The VoicePrivacy 2022 Challenge Evaluation Plan

Version 1.0

Natalia Tomashenko[1], Xin Wang[2], Xiaoxiao Miao[2], Hubert Nourtel[3], Pierre Champion[3,4], Massimiliano Todisco[5], Emmanuel Vincent[3], Nicholas Evans[5], Junichi Yamagishi[2,6], and Jean-François Bonastre[1]

[1]Laboratoire Informatique d'Avignon (LIA), Avignon Université, France
[2]National Institute of Informatics, Tokyo, Japan
[3]Université de Lorraine, CNRS, Inria, LORIA, France
[4]LIUM, Le Mans Université, France
[5]Audio Security and Privacy Group, EURECOM, France
[6]University of Edinburgh, UK

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, paralinguistic attributes, intelligibility and naturalness.
- Training, development and evaluation datasets are provided in addition to 3 different baseline anonymization systems, evaluation scripts, and metrics. Participants apply their developed anonymization systems, run evaluation scripts and submit objective evaluation results and anonymized speech data to the organizers.
- Results will be presented at a workshop held in conjunction with INTERSPEECH 2022 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. 2020

- A stronger, semi-informed **attack model** in the form of an automatic speaker verification (ASV) system trained on anonymized (per-utterance) speech data.
- Complementary **metrics** comprising the equal error rate (EER) as a privacy metric, the word error rate (WER) as a primary utility metric, and the pitch correlation ρ^{F_0} and gain of voice distinctiveness G_{VD} as secondary utility metrics.
- A new **ranking** policy based upon a set of minimum target privacy requirements.

1 Context

Recent years have seen mounting calls for the preservation of privacy when treating or storing personal data. This is not least the result of recent European privacy legislation, e.g., the general data protection regulation (GDPR). While there is no legal definition of privacy [1], speech data is likely to fall within the scope of privacy regulation. 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]. Speaker recognition systems can also reveal the speaker's identity. Formed in 2020, the VoicePrivacy initiative [3] is spearheading the effort to develop privacy preservation solutions for speech technology. We aim to foster progress in the development of anonymization and pseudonymization solutions which suppress personally identifiable information contained within recordings of speech while preserving linguistic content, paralinguistic attributes, intelligibility and naturalness. VoicePrivacy takes the form of a competitive benchmarking challenge, with common datasets, protocols and metrics. The first edition of VoicePrivacy was held in 2020 [3–6]. VoicePrivacy 2022, the second edition, starts in March 2022 and culminates in the VoicePrivacy Challenge workshop held in conjunction with the 2nd Symposium on Security and Privacy in Speech Communication (SPSC)[1], a joint event co-located with INTERSPEECH 2022[2] in Incheon, Korea.

Anonymization refers to the goal of suppressing personally identifiable information in the speech signal, leaving other attributes intact [4]. Note that, in the legal community, the term "anonymization" means that this goal has been achieved. Here, it refers to the task to be addressed, even when the method being evaluated has failed. Anonymization requires altering not only the speaker's voice, but also linguistic content, extralinguistic traits, and background sounds which might reveal the speaker's identity. As a step towards this goal, and in keeping with the inaugural VoicePrivacy 2020 Challenge, the second edition focuses on *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.

This document describes plans for the challenge, the datasets, protocol and the set of metrics that will be used for assessment, in addition to evaluation rules and guidelines for registration and submission.

2 Challenge objectives

The grand objective of the VoicePrivacy Challenge is to foster progress in the development of anonymization techniques for speech data. The specific technical goals are summarised as follows. They are to:

- facilitate the development of novel techniques which suppress speaker-discriminative information within speech signals;
- promote techniques which provide effective anonymization while protecting linguistic content, paralinguistic attributes, intelligibility and naturalness;
- provide a level playing field to facilitate the comparison of different anonymization solutions using a common dataset and protocol;
- investigate metrics for the evaluation and meaningful comparison of different anonymization solutions.

VoicePrivacy participants will be provided with common datasets, protocols and a suite of software packages that will enable them to evaluate anonymization performance.

3 Task

Privacy preservation is formulated as a game between *users* who share some data and *attackers* who access this data or data derived from it and wish to infer information about the users [3, 7, 8]. To protect their privacy, the users share data that contain as little personal information as possible

[1] 2nd Symposium on Security and Privacy in Speech Communication: https://symposium2022.spsc-sig.org/
[2] https://www.interspeech2022.org/

while still allowing one or more downstream goals to be achieved. To infer personal information, the attackers may use additional prior knowledge.

Focusing on speech data, a given privacy preservation scenario is specified by: (i) the nature of the data: waveform, features, etc.; (ii) the information seen as personal: speaker identity, traits, linguistic content, etc.; (iii) the downstream goal(s): human communication, automated processing, model training, etc.; (iv) the data accessed by the attackers: one or more utterances, derived data or models, etc.; (v) the attackers' prior knowledge: previously collected speech data, the applied privacy preservation method, etc. Different specifications lead to different privacy preservation methods from the users' point of view and different attacks from the attackers' point of view.

Here, we consider the scenario where speakers want to hide their identity to the greatest possible extent while allowing the desired downstream goals to be achieved, while attackers want to identify the speakers from their utterances.

3.1 Anonymization task

The utterances shared by the users are referred to as *trial* utterances. In order to hide his/her identity, each user passes these utterances through a voice anonymization system prior to sharing. The resulting utterances sound as if they were 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 anonymization system. It should:

(a) output a speech waveform;

(b) conceal the speaker identity;

(c) leave the linguistic content and paralinguistic attributes unchanged;

(d) ensure that all trial utterances from a given speaker are uttered by the same pseudo-speaker, while trial utterances from different speakers are uttered by different pseudo-speakers.[3]

The requirement (c) promotes the achievement of all possible downstream goals to the greatest possible extent. In practice, we restrict ourselves two use cases: automatic speech recognition (ASR) training and/or decoding, and multi-party human conversations. The achievement of these goals is assessed via a range of *utility* metrics. Specifically, we will measure ASR performance using a model trained on anonymized data. In addition, the pitch correlation $\boldsymbol{\rho}^{F_0}$ between original (unprocessed) and anonymized speech signals will be used as a secondary objective utility metric to measure the degree to which intonation is preserved in anonymized speech, and subjective speech intelligibility and naturalness will also be measured.

The requirement (d) is motivated by the fact that, in a multi-party human conversation, the anonymized voices of all speakers must be distinguishable from each other and should not change over time. It will be assessed via the gain of voice distinctiveness G_{VD} metric.

3.2 Attack model

For each speaker of interest, the attacker is assumed to have access to one or more utterances spoken by that speaker. These utterances may or may not have been anonymized and are referred to as *enrollment* utterances.

In this work, the attackers have access to:

(a) one or more anonymized trial utterances;

(b) possibly, original or anonymized enrollment utterances for each speaker;

(c) anonymized training data (and can retrain an automatic speaker verification system using this data).

[3]We refer to this type of anonymization as *speaker-level* anonymization. An alternative approach is *utterance-level* anonymization where different utterances of the same source speaker are anonymized using different parameters of the anonymization system, so that they may sound as if they were uttered by different pseudo-speakers. For evaluation, we assume that only *speaker-level* anonymization should be applied to trial and enrollment data, while *utterance-level* anonymization will be applied to the training data for training strong attack models.

Table 1: Number of speakers and utterances in the training, development, and evaluation sets [3].

Subset			Female	Male	Total	#Utterances
Training	VoxCeleb-1,2		2,912	4,451	7,363	1,281,762
	LibriSpeech train-clean-100		125	126	251	28,539
	LibriSpeech train-other-500		564	602	1,166	148,688
	LibriTTS train-clean-100		123	124	247	33,236
	LibriTTS train-other-500		560	600	1,160	205,044
Development	LibriSpeech dev-clean	Enrollment	15	14	29	343
		Trial	20	20	40	1,978
	VCTK-dev	Enrollment	15	15	30	600
		Trial (different)				10,677
		Trial (common)				695
Evaluation	LibriSpeech test-clean	Enrollment	16	13	29	438
		Trial	20	20	40	1,496
	VCTK-test	Enrollment	15	15	30	600
		Trial (different)				10,748
		Trial (common)				700

The protection of identity information is assessed via *privacy* metrics, including objective and subjective speaker verifiability. These metrics assume different attack models. The objective speaker verifiability metrics assume that the attacker has access to a single anonymized trial utterance and several anonymized enrollment utterances but, for subjective speaker verifiability evaluation, to a single anonymized trial utterance and a single original enrollment utterance.

4 Data

Several publicly available corpora are used for the training, development and evaluation of voice anonymization systems. They are the same as for the VoicePrivacy 2020 Challenge [9] and comprise subsets from the following corpora:

- ***LibriSpeech***[4] [10] is a corpus of read English speech derived from audiobooks and designed for ASR research. It contains approximately 1,000 hours of speech sampled at 16 kHz.
- ***LibriTTS***[5] [11] is a corpus of English speech derived from LibriSpeech and designed for research in text-to-speech (TTS). It contains approximately 585 hours of read English speech sampled at 24 kHz.
- ***VCTK***[6] [12] is a corpus of read speech collected from 109 native speakers of English with various accents. It was originally aimed for research in TTS and contains approximately 44 hours of speech sampled at 48 kHz.
- ***VoxCeleb-1,2***[7] [13, 14] is an audiovisual corpus extracted from videos uploaded to YouTube and designed for speaker verification research. It contains approximately 2,770 hours of speech sampled at 16 kHz collected from 7,363 speakers, covering a wide range of accents and languages.

A detailed description of the datasets provided for training, development and evaluation is presented below and in Table 1.

Training set – The training set comprises the *VoxCeleb-1,2* corpus [13, 14] and subsets of the *LibriSpeech* [10] and *LibriTTS* [11] corpora. The selected subsets are detailed in Table 1 (top).

[4]Librispeech: http://www.openslr.org/12
[5]LibriTTS: http://www.openslr.org/60/
[6]VCTK, release version 0.92: https://datashare.is.ed.ac.uk/handle/10283/3443
[7]VoxCeleb: http://www.openslr.org/60/

Development set – The development set comprises *LibriSpeech dev-clean* and a subset of the VCTK corpus [12], denoted *VCTK-dev* (see Table 1, middle). Both are split into trial and enrollment subsets. For *LibriSpeech dev-clean*, speakers in the enrollment set are a subset of those in the trial set. For *VCTK-dev*, we use the same speakers for enrollment and trial and consider two trial subsets: *common* and *different*. The *common* subset comprises utterances $\#1-24$ in the VCTK corpus that are identical for all speakers. This choice is intended to support subjective evaluation of speaker verifiability in a text-dependent manner. The enrollment and *different* subsets comprise distinct utterances for all speakers.

Evaluation set – Similarly, the evaluation set comprises *LibriSpeech test-clean* and a subset of VCTK, denoted *VCTK-test* (see Table 1, bottom).

5 Utility and privacy metrics

We consider objective and subjective privacy metrics to assess speaker verifiability. We also propose objective and subjective utility metrics to assess fulfillment of the user goals specified in Section 3.

5.1 Objective assessment of the privacy-utility tradeoff

A pair of metrics will be used for the objective ranking of submitted systems: the equal error rate (EER) as the privacy metric and the word error rate (WER) as the primary utility metric. These metrics rely on automatic speaker verification (ASV) and automatic speech recognition (ASR) systems, both trained on the *LibriSpeech-train-clean-360* dataset, statistics for which are presented in Table 2. Training and evaluation will be performed with the provided recipes.[8]

Table 2: Statistics of the training dataset for the objective evaluation systems.

Subset	**Size,h**	**Number of Speakers**			**Number of Utterances**
		Female	**Male**	**Total**	
LibriSpeech: train-clean-360	363.6	439	482	921	104,014

New to the 2022 edition is the use of multiple evaluation conditions specified with a set of minimum target privacy requirements. Submissions to each condition that meet each minimum target privacy requirement will then be ranked according to their protection of utility. The goal is to measure the privacy-utility trade-off of any given solution at multiple operating points, e.g. when they 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 primary 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 2022 test datasets greater than the corresponding minimum EER_i. The average EER is computed by averaging the three EERs obtained on *LibriSpeech-test-clean*, *VCTK-test (common)* and *VCTK-test (different)* with weights of 0.5, 0.1 and 0.4, respectively.[9] The set of valid submissions for each minimum EER_i will then be ranked according to the corresponding average WER results computed by averaging the two WERs obtained on *LibriSpeech-test-clean* and *VCTK-test* with equal weights. The VoicePrivacy 2022 Challenge involves $N = 4$ conditions with minimum target EERs of: $\text{EER}_1 = 15\%$, $\text{EER}_2 = 20\%$, $\text{EER}_3 = 25\%$, $\text{EER}_4 = 30\%$. The lower the WER for a given EER condition, the better the rank of the considered system. A depiction of example results and system rankings according to this methodology is shown in Figure 1.

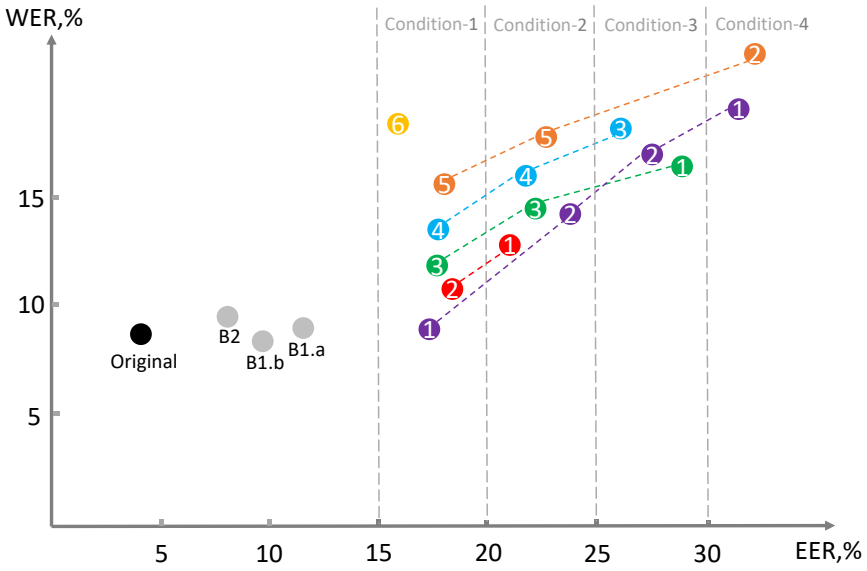

Figure 1: Example system rankings according to the privacy (EER) and utility (WER) results for 4 EER threshold conditions. Different colors correspond to 6 different teams. Numbers within each circle show system ranks for a given condition. Grey circles correspond to the baseline systems, and a black one – to the original (unprotected) system.

5.1.1 Privacy metric: equal error rate (EER)

The evaluation of objective speaker verifiability assumes that the attacker has access to one trial utterance and several enrollment utterances. The ASV system is based on x-vector speaker embeddings and probabilistic linear discriminant analysis (PLDA) [15]. For every pair of enrollment and trial x-vectors, it outputs a log-likelihood ratio (LLR) 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., $\text{EER} = P_{\text{fa}}(\theta_{\text{EER}}) = P_{\text{miss}}(\theta_{\text{EER}})$.

As seen in Figure 2, this metric is computed for two evaluation scenarios [4, 16]:

1. *Unprotected* — No anonymization is performed by users. The attacker has access to original (i.e., unprocessed) trial and enrollment data and uses an ASV system (denoted ASV_{eval}) trained on the original *LibriSpeech-train-clean-360* data.
2. *Semi-informed* [17] — Users anonymize their trial data. The attacker has access to original

[8]Evaluation scripts: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2022

[9]These weights assign the same importance to LibriSpeech and VCTK, and account for the different number of trials in the two VCTK subsets.

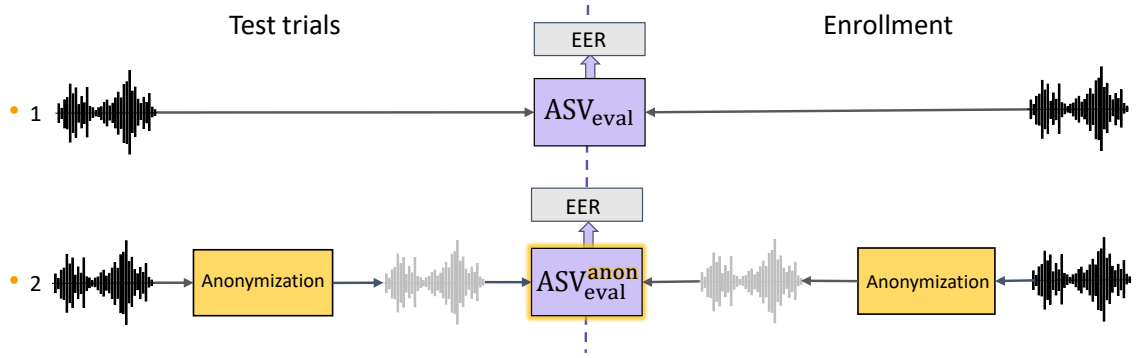

Figure 2: ASV evaluation (1) *unprotected*: original trial and enrollment data, ASV_{eval} trained on original data; (2) *semi-informed* attacker: *speaker-level* anonymized trial and enrollment data with different pseudo-speakers, $ASV_{\text{eval}}^{\text{anon}}$ trained on *utterance-level* anonymized data.

enrollment data and to the anonymization system, which is assumed to be publicly available. Using that system, the attacker performs *speaker-level* anonymization of the enrollment data so as to reduce the mismatch. The trial and enrollment data are anonymized using different pseudo-speakers, since the attacker does not know the pseudo-speaker chosen by each user. In addition, the attacker uses that system for *utterance-level* anonymization of the *LibriSpeech-train-clean-360* dataset and retrains the ASV system (now denoted $ASV_{\text{eval}}^{\text{anon}}$) on it. We found the latter to lead to a lower EER, i.e., a stronger attack, than *speaker-level* anonymization of the ASV training set [18, Table V]. This attack model is actually the strongest known to date, hence we consider it as the most reliable for privacy assessment.[10]

The number of same-speaker and different-speaker trials in the development and evaluation datasets is given in Table 3. For a given speaker, all enrollment utterances are used to compute an average x-vector for enrollment.

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
	VCTK-dev	Same-speaker (different)	1,781	2,015	3,796
		Same-speaker (common)	344	351	695
		Different-speaker (different)	13,219	12,985	26,204
		Different-speaker (common)	4,810	4,911	9,721
Evaluation	LibriSpeech test-clean	Same-speaker	548	449	997
		Different-speaker	11,196	9,457	20,653
	VCTK-test	Same-speaker (different)	1,944	1,742	3,686
		Same-speaker (common)	346	354	700
		Different-speaker (different)	13,056	13,258	26,314
		Different-speaker (common)	4,838	4,952	9,790

5.1.2 Primary utility metric: word error rate (WER)

The ability of the anonymization system to preserve linguistic information is assessed using an ASR system based on the Kaldi toolkit [20]. We adapted the Kaldi recipe for LibriSpeech involving an acoustic model with a factorized time delay neural network (TDNN-F) architecture [21, 22] and a *large* trigram language model. The ASR system outputs a word sequence and 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.

As shown in Figure 3, we consider two ASR evaluation scenarios: (1) the original trial data is decoded using the ASR model (denoted ASR_{eval}) trained on the original *LibriSpeech-train-clean-360* dataset, while (2) anonymized trial data is decoded using the ASR model (denoted $ASR_{\text{eval}}^{\text{anon}}$) trained on the anonymized (*utterance-level*) *LibriSpeech-train-clean-360* dataset. As demonstrated in [4, 19], the latter significantly decreases the WER on anonymized data compared to decoding with the ASR_{eval} model trained on original data.

5.2 Secondary utility metrics

In addition to the primary metrics, we consider two secondary utility metrics, namely pitch correlation ρ^{F_0} and the gain of voice distinctiveness G_{VD}. While these secondary metrics are not used for ranking, all submissions are expected to exceed a minimum pitch correlation threshold.

[10] In the VoicePrivacy 2020 Challenge, evaluation relied on three attack models referred to as *ignorant*, *lazy-informed*, and *semi-informed*, corresponding to attackers with different knowledge [4, 16, 19]. The *semi-informed* attack model differed from the one considered here, since it assumed *speaker-level* anonymization of the ASV training dataset.

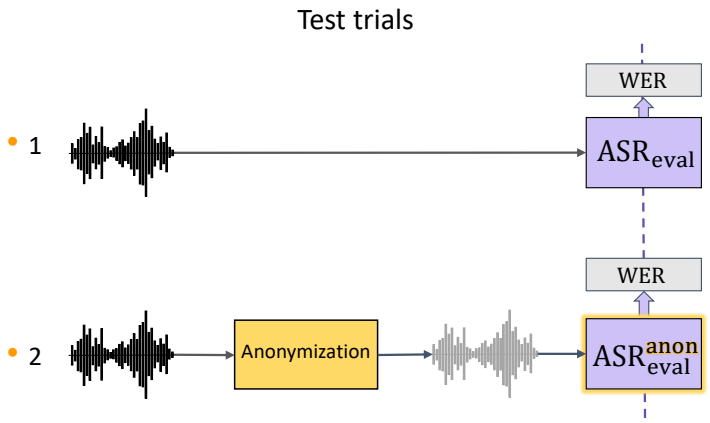

Figure 3: ASR evaluation (1) original data decoded with ASR_{eval} trained on original data; (2) *speaker-level* anonymized data decoded with $ASR_{\text{eval}}^{\text{anon}}$ trained on *utterance-level* anonymized data. WER is computed on the *trial* utterances of the development and evaluation datasets.

5.2.1 Pitch correlation metric ρ^{F_0}

The new pitch correlation metric has been introduced to provide a measure of how well anonymization preserves the intonation of the original utterance. Intonation is among *other speech attributes* that should remain intact following anonymisation.

The pitch correlation metric $\boldsymbol{\rho}^{F_0}$ is the Pearson correlation between the pitch sequences, estimated according to [23], of original and anonymized utterances. It is computed as follows. Pitch sequences are estimated for each pair of utterances and the shortest sequence is linearly interpolated so that its length matches that of the longest sequence. The temporal lag between original and anonymized utterances is then adjusted in order to maximise the Pearson cross-correlation. The latter is estimated using only segments where both original and anonymized utterances are voiced. Estimates of $\boldsymbol{\rho}^{F_0}$, calculated for development and evaluation datasets, are averages of the pitch correlation values for all utterances in each dataset.

While a secondary metric, **all** submissions should achieve a minimum average pitch correlation of $\rho^{F_0} > 0.3$ **for each dataset and for each condition**. Solutions that achieve lower average pitch correlations will be considered invalid. The threshold was set according to results derived from arbitrary anonymisation solutions that do not preserve intonation, e.g. ASR+TTS solutions. The threshold is a modest minimum correlation; all baseline solutions achieve average pitch correlations in the order of 0.7 hence submissions that make a reasonable attempt to preserve intonation should achieve correlation well above the minimum threshold.

5.2.2 Gain of voice distinctiveness $\mathbf{G_{VD}}$

The gain of voice distinctiveness metric aims to evaluate the requirement to preserve voice distinctiveness. It relies on voice similarity matrices [24, 25].

A voice similarity matrix $M = (M(i,j))_{1\leq i\leq N, 1\leq j\leq N}$ is defined for a set of N speakers using similarity values $M(i,j)$ computed for speakers i and j as follows:

$$M(i,j) = \text{sigmoid}\left(\frac{1}{n_i n_j} \sum_{\substack{1\leq k\leq n_i \text{ and } 1\leq l\leq n_j \\ k\neq l \text{ if } n_j=n_j}} \text{LLR}(x_k^{(i)}, x_l^{(j)})\right) \tag{1}$$

where $\text{LLR}(x_k^{(i)}, x_l^{(j)})$ is the log-likelihood-ratio obtained by comparing the k-th segment from the i-th speaker with the l-th segment from the j-th speaker, and where n_i and n_j are the numbers of segments for each speaker. Two matrices are computed using the ASV_{eval} model trained on original data: M_{oo} on original data and M_{aa} on anonymized data. For each of these two similarity matrices, the diagonal dominance $D_{\text{diag}}(M)$ is computed as the absolute difference between the mean values of diagonal and off-diagonal elements:

$$D_{\text{diag}}(M) = \left| \sum_{1\leq i\leq N} \frac{M(i,i)}{N} - \sum_{\substack{1\leq j\leq N \text{ and } 1\leq k\leq N \\ j\neq k}} \frac{M(j,k)}{N(N-1)} \right|. \tag{2}$$

The gain of voice distinctiveness metric (G_{VD}) [24] is defined as the ratio of diagonal dominance of the two matrices:

$$G_{\mathrm{VD}} = 10 \log_{10} \frac{D_{\mathrm{diag}}(M_{\mathrm{aa}})}{D_{\mathrm{diag}}(M_{\mathrm{oo}})}, \tag{3}$$

where a gain of $G_{\mathrm{VD}} = 0$ implies that the voice distinctiveness remains the same on average after anonymization, and a gain above or below 0 corresponds respectively to an average increase or decrease in voice distinctiveness.

5.3 Subjective metrics

Subjective metrics include: speaker verifiability; speech intelligibility; speech naturalness. They will each be evaluated via unified subjective evaluation tests carried out by the organizers as described below and as illustrated in Figure 4. The approach is similar to that used for the VoicePrivacy 2020 Challenge [4]. For naturalness and intelligibility assessments, evaluators will be asked to rate a single original or anonymized trial utterance at a time. For naturalness, the evaluator will assign a score from 1 ('totally unnatural') to 10 ('totally natural'). For intelligibility, the evaluator will assign a score from 1 ('totally unintelligible') to 10 ('totally intelligible'). Assessments of speaker verifiability will be performed with pairs of utterances, an original enrollment utterance and an original or anonymized trial utterance collected from the same or a different speaker. The evaluators will rate the similarity between the voices in enrollment and trial utterances using a scale of 1 to 10, where 1 denotes 'different speakers' and 10 denotes 'the same speaker'. Evaluators will be instructed to assign scores through a role-playing game. When an evaluator starts an evaluation session, the following instruction is displayed:

> "Please imagine that you are working at a TV or radio company. You wish to broadcast interviews of person X, but this person X does not want to disclose his/her identity. Therefore you need to modify speech signals in order to hide it. You have several automated tools to change speaker identity. Some of them hide the identity well, but severely degrade audio quality. Some of them hide the identity, but the resulting speech sounds very unnatural and may become less intelligible. In such cases, the privacy of person X is protected, but you will receive many complaints from the audience and listeners of TV/radio programs. You need to balance privacy of person X and satisfaction of TV/radio program audience and listeners. Your task is to evaluate such automated tools to change speaker identity and find out well-balanced tools."

Separate, detailed instructions are provided for the listening test involving each of the three subjective metrics. The evaluator is asked to imagine the described scene when evaluating the corresponding metric.

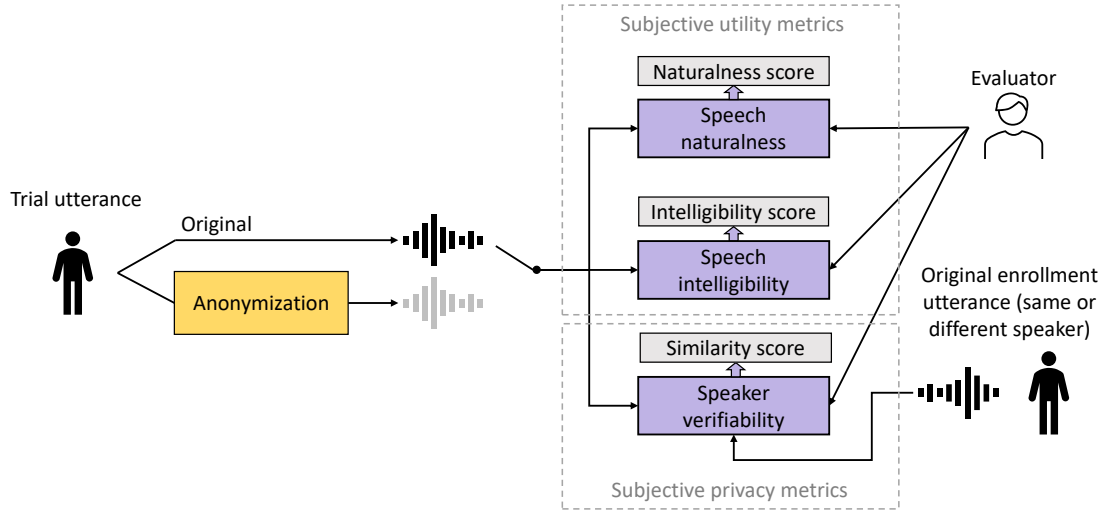

Figure 4: Subjective evaluation test for speech naturalness, intelligibility, and speaker verifiability [4].

“ **Subjective speech intelligibility**
For the final task, you are required to listen to audio A again and try to understand the audio content. Please judge how understandable audio A is.
You need to select one score between 1 and 10, where a higher score denotes higher intelligibility. In particular, 1 means “audio A is NOT understandable at all” and 10 means “audio A is perfectly understandable. ”

“ **Subjective speech naturalness**
You will listen to either original audio and audio modified by the above anonymization tools. Some of them result in artifacts and degradation due to poor audio processing.
Now, please listen to audio A and answer how much you can hear the audio degradation. Please judge based on the characteristics of the audio rather than what is being said.
You need to select a score between 1 and 10, where a higher score indicates less degradation. In particular, 1 means “audio A exhibits severe audio degradation” and 10 means “audio A does not exhibit any degradation”. Please note that the original audio includes background noise.”

“ **Subjective speaker verifiability**
Your next task is to compare the processed or unprocessed audio A with audio B. From the voices, you must determine whether they are from the same person or another person.
Now, please listen to audio A above and audio B below, and determine if they were uttered by the same speaker. Please judge based on the characteristics of the voice rather than what is being said.
You need to select one score between 1 and 10, where a higher score denotes higher speaker similarity. In particular, 1 means “audio A and B were uttered by different speakers for sure” and 10 means “audio A and B were uttered by the same speaker for sure. ”

By using clean speech of the same original speaker or a different speaker as Sample A, we will have anchors in the listening test and can visualize the performance of each participant system through detection error tradeoff (DET) curves [26] as described in [4]. These curves assume a detection task, where the decision for a given trial is made by comparing the score with a threshold. The false alarm and miss rates are computed as a function of the threshold and plotted against each other. For naturalness and intelligibility the task is to detect original data, while for speaker similarity the task is to detect whether the trial utterance is from the same speaker as the enrollment utterance. The closer the DET curves are to the top-right corner of each plot, the higher the naturalness, intelligibility, and privacy preservation.

6 Baselines

Three different baseline systems have been developed for the challenge.[11]

6.1 Anonymization using x-vectors and neural waveform models: B1.a and B1.b

The baseline anonymization systems **B1.a** and **B1.b** are based on a common approach to x-vector modification and on two different speech synthesis components.

6.1.1 B1.a

The first baseline **B1.a** is the primary baseline of the VoicePrivacy 2020 Challenge [3]. It is based on the voice anonymization method proposed in [27] and shown in Figure 5. Anonymization is performed in three steps:

- **Step 1 – Feature extraction:** extraction of the speaker x-vector [15], the fundamental frequency (F0) and bottleneck (BN) features from the original audio waveform.
- **Step 2 – X-vector anonymization:** anonymization of the source-speaker x-vector using an external pool of speakers.

[11] All baseline systems are available online: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2022

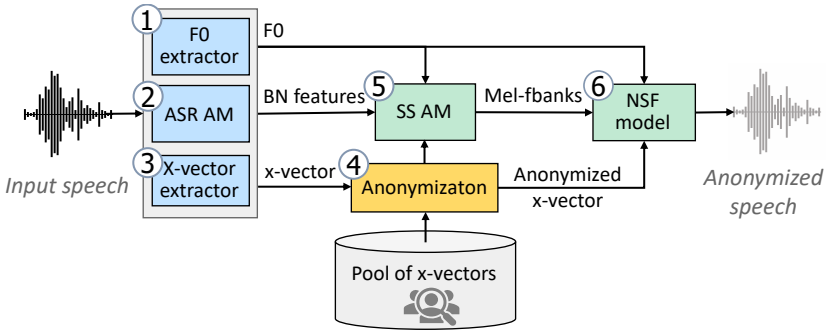

Figure 5: First baseline anonymization system **B1.a** [3].

- **Step 3 – Speech synthesis:** synthesis of a speech waveform from the anonymized x-vector and the original BN and F0 features using an acoustic model and a neural waveform model.

In order to implement these steps, four different models are required, as shown in Figure 5. Details for training these components are presented in Table 4.

In *Step 1*, to extract BN features, an ASR acoustic model (AM) is trained (#1 in Table 4). 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 [21, 22] and is trained using the Kaldi toolkit [20]. To encode speaker information, an x-vector extractor with a TDNN model topology (#2 in Table 4) 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)

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

Table 4: Modules and training corpora for the anonymization systems **B1.a** and **B1.b**. The module indexes are the same as in Figures 5 and 6. Superscript numbers represent feature dimensions.

#	**Module**	**Description**	**Output features**	**Data**
1	F0 extractor	pYAAPT[12], 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.a **B1.a**	Speech synthesis AM	Autoregressive (AR) network Input: $F0^{1}$+ BN^{256}+ x-$vectors^{512}$ FF * 2 + BLSTM + AR + LSTM * 2 + highway-postnet MSE criterion	Mel-$filterbanks^{80}$	LibriTTS: train-clean-100
5.b **B1.b**	Speech synthesis AM	sinc-hn-NSF in [28] + HiFi-GAN discriminators [29] Input: $F0^{1}$+ BN^{256} + x-$vectors^{512}$ Training criterion defined in Hifi-GAN [29]	speech waveform	LibriTTS: train-clean-100
6	NSF model	sinc-hn-NSF in [28] Input: $F0^{1}$ + Mel-$fbanks^{80}$ + x-$vectors^{512}$ STFT criterion	speech waveform	LibriTTS: train-clean-100

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.[13] The x-vectors for the speaker pool are extracted from a disjoint dataset (*LibriTTS-train-other-500*).

In *Step 3*, two modules are used to generate the speech waveform: a speech synthesis AM that generates Mel-filterbank features given the F0, the anonymized x-vector, and the BN features, and a neural source-filter (NSF) waveform model [28] that produces a speech waveform given the F0, the anonymized x-vector, and the generated Mel-filterbank outputs. Both models (#5.a and #6 in Table 4) are trained on the same corpus (*LibriTTS-train-clean-100*).

More details about the baseline recipe can be found in the provided scripts[11] and in [16, 30].

6.1.2 B1.b

The second baseline **B1.b**, shown in Figure 6, is based on the same idea as **B1.a** and has the same x-vector extractor and anonymization modules, as well as pitch extractor and ASR AM for linguistic feature extraction. The main difference between the two baselines is in the speech synthesis component (Step 3) of the anonymization system. While **B1.a** follows the traditional pipeline TTS approach and includes a speech synthesis AM and a separate waveform model, **B1.b** directly converts BN, F0, and x-vector features using an NSF model. It is unnecessary to generate Mel-filterbanks since BN-features already encode linguistic content. **B1.b** thus simplifies the system structure.

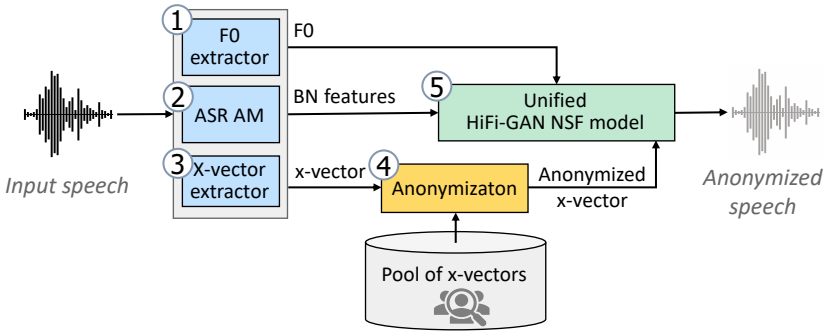

Figure 6: Second baseline anonymization system **B1.b**.

Another motivation for **B1.b** is to improve the quality of anonymized speech. Results from the VoicePrivacy 2020 Challenge [4] indicate that speech anonymized by **B1.a** is inferior to **B2** in terms of subjective quality. One possible reason is the over-smoothing effect caused by the maximum-likelihood-based NSF training criterion [28]. One solution is to adopt a generative adversarial network (GAN)-based framework. As shown in Figure 6, **B1.b** combines the NSF model (as the generator) with the discriminators of HiFi-GAN [29] and trains the model in the same manner as HiFi-GAN. **B1.b** is trained using the same data as **B1.a**, namely *LibriTTS-train-clean-100*. After training, the discriminators can be safely discarded, and only the trained NSF is used in the anonymization system. The implementation of the TTS modules is based on Pytorch [31].

6.2 Anonymization using the McAdams coefficient: B2

In contrast to **B1.a** and **B1.b**, the third baseline **B2** shown in Figure 7 does not require any training data and is based upon simple signal processing techniques. It employs the McAdams coefficient [32] to achieve anonymisation by shifting the pole positions derived from linear predictive coding (LPC) analysis of speech signals.

It 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 value of α implies a contraction or

[13] In the baselines **B1.a** and **B1.b**, the following parameter values are used: $N = 200$ and $N^* = 100$.

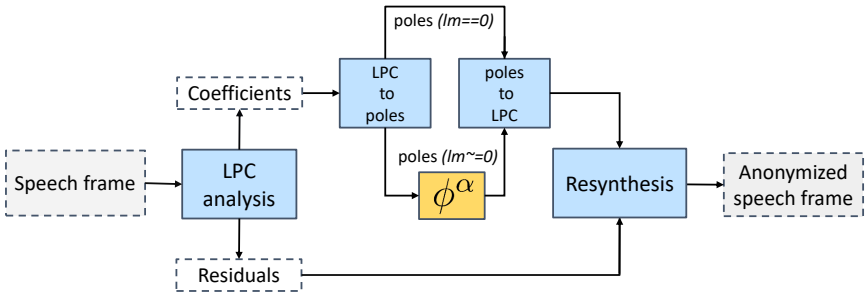

Figure 7: Third baseline anonymization system **B2**.

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 [33]. 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.

The baseline **B2** is a randomized version of the anonymisation algorithm proposed in [34], where the McAdams coefficient is sampled for each source speaker in the evaluation set from a uniform distribution: $\alpha \sim U(\alpha_{\min}, \alpha_{\max})$.[14] We use $\alpha_{\min} = 0.5$ and $\alpha_{\max} = 0.9$.

6.3 Results

Results for the three baselines are reported in Tables 5 and 6 in terms of the privacy (EER) and primary utility (WER) metrics, and in Table 7 in terms of the two secondary utility metrics. EER and WER results for the three baseline systems are also depicted in Figure 1. **B1.b** has better WER than the other baseline systems, while **B1.a** demonstrates the highest average EER.

6.4 Alternative anonymization systems

In addition to the proposed three baseline systems described above, several alternative solutions have been developed by the the challenge organizers and proposed as sources of additional inspiration for challenge participants.[15]

[14] This is different from the **B2** baseline of the 2020 challenge for which we used a constant value of $\alpha = 0.8$.
[15] Source code is available from https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2022.

Table 5: Primary privacy evaluation: EER,% achieved by $ASV_{\text{eval}}^{\text{anon}}$ on data processed by **B1.a**, **B1.b**, or **B2** vs. EER achieved by ASV_{eval} on the original (Orig.) unprocessed data.

Dataset	**Gender**	**Weight**	**EER,%**			
			Orig.	**B1.a**	**B1.b**	**B2**
LibriSpeech-dev	female	0.25	8.67	17.76	19.03	11.36
	male	0.25	1.24	6.37	5.59	1.40
VCTK-dev (different)	female	0.20	2.86	12.46	8.25	6.68
	male	0.20	1.44	9.33	6.01	6.35
VCTK-dev (common)	female	0.05	2.62	13.95	9.01	5.81
	male	0.05	1.43	13.11	9.40	8.83
Weighted average dev			3.54	11.74	9.93	6.53
LibriSpeech-test	female	0.25	7.66	12.04	9.49	7.12
	male	0.25	1.11	8.91	7.80	1.11
VCTK-test (different)	female	0.20	4.89	16.00	10.91	16.92
	male	0.20	2.07	10.05	7.52	7.69
VCTK-test (common)	female	0.05	2.89	17.34	15.32	10.98
	male	0.05	1.13	9.89	8.19	4.80
Weighted average test			3.79	11.81	9.18	7.77

Table 6: Primary utility evaluation: WER,% achieved by $ASR^{\text{anon}}_{\text{eval}}$ on data processed by **B1.a**, **B1.b**, or **B2** vs. WER achieved by ASR_{eval} on the original (Orig.) unprocessed data.

Dataset	**WER,%**			
	Orig.	**B1.a**	**B1.b**	**B2**
LibriSpeech-dev	3.82	4.34	4.19	4.32
VCTK-dev	10.79	11.54	10.98	11.76
Average dev	7.31	7.94	7.59	8.04
LibriSpeech-test	4.15	4.75	4.43	4.58
VCTK-test	12.82	11.82	10.69	13.48
Average test	8.49	8.29	7.56	9.03

Table 7: Secondary utility evaluation: pitch correlation ρ^{F_0} and gain of voice distinctiveness G_{VD} achieved on data processed by **B1.a**, **B1.b**, or **B2**.

Dataset	**Gender**	**Weight**	$\boldsymbol{\rho^{F_0}}$			$\boldsymbol{G_{\text{VD}}}$		
			B1.a	**B1.b**	**B2**	**B1.a**	**B1.b**	**B2**
LibriSpeech-dev	female	0.25	0.77	0.84	0.64	-9.15	-4.92	-1.94
	male	0.25	0.73	0.76	0.53	-8.94	-6.38	-1.65
VCTK-dev (different)	female	0.20	0.84	0.87	0.70	-8.82	-5.94	-1.32
	male	0.20	0.78	0.76	0.59	-12.61	-9.38	-2.18
VCTK-dev (common)	female	0.05	0.79	0.84	0.64	-7.56	-4.17	-1.14
	male	0.05	0.72	0.72	0.54	-10.37	-6.99	-1.32
Weighted average dev			0.77	0.80	0.61	-9.71	-6.44	-1.72
LibriSpeech-test	female	0.25	0.77	0.85	0.61	-10.04	-5.00	-1.71
	male	0.25	0.69	0.72	0.54	-9.01	-6.64	-1.74
VCTK-test (different)	female	0.20	0.84	0.87	0.68	-10.29	-6.09	-1.56
	male	0.20	0.79	0.77	0.66	-11.69	-8.64	-1.56
VCTK-test (common)	female	0.05	0.79	0.85	0.65	-9.31	-5.10	-1.59
	male	0.05	0.70	0.71	0.61	-10.43	-6.50	-1.36
Weighted average test			0.77	0.80	0.62	-10.15	-6.44	-1.63

6.4.1 Alternative x-vector extractor

The first alternative is based on the Sidekit toolkit [35].[16] It replaces the Kaldi-based x-vector extractor with a Pytorch-based implementation. This alternative system allows more straightforward modification of the extractor and adds more recent speaker verification loss functions, such as the additive angular margin loss [36]. X-vector representations are necessary to anonymize speech in the **B1** baselines. The provided x-vector extractor model is based on a ResNet-34 network and has an x-vector embedding size of 256.

Anonymization systems developed using the provided models will not be considered for evaluation and ranking in the challenge since they rely on additional data used for data augmentation[17] besides those specified in Section 4. Nevertheless, participants may experiment with these models and report the results in their challenge papers.

6.4.2 Alternative speech synthesis models

The second alternative includes:[18]

- `am_nsf_pytorch`: A Pytorch-based re-implementation of **B1.a**. All the scripts and codes are unified under a Pytorch-based project. This is expected to be easier to customize and revise than the C++/CUDA-based **B1.a**;

[16] The code is available in the sidekit branch.

[17] Room impulse response and noise database: `http://www.openslr.org/resources/28/`; MUSAN corpus of music, speech, and noise recordings: `http://www.openslr.org/resources/17/`.

[18] The code is available in the master branch and options for different speech synthesis models are setup by parameter `tts_type`.

- `joint_hifigan`: A variant of **B1.a**. The only difference is that the HiFi-GAN NSF model is replaced with the original HiFi-GAN.

6.4.3 Using self-supervised learning models

The third alternative is based on self-supervised learning.[19] Compared to **B1.b**, it has two main differences: 1) The BN features are replaced by the representation obtained from the last layer of a fine-tuned wav2vec 2.0 model [37]. Specifically, a wav2vec 2.0 Base model released by Facebook Research[20] was trained with 10k hours of unlabeled cross-lingual speech [38] and finetuned on the labeled *LibriSpeech-train-clean-100* data. The fine-tuning was conducted using the Fairseq toolkit[21] with default settings; 2) HiFi-GAN is used as the speech waveform generation model.[22] The other settings remain the same as in **B1.b**.

Anonymization systems developed using the provided SSL models will not be considered for evaluation and ranking in the challenge since they rely on additional data besides those specified in Section 4. Nevertheless, participants may experiment with these or other self-supervised models [39, 40] and report the results in their challenge papers.

7 Evaluation rules

- Participants are free to develop their own anonymization systems, using components of the baselines or not. They are strongly encouraged to make multiple submissions corresponding to different EER thresholds (see Section 5.1.1). Thresholds are applied to the weighted average of EER across the VoicePrivacy test datasets (with weights of 0.5, 0.1 and 0.4 for *LibriSpeech-test-clean*, *VCTK-test (common)* and *VCTK-test (different)*, respectively).

- The primary metrics (EER, WER) will be used for system ranking. Within each interval for EER – [15,20), [20,25), [25,30), [30,100) – systems will be ranked in order of increasing WER (averaged over *LibriSpeech-test-clean* and *VCTK-test*). All submissions considered for ranking should achieve a minimum average pitch correlation of $\rho^{F_0} > 0.3$ for each development and test dataset.

- Participants can use only the training and development datasets specified in Section 4 in order to train their system and tune hyperparameters. The use of any additional speech data is strictly prohibited.

- Participants must anonymize the development and test data in a *speaker-level* manner. All enrollment (resp. trial) utterances from a given speaker must be converted into the same pseudo-speaker, and enrollment (resp. trial) utterances from different speakers must be converted into different pseudo-speakers. Also, the pseudo-speaker corresponding to a given speaker in the enrollment set must be different from the pseudo-speaker corresponding to that same speaker in the trial set.

- Participants must anonymize the dataset (*LibriSpeech-train-clean-360*) used for training of the evaluation models $ASV_{\text{eval}}^{\text{anon}}$ and $ASR_{\text{eval}}^{\text{anon}}$ using the same anonymization system applied to the development and test data, albeit in an *utterance-level* manner. They must then train the evaluation models on the anonymized training data and apply them to the anonymized development and test data using the provided scripts. Modifications to the training or evaluation recipes (e.g., changing the network architecture, the hyperparameters, etc.) are prohibited.

- For every submitted system, participants must compute the primary objective metrics (WER, EER) and the secondary objective metrics (ρ^{F_0}, G_{VD}) for the two development datasets and the two test datasets using the provided evaluation scripts and retrained evaluation models. The organizers will be responsible for subjective evaluation only.

[19] The code is available in the ssl_anon_v2 branch.
[20] https://github.com/facebookresearch/voxpopuli
[21] https://github.com/pytorch/fairseq/
[22] The fine-tuned wav2vec 2.0 model and HiFi-GAN models are available at https://zenodo.org/record/6350122/files/ssl_models.tar.gz

8 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 will ask all participants to share with us the anonymized speech data obtained when running their anonymization system on the training, development and test datasets. Further details will follow in due course.

9 Registration and submission of results

9.1 General mailing list

All participants and team members are encouraged to subscribe to the general mailing list. Subscription can be done by sending an email to:

sympa@lists.voiceprivacychallenge.org

with *'subscribe 2022'* as the subject line. Successful subscriptions are confirmed by return email. To post messages to the mailing list itself, emails should be addressed to:

2022@lists.voiceprivacychallenge.org

9.2 Registration

Participants/teams are requested to register for the evaluation. Registration should be performed **once only** for each participating entity and by sending an email to:

organisers@lists.voiceprivacychallenge.org

with *'VoicePrivacy 2022 registration'* as the subject line. The mail body should include: (i) the name of the team; (ii) the name of the contact person; (iii) their affiliation; (iv) their country; (v) their status (academic/non-academic).

9.3 Submission of results

Each participant may submit as many systems as they wish for each EER threshold provided in Section 5.1.1. In the case of multiple submissions for each condition, the organisers will use the single system with the lowest WER for ranking. Participants should submit audio data for only a single system per condition. Audio data for this system will be used for subjective evaluation.

Each single submission should include:

1. The *results* files generated by the evaluation scripts, which contain EER, WER, ρ^{F_0} and G_{VD} estimates for both development and test datasets,[23] along with the full contents of the two *results* directories `exp/results-<date>-<time>` and `exp/results-<date>-<time>.orig` also generated by the evaluation scripts;

2. The corresponding PLDA (LLR) scores in Kaldi format (for the development and test data) obtained with the provided scripts;

[23] Example *results* files for the baseline system **B1.b**:

- Primary and secondary metrics: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2022/blob/master/baseline/results/RESULTS_summary_tts_joint_nsf_hifigan (saved in exp/results-<date>-<time>/results_summary.txt)
- Additional metrics obtained using ASV_{eval} and ASR_{eval}: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2022/blob/master/baseline/results/results.orig_tts_joint_nsf_hifigan (saved in exp/results-<date>-<time>.orig/results.txt)
- Additional metrics obtained using $ASV_{\text{eval}}^{\text{anon}}$ and $ASR_{\text{eval}}^{\text{anon}}$: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2022/blob/master/baseline/results/results.anon_tts_joint_nsf_hifigan (saved in exp/results-<date>-<time>/results.txt)

3. The corresponding anonymized speech data (wav files, 16 kHz, with the same names as in the original corpus) generated from the development and test datasets. 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.

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 Challenge website.

10 VoicePrivacy Challenge workshop at INTERSPEECH 2022

The VoicePrivacy 2022 Challenge will culminate in a joint workshop held in Incheon, Korea in conjunction with **INTERSPEECH 2022** and in cooperation with the ISCA SPSC Symposium.[1] VoicePrivacy 2022 Challenge participants are encouraged to submit papers on the topic of their challenge entry according to the paper submission schedule (see Section 11). Paper submissions must conform to the format of the ISCA SPSC Symposium proceedings, detailed in the author's kit[24], 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.

11 Schedule

The result submission deadline is 31st July 2022. 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 2022.

Table 8: Important dates

Release of training, development and evaluation data	Done
Release of evaluation software and baselines	19th March 2022
Submission of challenge papers to the joint SPSC Symposium and VoicePrivacy Challenge workshop	15th June 2022
Author notification for challenge papers	1st July 2022
Early bird registration to the joint SPSC Symposium and VoicePrivacy Challenge workshop	7th July 2022
Deadline for participants to submit objective evaluation results, anonymized data, and system descriptions	31st July 2022
Final paper upload	5th September 2022
Joint SPSC Symposium and VoicePrivacy Challenge workshop	23rd–24th September 2022

[24] https://interspeech2022.org/files/IS2022_paper_kit.zip

12 Acknowledgement

This work was supported in part by the French National Research Agency under project DEEP-PRIVACY (ANR-18-CE23-0018) and jointly by the French National Research Agency and the Japan Science and Technology Agency under project VoicePersonae.

References

[1] A. Nautsch, C. Jasserand, E. Kindt, M. Todisco, I. Trancoso, and N. Evans, "The GDPR & speech data: Reflections of legal and technology communities, first steps towards a common understanding," in *Interspeech*, 2019, pp. 3695–3699.

[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, 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

[4] 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

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

[6] J.-F. Bonastre, H. Delgado, N. Evans, T. Kinnunen, K. A. Lee, X. Liu, A. Nautsch, P.-G. Noe, J. Patino, M. Sahidullah *et al.*, "Benchmarking and challenges in security and privacy for voice biometrics," *arXiv preprint arXiv:2109.00281*, 2021.

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

[8] 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.

[9] 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/docs/VoicePrivacy_2020_Eval_Plan_v1_3.pdf, 2020.

[10] 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.

[11] 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.

[12] 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.

[13] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in *Interspeech*, 2017, pp. 2616–2620.

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

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

[16] 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," 2021. [Online]. Available: https://hal.inria.fr/hal-03197376/file/design_choices_informed.pdf

[17] M. Maouche, B. M. L. Srivastava, N. Vauquier, A. Bellet, M. Tommasi, and E. Vincent, "Enhancing speech privacy with slicing," 2021. [Online]. Available: https://hal.inria.fr/hal-03369137

[18] A. S. Shamsabadi, B. M. L. Srivastava, A. Bellet, N. Vauquier, E. Vincent, M. Maouche, M. Tommasi, and N. Papernot, "Differentially private speaker anonymization," *arXiv preprint arXiv:2202.11823*, 2022.

[19] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans *et al.*, "Post-evaluation analysis for the VoicePrivacy 2020 challenge: Using anonymized speech data to train attack models and ASR," https://www.voiceprivacychallenge.org/docs/VoicePrivacy2020_post_evaluation.pdf, 2020.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel *et al.*, "The Kaldi speech recognition toolkit," 2011.

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

[22] 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.

[23] D. Hirst, "A Praat plugin for momel and intsint with improved algorithms for modelling and coding intonation. icphs xvi, saabrücken," 2007.

[24] P.-G. Noé, J.-F. Bonastre, D. Matrouf, N. Tomashenko, A. Nautsch, and N. Evans, "Speech pseudonymisation assessment using voice similarity matrices," in *Interspeech*, 2020, pp. 1718–1722.

[25] P.-G. Noé, A. Nautsch, N. Evans, J. Patino, J.-F. Bonastre, N. Tomashenko, and D. Matrouf, "Towards a unified assessment framework of speech pseudonymisation," *Computer Speech & Language*, vol. 72, p. 101299, 2022.

[26] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," National Inst of Standards and Technology Gaithersburg MD, Tech. Rep., 1997.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, H. Wallach *et al.*, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

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

[33] 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.

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

[35] A. Larcher, K. A. Lee, and S. Meignier, "An extensible speaker identification sidekit in python," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5095–5099.

[36] J. Deng, J. Guo, J. Yang, N. Xue, I. Cotsia, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition." *IEEE transactions on pattern analysis and machine intelligence*, vol. PP, 2021.

[37] 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.

[38] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 993–1003. [Online]. Available: https://aclanthology.org/2021.acl-long.80

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

[40] ——, "Analyzing language-independent speaker anonymization framework under unseen conditions," 2022. [Online]. Available: https://arxiv.org/abs/2203.14834