprehensive evaluation of each solution, while also providing participants with a set of clear optimisation criteria. The privacy and utility metrics will be used for this purpose.

Minimum target privacy requirements are specified with a set of N minimum target EERs: $\{\text{EER}_1, \ldots, \text{EER}_N\}$. Each minimum target EER constitutes a separate evaluation condition. Participants are encouraged to submit solutions to as many conditions as possible. Submissions to any one condition i should achieve an average EER on the VoicePrivacy 2024 evaluation set greater than the corresponding EER_i. The set of valid submissions for each EER_i will then be ranked according to the corresponding WER and UAR. The VoicePrivacy 2024 Challenge involves $N = 4$ conditions with minimum target EERs of: $\text{EER}_1 = 10\%$, $\text{EER}_2 = 20\%$, $\text{EER}_3 = 30\%$, $\text{EER}_4 = 40\%$.

The lower the WER for a given EER condition, the better the rank of the considered system in ASR results ranking. Similarly, the higher the UAR for a given EER condition, the better the rank of the considered system in SER results ranking. A depiction of example results and system rankings according to this methodology is shown in Figure 1.

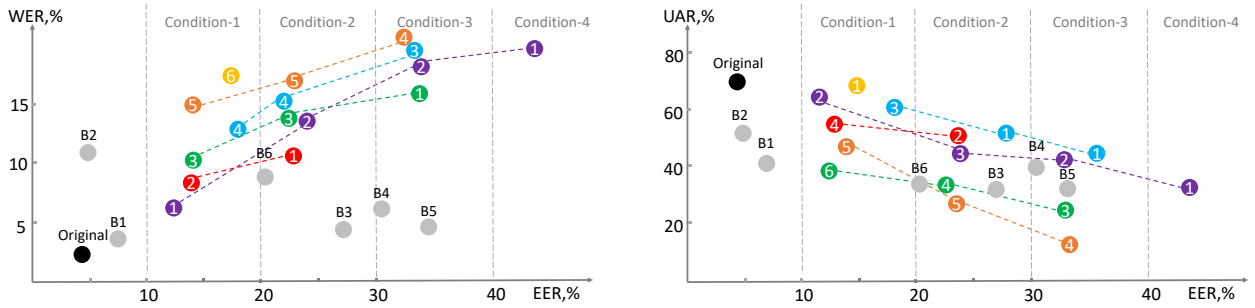

Figure 1: Example system rankings according to the privacy (EER) and utility (WER and UAR) results for 4 minimum target EERs. Different colors correspond to 6 different teams. Numbers within each circle show system ranks for a given condition. Grey circles correspond to the baseline systems, and the black circle to the original (unprocessed) data.

4.2 Privacy metric: equal error rate (EER)

The ASV system used for privacy evaluation is an ECAPA-TDNN [20] with 512 channels in the convolution frame layers, implemented by adapting the *SpeechBrain* [45] *VoxCeleb* recipe to *LibriSpeech*. As seen in Figure 2, we consider a *semi-informed* attacker, who has access to the anonymization system under evaluation [3,9]. Using that system, the attacker anonymizes the original enrollment data so as to reduce the mismatch with the anonymized trial data. In addition, the attacker anonymizes the *LibriSpeech-train-clean-360* dataset and retrains the ASV system (denoted $ASV_{\text{eval}}^{\text{anon}}$) on it, so that it is adapted to this specific anonymization system.[6] Anonymization is conducted on the *utterance level*, using the same pseudo-speaker assignment process as the trial data. For a given speaker, all enrollment utterances are used to compute an average speaker vector for enrollment.

For every pair of enrollment and trial speaker vectors in the *LibriSpeech* development and evaluation sets, the cosine similarity score is computed 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,

[6]It is critical that the $ASV_{\text{eval}}^{\text{anon}}$ system is well trained, indeed a badly trained system can overestimate the EER and give a false sense of privacy [46]. The organizers will use the anonymized data submitted by the participants to check it. In the event when some submissions do not satisfy it, the organizers reserve the right to modify the ASV evaluation scripts or to mark those submissions accordingly to ensure a fair competition.

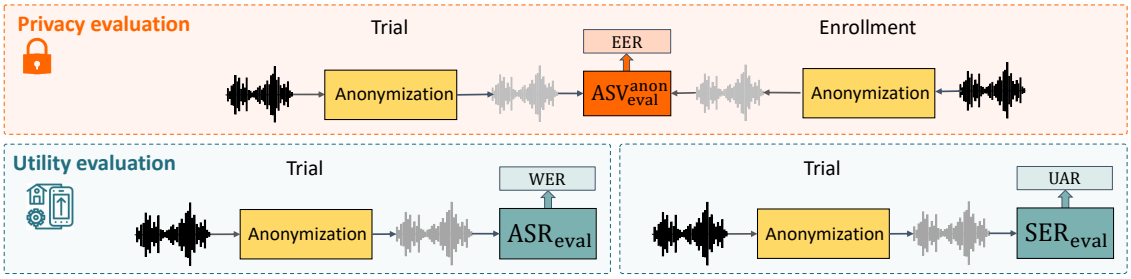

Figure 2: Privacy and utility evaluation.

Table 5: Number of same-speaker and different-speaker pairs considered for evaluation.

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

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 higher the WER, the greater the privacy. The number of same-speaker and different-speaker pairs is given in Table 5.

4.3 Utility metrics

4.3.1 Word error rate (WER)

The ability of the anonymization system to leave the linguistic content undistorted is assessed using an ASR system[7] (denoted ASR_{eval}) fine-tuned on *LibriSpeech-train-960* from *wav2vec2-large-960h-lv60-self*[8] using a *SpeechBrain* recipe. Unlike the 2022 challenge edition, this ASR evaluation model is fixed, and trained and fine-tuned on original (unprocessed) data.

For every anonymized trial utterance in the *LibriSpeech* development and evaluation sets, the ASR system outputs a word sequence. The WER is calculated as

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

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

4.3.2 Unweighted average recall (UAR)

The ability of the anonymization system to leave the emotional content undistorted is assessed using an SER system (denoted SER_{eval}) trained using the *SpeechBrain* recipe for SER on *IEMOCAP*. It is a wav2vec2-based model that has been trained separately for each of the training folds in Table 3.

Within each fold, emotion recognition performance is quantified on the anonymized *IEMOCAP* development and evaluation sets using the standard UAR metric calculated as the sum of class-wise recalls R_i divided by the number of classes N_{class}:

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

The recall R_i for each class i is computed as number of true positives divided by the total number of samples in that class. The obtained UARs are then averaged across the five folds. The higher the UAR, the greater the utility.

5 Baseline voice anonymization systems

Baseline voice anonymization systems are released to help participants develop their own system. We provide a description and evaluation results for two established baseline systems inspired from past challenge editions,

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

that will be used to gauge progress with respect to these editions (**B1** and **B2**). In addition, we also have released several new baseline systems (**B3**, **B4**, **B5**, and **B6**) that better protect privacy and have different performance in utility. Note that the training data for the new baseline systems may differ for each method. All the data and models used in their development can be used by the challenge participants in the training of their anonymization system. These data and models are included in the list of training resources (see Section 3).

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

The baseline anonymization system **B1** is based on a common approach to x-vector modification and speech synthesis. It is identical to the **B1.b** baseline from the VoicePrivacy 2022 Challenge [6], except that anonymization is now performed on the utterance level instead of the speaker level.

B1 is based on the voice anonymization method proposed in [47] and shown in Figure 3. Anonymization is performed in three steps:

- **Step 1 – Feature extraction:** extraction of the speaker x-vector [48], the fundamental frequency (F0) and bottleneck (BN) features from the original audio waveform.
- **Step 2 – X-vector anonymization:** generation of an anonymized (pseudo-speaker) x-vector using an external pool of speakers.
- **Step 3 – Speech synthesis:** synthesis of an anonymized speech waveform from the anonymized x-vector and the original BN and F0 features using a neural source-filter (NSF) model.

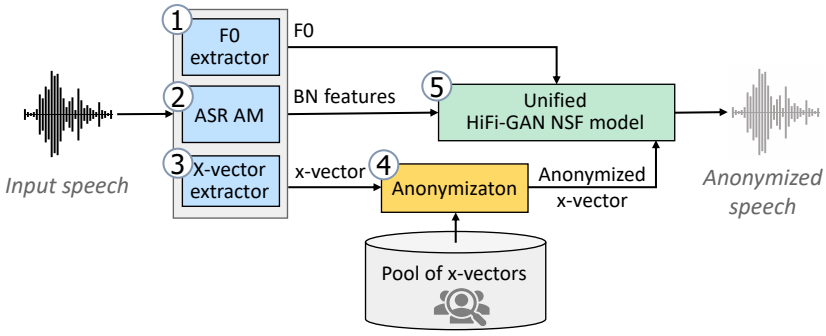

Figure 3: Baseline anonymization system **B1**.

In order to implement these steps, four different models are required, as shown in Figure 3. Details for training these components are presented in Table 6.

In *Step 1*, to extract BN features, an ASR acoustic model (AM) is trained (#1 in Table 6). We assume that the BN features represent the linguistic content of the speech signal. The ASR AM has a factorized time delay neural network (TDNN-F) model architecture [49, 50] and is trained using the Kaldi toolkit [51]. To encode speaker information, an x-vector extractor with a TDNN model topology (#2 in Table 6) is also trained using Kaldi.

In *Step 2*, for a given source speaker, a new anonymized x-vector is computed by averaging a set of candidate x-vectors from the speaker pool. Probabilistic linear discriminant analysis (PLDA) is used as a distance measure between these vectors and the x-vector of the source speaker. The candidate x-vectors for averaging are chosen in two steps. First, for a given source x-vector, the N farthest x-vector candidates in the speaker pool are selected. Second, a smaller subset of N^* candidates are chosen randomly among those N vectors ($N = 200$ and $N^* = 100$). The x-vectors for the speaker pool are extracted from a disjoint dataset (*LibriTTS-train-other-500*).

In *Step 3*, the NSF model used to synthesize the anonymized speech waveform is trained on *LibriTTS-train-clean-100* in the same manner as HiFi-GAN [22] using the HiFi-GAN discriminators. After training, the discriminators can be safely discarded, and only the trained NSF model is used in the anonymization system.

More details about this model can be found in the scripts for VoicePrivacy 2022[9] and in [52, 53].

[9]To perform *utterance-level* (in contrast to *speaker-level*) anonymization of the enrollment and trial data for **B1**, the corresponding parameters should be setup in config.sh: `anon_level_trials=utt` and `anon_level_enroll=utt`.

[10]pYAAPT: `http://bjbschmitt.github.io/AMFM_decompy/pYAAPT.html`

Table 6: Modules and training corpora for the anonymization system **B1**. The module indexes are the same as in Figure 3. Superscript numbers represent feature dimensions.

#	Module	Description	Output features	Data
1	F0 extractor	pYAAPT[10], uninterpolated	$F0^{1}$	-
2	ASR AM	TDNN-F Input: $MFCC^{40}$ + i-$vectors^{100}$ 17 TDNN-F hidden layers Output: 6032 triphone ids LF-MMI and CE criteria	BN^{256} features extracted from the final hidden layer	LibriSpeech: train-clean-100 train-other-500
3	X-vector extractor	TDNN Input: $MFCC^{30}$ 7 hidden layers + 1 stats pooling layer Output: 7232 speaker ids CE criterion	speaker x-$vectors^{512}$	VoxCeleb-1,2
4	X-vector anonymization module		pseudo-speaker x-$vectors^{512}$	(Pool of speakers) LibriTTS: train-other-500
5	NSF model	sinc-hn-NSF in [54] + HiFi-GAN discriminators [22] Input: $F0^{1}$+ BN^{256} + x-$vectors^{512}$ Training criterion defined in Hifi-GAN [22]	speech waveform	LibriTTS: train-clean-100

5.2 Anonymization using the McAdams coefficient: B2

The second baseline anonymization system **B2** shown in Figure 4 is identical to the **B2** baseline from the VoicePrivacy 2022 Challenge [6]. In contrast to **B1**, it does not require any training data and is based upon simple signal processing techniques. It is a randomized version of the anonymization method proposed in [55], which employs the McAdams coefficient [56] to shift the pole positions derived from linear predictive coding (LPC) analysis of speech signals.

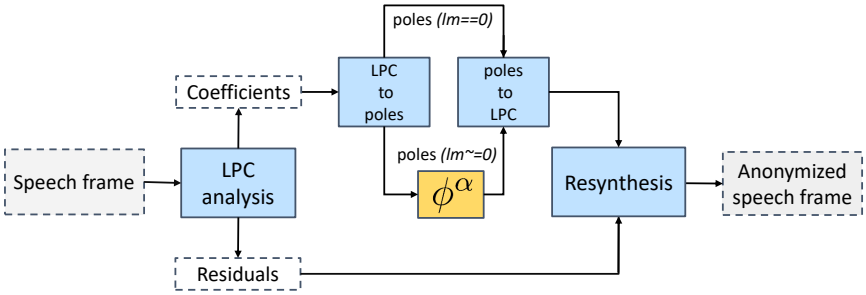

Figure 4: Baseline anonymization system **B2**.

B2 starts with the application of frame-by-frame LPC source-filter analysis to derive LPC coefficients and residuals. The residuals are set aside for later resynthesis, whereas the LPC coefficients are converted into pole positions in the z-plane by polynomial root-finding. Each pole corresponds to a peak in the spectrum, resembling a formant position. The McAdams' transformation is applied to the phase of each pole: while real-valued poles are left unmodified, the phase ϕ (between 0 and π radians) of poles with non-zero imaginary parts is raised to the power of the McAdams' coefficient α so that transformed poles have new, shifted phases of ϕ^{α}. The coefficient α is sampled for each utterance from a uniform distribution: $\alpha \sim U(\alpha_{\min}, \alpha_{\max})$, with $\alpha_{\min} = 0.5$ and $\alpha_{\max} = 0.9$. It implies a contraction or expansion of the pole positions around $\phi = 1$ radian. For a sampling rate of 16 kHz, i.e. for the data used in this challenge, $\phi = 1$ radian corresponds to approximately 2.5 kHz which is the approximate mean formant position [57]. The corresponding complex conjugate poles are similarly shifted in the opposite direction and the new set of poles, including original real-valued poles, are converted back into LPC coefficients. Finally, the LPC coefficients and the residuals are used to resynthesise a new speech frame in the time domain.

5.3 Anonymization using phonetic transcriptions and GAN: B3

The baseline **B3**, shown in Figure 5, is a system based on speech synthesis conditioned on keeping linguistic and general prosodic information while replacing the speaker embedding [58]. The core part of the baseline is a generative adversarial network (GAN) that generates artificial pseudo-speaker embeddings [59]. Anonymization is performed in three steps:

- **Step 1 – Feature extraction:** extraction of the speaker embedding, phonetic transcription, fundamental frequency (F0), energy, and phone duration from the original audio waveform.
- **Step 2 – Speaker embedding anonymization; pitch and energy modification.**
- **Step 3 – Speech synthesis:** synthesis of an anonymized speech waveform from the anonymized speaker embedding, modified F0 and energy features, original phonetic transcripts and original phone durations.

Different models are required to implement these steps, as shown in Figure 5. Details for training these components are presented in Table 7.

In *Step 1*, the speaker embedding is extracted using an adapted global style tokens model [60]. The phonetic transcription is obtained using an end-to-end ASR model with a hybrid CTC-attention architecture, a Branchformer encoder and a Transformer decoder.

In *Step 2*, the original speaker embedding is replaced by an artificial one generated by a Wasserstein GAN [61]. If the cosine distance between the artificial and the original embeddings exceeds 0.3, they are assumed to be dissimilar enough. Otherwise another artificial embedding is generated until this condition is satisfied. Furthermore, the pitch and energy values of each phone are multiplied by random values generated uniformly and independently between 0.6 and 1.4 to remove individual prosodic patterns while keeping the general prosody of the utterance. The random values are chosen for each phone individually.

In *Step 3*, the anonymized speaker embedding, modified prosody, and original phonetic transcription are fed into a speech synthesis system based on *FastSpeech2* [62] and HiFi-GAN [22] as implemented in *IMS-Toucan* [63] to synthesize the anonymized speech.

Table 7: Modules and training corpora for the anonymization system **B3**. The module indexes are the same as in Figure 5. Superscript numbers represent feature dimensions.

#	Module	Description	Output features	Data
1	Prosody extractor	Phone aligner: 6-layer CNN + LSTM with CTC loss F0 estimation using Praat F0, energy, durations normalized by each vector's mean	F0^{1}, energy^{1} phone durations^{1}	LibriTTS: train-clean-100
2	ASR	End-to-end with hybrid CTC-attention [64] Input: log mel Fbank^{80} Encoder: Branchformer [65] Decoder: Transformer Output: phone sequences CTC and attention criteria	phonetic transcript with pauses and punctuation	LibriTTS: train-clean-100 train-other-500
3	Speaker embedding extractor	GST [60], trained jointly with SS model Input: mel spectrogram^{80} 6 hidden layers + 4-head attention Output: GST speaker embedding^{128}	GST speaker embedding^{128}	LibriTTS: train-clean-100
4	Prosody modification module	Value-wise multiplication of F0 and energy with random values in $[0.6, 1.4)$	F0^{1}, energy^{1}	-
5	Speaker anonymization module	Wasserstein GAN Input: Random noise^{16} from normal distribution Generator: ResNet with three residual blocks, 150k params Critic: ResNet with three residual blocks, 150k params Output: MSE and Quadratic Transport Cost [66] criteria	pseudo-speaker GST embeddings^{128}	LibriTTS: train-clean-100 RAVDESS [27] ESD [24]
6	SS model	*IMS Toucan* [63] implementation of *FastSpeech2* [62] Input: F0^{1} + energy^{1} + phone duration^{1} + phonetic transcript + GST embeddings^{128} Training criterion defined in *FastSpeech2* [62]	mel spectrogram^{80}	LibriTTS: train-clean-100
7	Vocoder	HiFi-GAN vocoder [22] Input: mel spectrogram^{80} Training criterion defined in Hifi-GAN [22]	speech waveform	LibriTTS: train-clean-100

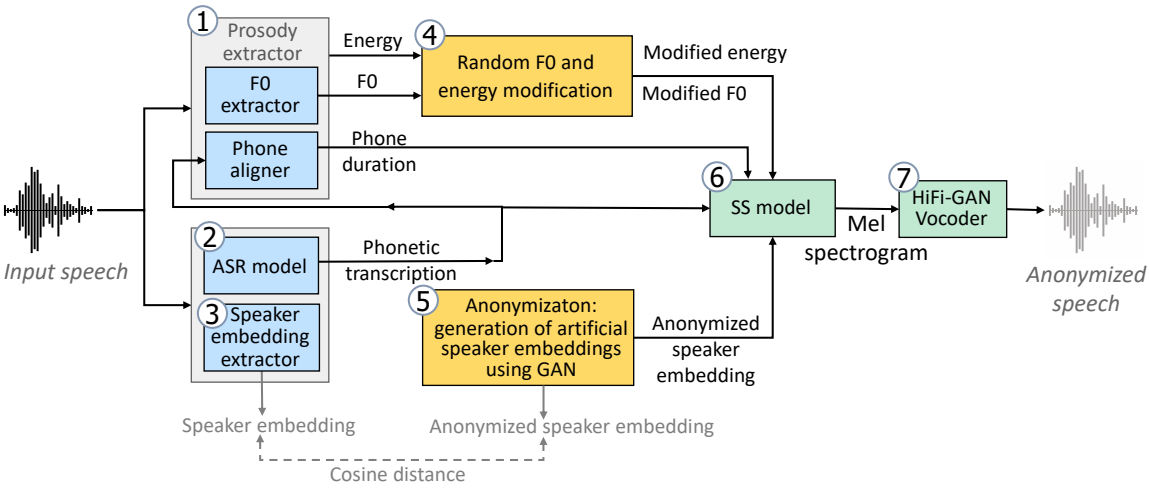

Figure 5: Baseline anonymization system **B3**.

5.4 Anonymization using neural audio codec (NAC) language modeling: B4

System **B4**, proposed in [67], is based upon the technique of neural audio codec (NAC) language modeling [68, 69]. A NAC is an encoder-decoder neural network whereby an audio signal can be encoded into a sequence of discrete *acoustic tokens* $\mathbf{a} \in \{1, \ldots, N_Q\}^{Q \times T_A}$ and subsequently decoded to a waveform again; here, T_A is the number of time frames the input waveform is divided into, and Q is the number of tokens associated to a time frame, each drawn from one of Q different token dictionaries. Tokens are integers ranging from 1 to N_Q. In the context of **B4**, acoustic tokens are assumed to capture the characteristics of an individual's speech. Several sets of acoustic tokens are extracted from the speech of a pool of pseudo-speakers, obtaining a *pool of acoustic prompts* $\mathbf{A}$. Given a speech signal to anonymize, system **B4** uses a *semantic extractor* to extract from it a sequence of discrete *semantic tokens* $\mathbf{s} \in \{1, \ldots, N_S\}^{T_S}$, where T_S is the number of time frames and N_S is the maximum integer value that a semantic token can take. The semantic tokens encode the spoken content of the utterance. They are concatenated with a randomly chosen sequence of acoustic tokens $\tilde{\mathbf{a}} \in \mathbf{A}$ to form a single sequence $(\mathbf{s}, \tilde{\mathbf{a}})$. A GPT-like, decoder-only Transformer then uses said sequence as a prompt and auto-regressively generates a continuation of acoustic tokens $\mathbf{a}$ that respects both the semantics encoded in $\mathbf{s}$ and the speech style encoded in $\tilde{\mathbf{a}}$. The decoder module of the NAC is used to convert $\mathbf{a}$ to a waveform that preserves the semantic content of the original input audio, but is associated to a different pseudo-speaker. An outline of system **B4** is shown in Figure 6.

The NAC is EnCodec[11] [23], which is trained with speech segments from the DNS Challenge [70] and *Common Voice* [71], along with non-speech audio data from *AudioSet* [72], *FSD50K* [73], and *Jamendo* [74]. The semantic extractor is composed of a HuBERT feature extractor [13] trained on *LibriSpeech-train-960*, and a LSTM back-end that predicts a token index from the feature vector at each time frame. The decoder-only model is a publicly available checkpoint from Bark[12], although its authors do not disclose its training data. Further architectural details of the implementation are provided in Table 8.

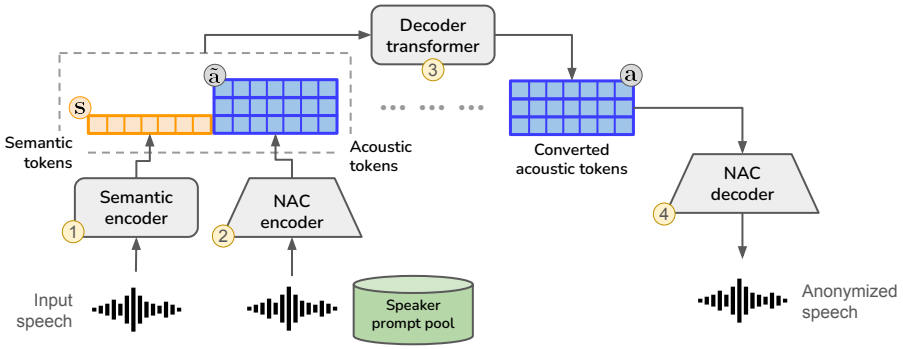

Figure 6: Baseline anonymization system **B4**.

[10] pYAAPT: http://bjbschmitt.github.io/AMFM_decompy/pYAAPT.html
[11] https://github.com/facebookresearch/encodec
[12] https://github.com/suno-ai/bark

Table 8: Modules and training corpora for anonymization system **B4**.

#	Module	Description	Output features
1	Semantic encoder	HuBERT Base quantized https://dl.fbaipublicfiles.com/hubert/hubert_base_ls960.pt	1 semantic token per time step
2	NAC encoder	EnCodec 24 KHz [23] encoder https://huggingface.co/facebook/encodec_24khz	8 acoustic tokens per time step
3	Decoder transformer	Two 12-layer, decoder-only transformers operating on different ranges of acoustic tokens Taken from Bark[12] [67] https://huggingface.co/erogol/bark/tree/main	8 acoustic tokens per time step
4	NAC decoder	EnCodec 24 KHz [23] decoder	speech waveform

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

This anonymization pipeline elaborated in [46] shares similarities with **B1** and showcases some improvements. In particular, it exclusively relies on PyTorch for execution and has been optimized for fast inference. Furthermore, it incorporates vector quantization (VQ) to enhance the disentanglement of linguistic and speaker attributes.

The pipeline leverages feature extractors to capture the fundamental frequency (F0) utilizing a Torch version of YAAPT, and acoustic VQ bottleneck (VQ-BN) features from an ASR AM specifically trained to identify left-biphones. Subsequently, the VQ-BN features, F0, and a designated speaker (represented as a one-hot vector corresponding to a speaker encountered during training) are directly used to synthesize an anonymized speech waveform via a HiFi-GAN network (see Figure 7).

We consider two different ASR AMs for feature extraction that correspond to two baselines:

B5: the AM combines a pretrained wav2vec2 model with three additional TDNN-F layers;

B6: the AM consists solely of 12 TDNN-F layers.

On the final TDNN-F layer, following the first activation (inner bottleneck of the TDNN-F with 256 dimensions), vector quantization (VQ) is applied. This process approximates a continuous vector with another vector of equivalent dimensions, the latter belonging to a finite set of vectors. The incorporation of VQ into this framework serves the purpose of minimizing the encoding of speaker information within the BN features, thereby enhancing the disentanglement property.

The wav2vec2 model is pre-trained on 24.1k hours of unlabeled multilingual west Germanic speech from *VoxPopuli*[13], then fine-tuned on *LibriSpeech train-clean-100*. The TDNN-F-only ASR-BN extractor takes Mel filterbank features as inputs and is trained on *LibriSpeech train-other-500* and *train-clean-100*. The HiFi-GAN model is trained on *LibriTTS-train-clean-100* for each ASR-BN extractor. *LibriTTS-train-clean-100* contains 247 speakers, so 247 possible one-hot vectors.

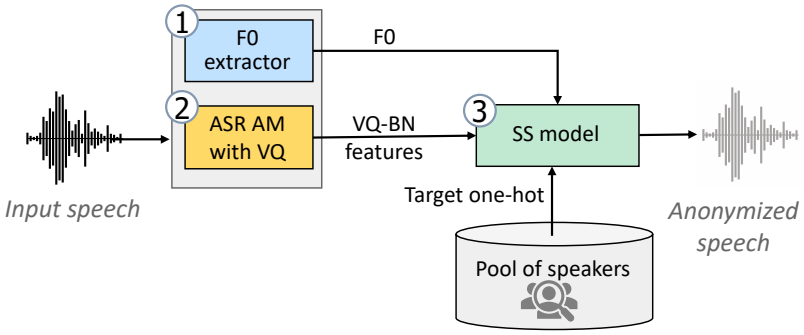

Figure 7: Baseline anonymization systems **B5** and **B6**.

5.6 Results

Results for the baselines are reported in Tables 9, 10, and 11. For the two old baselines, we can see that **B1** achieves a better average EER and WER than **B2**, while **B2** achieves a better UAR than **B1**. Results

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

for the four new baselines **B3**, **B4**, **B5**, and **B6** show better privacy protection than the two old baselines and fall into two privacy categories according to the privacy thresholds (Section 4.1). The highest EER is achieved by **B5**. The best result among all the baseline systems in terms of WER is obtained by **B1**, and the best result in terms of UAR by **B2**.

6 Evaluation 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*.
- Participants are strongly encouraged to make multiple submissions corresponding to different privacy-utility tradeoffs (see Section 4.2).
- The three metrics (EER, WER, UAR) will be used for system ranking on the provided development and evaluation sets. Within each EER interval – [10,20), [20,30), [30,40), [40,100) – systems will be ranked separately in order of (1) increasing WER and (2) decreasing UAR.
- 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 the development and evaluation sets and the *LibriSpeech-train-clean-360* dataset used to train the ASV evaluation model using the same anonymization system. They must then train the 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 Post-evaluation analysis

The organizers will run additional post-evaluation experiments in order to further characterize the performance of submitted systems. To do so, we ask all participants to share with us the anonymized speech data obtained when running their anonymization system on the training, development and evaluation sets. Further details about these experiments will follow in due course.

8 Registration and submission of results

8.1 Registration

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

organisers@lists.voiceprivacychallenge.org.

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

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

Table 9: EER (%) achieved on data anonymized by the baselines vs. original (Orig.) data.

Dataset	**Gender**	**EER,%**						
		Orig.	**B1**	**B2**	**B3**	**B4**	**B5**	**B6**
LibriSpeech-dev	female	10.51	10.94	12.91	28.43	34.37	35.82	25.14
	male	0.93	7.45	2.05	22.04	31.06	32.92	20.96
Average dev		5.72	9.20	7.48	25.24	32.71	34.37	23.05
LibriSpeech-test	female	8.76	7.47	7.48	27.92	29.37	33.95	21.15
	male	0.42	4.68	1.56	26.72	31.16	34.73	21.14
Average test		4.59	6.07	4.52	27.32	30.26	34.34	21.14

Table 10: WER (%) achieved on data processed by the baselines vs. original (Orig.) data.

Dataset	WER,%						
	Orig.	B1	B2	B3	B4	B5	B6
LibriSpeech-dev	1.80	3.07	10.44	4.29	6.15	4.73	9.69
LibriSpeech-test	1.85	2.91	9.95	4.35	5.90	4.37	9.09

Table 11: UAR (%) achieved on data processed by the baselines vs. original (Orig.) data.

Dataset	UAR,%						
	Orig.	B1	B2	B3	B4	B5	B6
IEMOCAP-dev	69.08	42.71	55.61	38.09	41.97	38.08	36.39
IEMOCAP-test	71.06	42.78	53.49	37.57	42.78	38.17	36.13

8.2 Submission of results

Each participant may submit as many systems as they wish for each minimum target EER provided in Section 4.2. In the case of three or more submissions per condition, the organisers will only include the system with the lowest WER and the system with the highest UAR in the official ranking. These two systems (or this system in case it's the same one) will be ranked in terms of both WER and UAR.

Each single submission should include a compressed archive containing:

1. Directories with the result files, the corresponding cosine similarity scores (saved in `exp/asv_orig/cosine_out` and `exp/asv_anon<anon_data_suffix>/cosine_out`), and additional information generated by the evaluation scripts:
   - `exp/results_summary`
   - `exp/asv_orig`
   - `exp/asv_anon<anon_data_suffix>`
   - `exp/asr`
   - `exp/ser/*csv`.
2. The corresponding anonymized speech data (wav files, 16 kHz, with the same names as in the original corpus) generated from the development and evaluation sets and from the *LibriSpeech-train-clean-360* dataset used to train the ASV evaluation model. For evaluation, the wav files will be converted to 16-bit signed integer PCM format, and this format is recommended for submission. 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 on the development and evaluation sets is saved in a single file `exp/results_summary`) [14].

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.

[14] Example *results* files for the baseline systems:

- **B1**: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2024/blob/main/results/result_for_rank_b1b
- **B2**: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2024/blob/main/results/result_for_rank_mcadams
- **B3**: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2024/blob/main/results/result_for_rank_sttts
- **B4**: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2024/blob/main/results/result_for_rank_nac
- **B5**: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2024/blob/main/results/result_for_rank_asrbn_hifigan_bn_tdnnf_wav2vec2_vq_48_v1
- **B6**: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2024/blob/main/results/result_for_rank_asrbn_hifigan_bn_tdnnf_600h_vq_48_v1

9 VoicePrivacy Challenge workshop at Interspeech 2024

The VoicePrivacy 2024 Challenge will culminate in a joint workshop held in Kos Island, Greece in conjunction with **Interspeech 2024** and in cooperation with the ISCA SPSC Symposium.[1] VoicePrivacy 2024 Challenge participants are encouraged to submit papers describing their challenge entry according to the paper submission schedule (see Section 10). Paper submissions must conform to the format of the ISCA SPSC Symposium proceedings, detailed in the author's kit[15], 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. Since subjective evaluation results will be released only after the submission deadline, challenge papers should report only objective evaluation results. 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.

10 Schedule

The result submission deadline is **15th June 2024**. 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 2024.

Table 12: Important dates

Release of evaluation data, software and baselines	8th March 2024
Deadline for participants to submit a list for training data and models	20th March 2024
Publication of the full final list of training data and models	21st March 2024
Submission of challenge papers to the joint SPSC Symposium and VoicePrivacy Challenge workshop	15th June 2024
Deadline for participants to submit objective evaluation results, anonymized data, and system descriptions	15th June 2024
Author notification for challenge papers	5th July 2024
Final paper upload	25th July 2024
Joint SPSC Symposium and VoicePrivacy Challenge workshop	6th September 2024

11 Acknowledgement

This work was supported by the French National Research Agency under project Speech Privacy and project IPoP of the Cybersecurity PEPR and jointly by the French National Research Agency and the Japan Science and Technology Agency under project VoicePersonae. The challenge organizers thank Ünal Ege Gaznepoğlu for his help with the code base.

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.

[15] https://interspeech2024.org/author-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] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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*, 2023.

[17] 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*. PMLR, 2022, pp. 18 003–18 017.

[18] 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 *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 244–250.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

[32] 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.

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

[34] 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.

[35] 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.

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

[37] 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.

[38] 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.

[39] 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.

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

[41] 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.

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

[43] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," in *Interspeech*, 2018, pp. 247–251.

[44] H. Nourtel, P. Champion, D. Jouvet, A. Larcher, and M. Tahon, "Evaluation of speaker anonymization on emotional speech," in *1st ISCA Symposium on Security and Privacy in Speech Communication (SPSC)*, 2021, pp. 62–66.

[45] 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.

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

[47] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker anonymization using x-vector and neural waveform models," in *Speech Synthesis Workshop*, 2019, pp. 155–160.

[48] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

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

[50] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Interspeech*, 2015, pp. 3214–3218.

[51] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011.

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

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

[54] X. Wang and J. Yamagishi, "Neural harmonic-plus-noise waveform model with trainable maximum voice frequency for text-to-speech synthesis," in *Speech Synthesis Workshop*, 2019, pp. 1–6.

[55] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, "Speaker anonymisation using the McAdams coefficient," in *Interspeech*, 2021, pp. 1099–1103.

[56] S. McAdams, "Spectral fusion, spectral parsing and the formation of the auditory image," Ph.D. dissertation, Stanford University, 1984.

[57] S. Ghorshi, S. Vaseghi, and Q. Yan, "Cross-entropic comparison of formants of British, Australian and American English accents," *Speech Communication*, vol. 50, no. 7, pp. 564–579, 2008.

[58] 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.

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

[60] 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.

[61] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning (ICML)*, 2017, pp. 214–223.

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

[63] F. Lux, J. Koch, A. Schweitzer, and N. T. Vu, "The IMS Toucan system for the Blizzard Challenge 2021," in *Blizzard Challenge Workshop*, 2021.

[64] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[65] Y. Peng, S. Dalmia, I. Lane, and S. Watanabe, "Branchformer: Parallel MLP-attention architectures to capture local and global context for speech recognition and understanding," in *International Conference on Machine Learning (ICML)*, 2022, pp. 17 627–17 643.

[66] H. Liu, X. Gu, and D. Samaras, "Wasserstein GAN with quadratic transport cost," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4832–4841.

[67] 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.

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

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

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

[71] 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 *12th Language Resources and Evaluation Conference (LREC)*, 2020, p. 4218-4222.

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

[73] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: An open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.

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