The VoicePrivacy 2020 Challenge Evaluation Plan

Version 1.2

Natalia Tomashenko[1], Brij Mohan Lal Srivastava[2], Xin Wang[3], Emmanuel Vincent[4], Andreas Nautsch[5], Junichi Yamagishi[3,6], Nicholas Evans[5], Jose Patino[5], Jean-François Bonastre[1], Paul-Gauthier Noé[1], and Massimiliano Todisco[5]

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

https://voiceprivacychallenge.org

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. This is because speech data contains much more than just the spoken words. Speech encapsulates a wealth of personal, personal 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. Since machines can now decipher spoken language with ever-impressive accuracy, there is no reason why political orientations, religious and philosophical beliefs could not also be derived from speech data.

Given that speech data can qualify as personal data, it is thus of no surprise that efforts to develop privacy preservation solutions for speech technology are starting to emerge. The study of privacy is also gaining traction; the Security and Privacy in Speech Communication (SPSC)[1] Special Interest Group (SIG) of the International Speech Communication Association (ISCA) has recently been formed to champion research in the domain.

There are arguably two general solutions to preserve privacy in speech data: cryptography, and anonymization, also referred to as de-identification. Cryptographic solutions can render speech data unreadable and can only be read in plain text with access to a private key. Some techniques, e.g.

[1] https://www.spsc-sig.org

homomorphic encryption, even support computation upon data in the encrypted domain. These methods are typically specific to the given application, are challenging to integrate within existing systems and almost always result in significant increases to computational complexity and/or communication overheads. As such, cryptographic solutions can be relatively inflexible and cumbersome. Even then, they can only be implemented with specialist cryptography expertise.

Anonymization is different. Instead of conserving and protecting sensitive data through encryption, anonymization techniques, e.g. [3], have the goal of suppressing personally identifiable information within a speech signal, leaving all other aspects of the speech signal intact. Anonymization techniques can be highly flexible and can be integrated easily within existing systems. They usually have low computational complexity and rarely require any additional expertise outside of the speech domain. Despite the appeal of anonymization techniques and the urgency to address privacy concerns, there are currently only few solutions. In addition, a formal definition of anonymization is missing, the level of anonymization offered by some solutions is somewhat unclear, and there are no common datasets, protocols and metrics to support further work.

The VoicePrivacy initiative is spearheading the effort to develop privacy preservation solutions for speech technology. It aims to gather a new community to define the task and metrics and to benchmark initial solutions using the first common datasets and protocols. In following best practice, VoicePrivacy takes the form of a competitive challenge. Challenge participants are required to process a dataset of speech signals in order to anonymize them, while protecting the linguistic content and speech naturalness. The challenge will run from early 2020 and conclude with a special session/event held in conjunction with Interspeech 2020, at which challenge results will be made publicly available. This document describes plans for the challenge, the dataset, 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 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 a common datasets, protocols and a suite of software packages that will enable them to gauge anonymization performance. While it is likely that this will evolve in future challenges, no formal ranking of systems will be made for this first edition. This is because recommendations for the assessment of anonymization solutions form one of the technical goals.

3 Task

A privacy preservation task is typically formulated as a game involving one or more *users* who publish some data and an *attacker* (also called an adversary) who gains legal or illegal access to this data or to derived data and attempts to infer personal information about the users. To protect their privacy, the users publish data that contain as little personal information as possible while allowing one or more desired goals to be fulfilled. To infer personal information about the users, the attacker may use additional knowledge.

Considering speech data as an example, the definition of the privacy preservation task depends on the following scenario specifications [4]:

- the nature of the data (e.g., speech waveform, speech features, text, etc.)
- the pieces of information considered as personal (e.g., speaker identity, speaker traits, spoken contents, etc.)
- the desired goal(s) (e.g., human dialogue, multi-party human conversation, automatic speech recognition (ASR) training, emotion recognition, etc.)
- the data accessed by the attacker (e.g., a single utterance, multiple utterances, features computed from these utterances, an ASR model trained on these utterances, etc.)
- the additional knowledge available to the attacker (e.g., utterances previously released by the user, the measures adopted by users to protect their privacy, etc.)

Different scenario specifications lead to different privacy preservation methods from the user's point of view and different attack methods from the attacker's point of view.

3.1 Scenario

In VoicePrivacy 2020, we consider the following specific scenario, where the terms "user" and "speaker" are used interchangeably:

- Speakers want to hide their identity while allowing any desired goal to be potentially achieved.
- The attacker has access to a single utterance and wants to identify the corresponding speaker.

3.2 Anonymization task

In order to hide his/her identity, each speaker passes his/her utterances through an anonymization system before publication. The resulting anonymized utterances are called *trial* utterances. They sound as if they were uttered by another speaker, which we call *pseudo-speaker* which may be an artificial voice not corresponding to any real speaker.

The task of challenge participants is to develop this anonymization system. In order to allow any desired goal to be potentially achieved, this system should satisfy the following requirements:

- output a speech waveform,
- hide speaker identity as much as possible,
- distort other speech characteristics as little as possible,

- ensure that all trial utterances from a given speaker appear to be uttered by the same pseudo-speaker, while trial utterances from different speakers appear to be uttered by different pseudo-speakers.

The third requirement will be assessed via a range of *utility* metrics (also called usability metrics). Specifically, ASR performance using a model trained on clean (non-anonymized) data and subjective speech intelligibility and naturalness will be measured during the challenge (see Section 5.1 and 5.2), and additional desired goals including ASR training will be assessed in a post-evaluation stage (see Section 8).

The fourth requirement is motivated by the fact that, in a multi-party human conversation, each speaker cannot change his/her anonymized voice over time and the anonymized voices of all speakers must be distinguishable from each other. This will be assessed via a new, specifically designed metric in the post-evaluation stage.

3.3 Attack model

We assume that the attacker has access to various amounts of data:

- one or more anonymized trial utterances,
- possibly, several additional utterances for each speaker, which may or may not have been anonymized and are called *enrollment* utterances.

The attacker does not have access to the anonymization system applied by the user though.

The level of protection of personal information will be assessed via a range of *privacy* metrics, including objective speaker verifiability metrics and subjective speaker verifiability and linkability metrics. 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 enrollment utterances. During the challenge, two sets of metrics will be computed, corresponding to the two situations when the enrollment utterances are clean or they have been anonymized (see Section 5.1). In the latter case, we assume that the utterances have been anonymized in the same way as the trial data using the same anonymization system, i.e., all enrollment utterances from a given speaker are converted into the same pseudo-speaker, and enrollment utterances from different speakers are converted into different pseudo-speakers. We also assume that the pseudo-speaker corresponding to a given speaker in the enrollment set is different from the pseudo-speaker corresponding to that same speaker in the trial set. In the post-evaluation stage, we will consider alternative anonymization procedures for the enrollement data, corresponding to stronger attack models (see Section 8).
- The subjective speaker veriability metric assumes that the attacker has access to a single anonymized trial utterance and a single clean enrollment utterance (see Section 5.2).
- The subjective speaker linkability metric assumes that the attacker has access to a several anonymized trial utterances and several clean enrollment utterances (see Section 5.2).

4 Data

Several publicly available corpora are used for training, development and evaluation of speaker anonymization systems. They are comprised of subsets from the following corpora:

- ***LibriSpeech***[2] [5] is a corpus of read English speech derived from audiobooks and designed for ASR research. It contains about 1,000 hours of speech sampled at 16 kHz.
- ***LibriTTS***[3] [6] is a corpus of English speech derived from the original LibriSpeech corpus and designed for text-to-speech (TTS). It contains approximately 585 hours of read English speech sampled at 24 kHz.
- ***VCTK***[4] [7] is a corpus of read speech of 109 native speakers of English with various accents. It was originally aimed for TTS synthesis systems and contains about 44 hours of speech sampled at 48 kHz.
- ***VoxCeleb-1,2***[5] [8, 9] is an audiovisual corpus extracted from videos uploaded to YouTube and designed for speaker verification research. It contains about 2,770 hours of speech (16 kHz) from about 7,360 speakers, covering a wide range of accents and languages.

The detailed description of datasets provided for training, development and evaluation is given below.

4.1 Training data

For training an anonymization system the following corpora can be used: *VoxCeleb-1,2*, *LibriSpeech-train-clean-100*, *LibriSpeech-train-other-500*, *LibriTTS-train-clean-100* and *LibriTTS-train-other-500*. Summary statistics for these data are given in Table 1.

Table 1: Statistics of the **training** datasets.

Subset	**Size,h**	**Number of Speakers**			**Number of Utterances**
		Female	**Male**	**Total**	
VoxCeleb-1,2	2,794	2,912	4,451	7,363	1,281,762
LibriSpeech: train-clean-100	100	125	126	251	28,539
LibriSpeech: train-other-500	497	564	602	1,166	148,688
LibriTTS: train-clean-100	54	123	124	247	33,236
LibriTTS: train-other-500	310	560	600	1,160	205,044

4.2 Development set

Subsets from two different corpora are provided as development sets for anonymization systems. The first one is a subset of the *LibriSpeech-dev-clean* dataset. The second one (denoted as *VCTK-dev*) is obtained from the VCTK corpus. As explained in Section 3, we split these datasets into a trial subset, which is to be anonymized, and an enrollment subset, which is to be anonymized with different pseudo-speakers in order to compute objective speaker verification metrics. The anonymized enrollment data shall be stored in a different folder than the clean enrollment data, since both clean and anonymized enrollment data will be used to computed two different sets of metrics.

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

The statistics of these subsets are summarized in Table 2. For the *LibriSpeech-dev-clean* dataset, the speakers in the enrollment set are a subset of those in the trial set. For the *VCTK-dev* dataset, we created two subsets of trial utterances, denoted as *common part* and *different part*. Both include trials from the same set of speakers, but from disjoint subsets of utterances. The *common part* of the trials is composed of utterances $\#1 - 24$ in the VCTK corpus that are identical for all speakers: the *elicitation paragraph*[6] (utterances $\#1 - 5$) and *rainbow passage*[7] (utterances $\#6 - 24$). This part is meant for subjective evaluation of speaker verifiability/linkability in a text-dependent manner. The enrollment subset and the *different part* of the trials are composed of distinct utterances for all speakers (utterances with indexes ≥ 25).

Table 2: Statistics of the **development** datasets.

Subset		**Female**	**Male**	**Total**
Librispeech: dev-clean	Speakers in enrollment	15	14	29
	Speakers in trials	20	20	40
	Enrollment utterances	167	176	343
	Trial utterances	1,018	960	1,978
VCTK-dev	Speakers (same in enrollment and trials)	15	15	30
	Enrollment utterances	300	300	600
	Trial utterances (common part)	344	351	695
	Trial utterances (different part)	5,422	5,255	10,677

4.3 Evaluation data

Similarly to the development data, subsets from two different corpora are used for evaluation. The first one is a subset of the *LibriSpeech-test-clean* dataset. The second one (denoted as *VCTK-test*) is obtained from the VCTK corpus. We split those datasets into enrollment and trial subsets as summarized in Table 3. For the VCTK dataset, we created two subsets of trial utterances, denoted as *common part* and *different part*, in a similar manner as for the development set.

Table 3: Statistics of the **evaluation** datasets.

Subset		**Female**	**Male**	**Total**
Librispeech: test-clean	Speakers in enrollment	16	13	29
	Speakers in trials	20	20	40
	Enrollment utterances	254	184	438
	Trial utterances	734	762	1496
VCTK-test	Speakers (same in enrollment and trials)	15	15	30
	Enrollment utterances	300	300	600
	Trial utterances (common part)	346	354	700
	Trial utterances (different part)	5,328	5,420	10,748

[6]Elicitation paragraph: http://accent.gmu.edu/pdfs/elicitation.pdf
[7]Rainbow passage: https://www.dialectsarchive.com/the-rainbow-passage

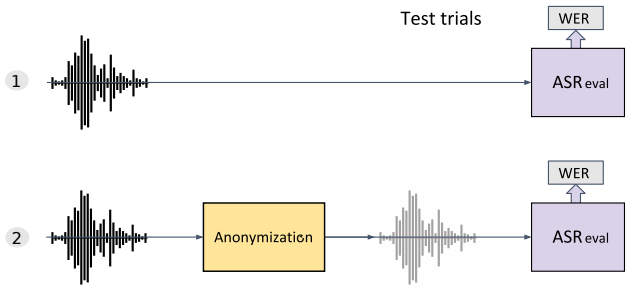

Figure 1: ASR evaluation for (1) clean speech data and (2) anonymized speech data. WER is estimated on the *trial* utterances of the development and evaluation datasets.

5 Utility and privacy metrics

Following the attack model assumed in Section 3.3, we consider objective and subjective privacy metrics to assess speaker re-identification and linkability. We also propose objective and subjective utility metrics in order to assess the fulfillment of the user goals specified in Section 3.2. Specifically, we consider ASR performance using a model trained on clean data and subjective speech intelligibility and naturalness.

Objective metrics will be computed by the participants themselves using the provided evaluation scripts and evaluation data, while subjective metrics will be computed the organizers using data provided by the participants.

5.1 Objective metrics

For objective evaluation of anonymization performance, two systems will be trained to assess the following characteristics: (1) speaker verifiability and (2) ability of the anonymization system to preserve linguistic information in the anonymized speech. The first system denoted as ASV_{eval} is an automatic speaker verification (ASV) system, which produces log-likelihood ratio (LLR) scores. The second system denoted as ASR_{eval} is an automatic speech recognition (ASR) system which outputs a word error rate (WER) metric.

Both ASR_{eval} and ASV_{eval} are trained on the *LibriSpeech-train-clean-360* dataset using the Kaldi speech recognition toolkit [10]. The statistics of this dataset are given in Table 4. Training of ASR_{eval} and ASV_{eval} and evaluation will be done with the provided recipes[8].

Table 4: Statistics of the training dataset for the ASV_{eval} and ASR_{eval} 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

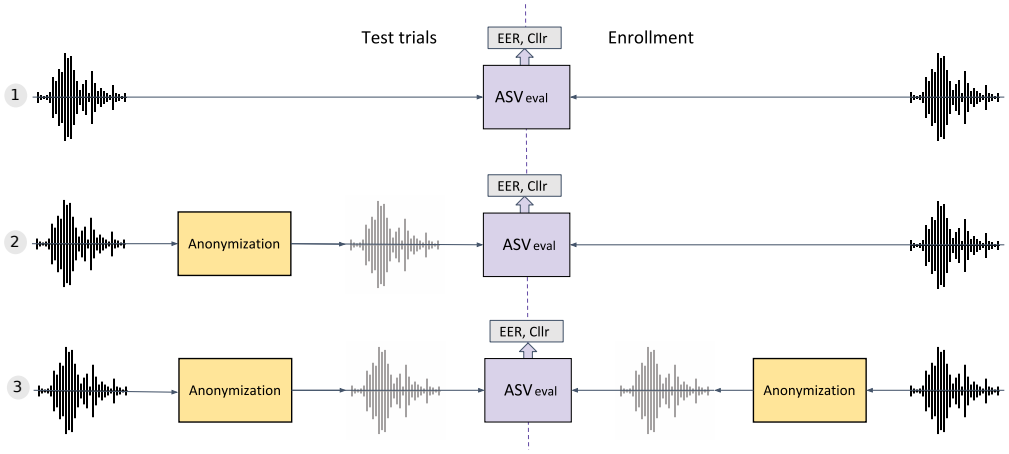

Figure 2: ASV evaluation for (1) clean trial and enrollment data; (2) anomymized trial data and clean enrollment data; and (3) anomymized trial and enrollment data.

5.1.1 Word error rate (WER)

ASR performance will be assessed using ASR_{eval} which is based on the state-of-the-art Kaldi recipe for LibriSpeech involving a TDNN-F acoustic model and a trigram language model. The recipe has been adapted to run on the *LibriSpeech-train-clean-360* dataset rather than the full LibriSpeech training corpus in order to speed up the evaluation process.

As shown in Figure 1, the anonymized development and evaluation data will be decoded using the provided pretrained system ASR_{eval} and the word error rate (WER) will be calculated:

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

with N_{sub}, N_{del}, and N_{ins}, the number of substitution, deletion, and insertion errors, respectively, and N_{ref} the number of words in the reference. This will be compared to the WER achieved on clean data. The WER is calculated only on the trial part of the development and evaluation datasets.

5.1.2 Speaker verifiability metrics

The ASV_{eval} system for speaker verification is based on state-of-the-art x-vector speaker embeddings with a probabilistic linear discriminant analysis (PLDA) backend [11] as implemented in Kaldi [10]. It will be applied as shown in Figure 2:

1. Compute PLDA (LLR) scores for (a) clean enrollment data and (b) anonymized trial data;
2. Compute PLDA (LLR) scores for (a) anonymized enrollment data and (b) anonymized trial data;

[8]Evaluation scripts: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2020

3. For steps 1 and 2, estimate the corresponding speaker verifiability metrics: equal error rate (EER) and log-likelihood-ratio cost function (C_{llr}).

The number of target and impostor trials in the development and evaluation datasets are given in Tables 5 and 6, respectively.

Table 5: Number of trials in the **development** datasets.

Subset	**Trials**	**Female**	**Male**	**Total**
Librispeech: dev-clean	Target	704	644	1,348
	Impostor	14,566	12,796	27,362
VCTK-dev	Target (common part)	344	351	695
	Target (different part)	1,781	2,015	3,796
	Impostor (common part)	4,810	4,911	9,721
	Impostor (different part)	13,219	12,985	26,204

Table 6: Number of trials in the **evaluation** datasets.

Subset	**Trials**	**Female**	**Male**	**Total**
Librispeech: test-clean	Target	548	449	997
	Impostor	11,196	9,457	20,653
VCTK-test	Target (common part)	346	354	700
	Target (different part)	1,944	1,742	3,686
	Impostor (common part)	4,838	4,952	9,790
	Impostor (different part)	13,056	13,258	26,314

Equal error rate (EER)

The EER is computed from the PLDA scores as follows. Let $P_{\text{fa}}(\theta)$ and $P_{\text{miss}}(\theta)$ denote the false alarm and miss rates at threshold θ:

$$P_{\text{fa}}(\theta) = \frac{\#\{\text{impostor trials with score} > \theta\}}{\#\{\text{total impostor trials}\}}$$
$$P_{\text{miss}}(\theta) = \frac{\#\{\text{target trials with score} \leq \theta\}}{\#\{\text{total target trials}\}},$$

so that $P_{\text{fa}}(\theta)$ and $P_{\text{miss}}(\theta)$ are, respectively, monotonically decreasing and increasing functions of θ. EER corresponds to the threshold θ_{EER} at which the two detection error rates are equal, i.e., $\text{EER} = P_{\text{fa}}(\theta_{\text{EER}}) = P_{\text{miss}}(\theta_{\text{EER}})$.

Log-likelihood-ratio cost function (C_{llr} and $C_{\text{llr}}^{\text{min}}$)

In addition to the EER, an alternative performance measure, the log-likelihood-ratio cost function (C_{llr}), will be estimated. It was proposed in [12] as an *application-independent*[9] evaluation objective and is defined as follows:

[9]In this context, *application* is defined as a set of prior probabilities and decision costs involved in the inferential process [12].

$$C_{\text{llr}} = \frac{1}{2}\left(\frac{1}{N_{\text{tar}}}\sum_{i\in\text{targets}}\log_2\left(1+e^{-\text{LLR}_i}\right) + \frac{1}{N_{\text{imp}}}\sum_{j\in\text{impostors}}\log_2\left(1+e^{\text{LLR}_j}\right)\right),$$

where N_{tar} and N_{imp} are respectively the number of target and impostor LLR values in the evaluation set.

C_{llr} can be decomposed into a discrimination loss ($C_{\text{llr}}^{\text{min}}$)[10] and a calibration loss ($C_{\text{llr}} - C_{\text{llr}}^{\text{min}}$)[11] [12]. The $C_{\text{llr}}^{\text{min}}$ is estimated by optimal calibration using monotonic transformation of scores to their empirical LLR values. To obtain this monotonic transformation, the pool adjacent violators (PAV) to LLR algorithm is used [12, 13].

5.2 Subjective metrics

Several subjective metrics will be used in evaluation: speaker verifiability, speaker linkability, speech intelligibility, and speech naturalness. They will be evaluated using listening tests that will be carried out by the organizers as described below.

5.2.1 Subjective speaker verifiability

To evaluate subjective speaker verifiability, the following two approaches can be used. The first approach is to measure the speaker similarity between one anonymized trial utterance and one clean enrollment utterance through a large-scale crowdsourced evaluation, following the approach used in a recent study on speech anti-spoofing [14]. Given one anonymized trial utterance and one randomly selected clean enrollment utterance from the same original speaker, the subjects will be instructed to imagine a scenario in which the anonymized sample is from an incoming telephone call, and the subjects need to judge whether the voice in the call is similar to the clean voice of the claimed speaker. The instruction given to the subjects is as follows:

> Imagine you are working for a bank call center. Your task is to compare customer inquiries with voices recorded when the same customer made inquiries in the past. You must determine whether the voices are of the same person or another person who is impersonating the original voices. However, if you choose 'spoofing by someone else' more than necessary, there will be many complaints from real customers, which should be avoided. Imagine a situation in which you are working to protect bank accounts and balance convenience. Now press the 'Sample A' and 'Sample B' buttons below and listen to the samples. You can listen to them as many times as you like. Use only the audio you hear to determine if the speakers are the same or not. The content of the conversation in English is irrelevant and does not need to be heard. Please judge on the basis of the characteristics of the voice, not the content of the words.

As the instruction describes, the subject will listen to two samples (i.e., Sample A and Sample B) in one evaluation page. Sample B is always a randomly selected enrollment utterance of a given original speaker. Sample A may be an anonymized trial utterance or a clean trial utterance of the same original speaker or a different speaker. The subject will be asked to evaluate the similarity

[10] How good are two classes separated for any threshold?

[11] In the light of domain shifts (e.g., changing speech signal quality between training and evaluation sets), scores loose the capability of correctly encoding the expected class ratio for what their value suggests; scores are not LLRs anymore and thus loose their goodness for ideal decision making. The calibration loss quantifies this gap; scores are considered to be LLRs if $C_{\text{llr}} \leq 1$, and for $C_{\text{llr}} = 1$, a coin toss provides the same information.

using a scale of 1 to 10, where 1 denotes 'different speakers' and 10 denotes 'the same speaker' with highest confidence. Note that, 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.

5.2.2 Subjective speaker linkability

The second subjective evaluation metric is based on speaker linkage, i.e., clustering. The listeners will be asked to place a set of anonymized trial utterances and clean enrollment utterances from different speakers in a 1- or 2-dimensional space according to speaker similarity. This will be done using a specially developed graphical interface, where each utterance is represented by a point in the space. The distance between two points expresses the subjective perception of speaker dissimilarity between the corresponding utterances.

5.2.3 Subjective speech intelligibility

Naturalness of the anonymized speech will also be evaluated through a large-scale crowdsourced evaluation. For one subjective evaluation round, the subject will listen to a sample and evaluate its naturalness using a scale from 1 (completely unintelligible) to 10 (completely intelligible and all content is clear). The sample can be an anonymized trial utterance or a clean enrollment utterance from a randomly selected speaker. The results can be visualized through DET curves for comprehensive comparison.

5.2.4 Subjective speech naturalness

Similar to the subjective speech intelligibility evaluation, naturalness of the anonymized speech will also be evaluated through a large-scale crowdsourced evaluation. The subject will listen to a sample and evaluate its naturalness using a scale from 1 (completely unnatural) to 10 (completely natural). The sample can be an anonymized trial utterance or a clean enrollment utterance from a randomly selected speaker. The performance of the participants can be visualized through DET curves for comprehensive comparison.

6 Baseline

Two different baseline systems have been developed for the challenge[12]: (1) anonymization using x-vectors and neural waveform models, and (2) anonymization using McAdams coefficient.

6.1 Baseline-1: Anonymization using x-vectors and neural waveform models

Our primary baseline system is based on the speaker anonymization method proposed in [3] and shown in Figure 3. The anonymization is performed in three steps:

- **Step 1: Feature extraction:** extract the speaker x-vector [11], the fundamental frequency (F0) and bottleneck (BN) features from the original audio waveform.

[12]Both baseline systems are available online: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2020

- **Step 2: X-vector anonymization:** anonymize the x-vector of the source speaker using an external pool of speakers.
- **Step 3: Speech synthesis:** synthesize the speech waveform from the anonymized x-vector and the original BN and F0 features using an acoustic model and a neural waveform model.

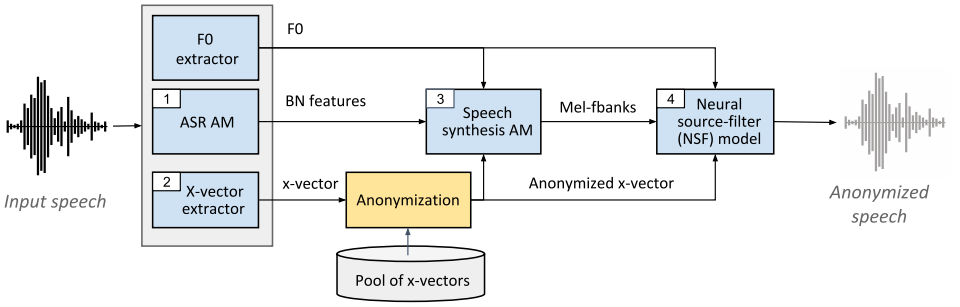

Figure 3: Baseline anonymization system.

Table 7: Baseline anonymization system: models and corpora. The model indexes are the same as in Figure 3. Superscript numbers represent feature dimensions.

#	Model	Description	Output features	Training dataset
1	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
2	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
3	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
4	NSF model	sinc1-h-NSF in [15] Input: F0^{1} + Mel-fbanks^{80} + x-vectors^{512} STFT criterion	speech waveform	LibriTTS: train-clean-100
Pool of speaker x-vectors				LibriTTS: train-other-500

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 7.

In *Step 1*, to extract BN features, an ASR acoustic model (AM) is trained (#1 in Table 1). We assume that these BN features represent the linguistic content of the speech signal. The ASR

AM has a factorized time delay neural network (TDNN-F) model architecture [16,17] and is trained using the Kaldi toolkit [10]. To encode speaker information, an x-vector extractor with a TDNN model topology (#2 in Table 1) 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 for which the similarity to the x-vector of the source speaker is in the given range. The cosine distance $\cos(x_1, x_2)$ or, optionally, probabilistic linear discriminant analysis (PLDA)) distance is used as a similarity measure between two x-vectors x_1 and x_2. The candidate x-vectors for averaging are chosen in two steps. First, for a given x-vector, N the most farthest candidates from the speaker pool are selected. Second, a smaller subset of N^* x-vector candidates from this set are chosen randomly[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 [15] that produces a speech waveform given the F0, the anonymized x-vector, and the generated Mel-filterbanks. Both models (#3 and #4 in Table 1) are trained on the same corpus (*LibriTTS-train-clean-100*).

More details about the baseline recipe can be found in the provided scripts.

Results for the ASV objective evaluation are provided in Table 8 for the development and evaluation datasets. Results for ASR evaluation are presented in Table 9 in terms of WER.

Table 8: ASV results for ***Baseline-1*** for development and test data (**o** – original, **a** – anonymized speech data for enrollment (**Enr**) and trial (**Trl**) parts; **Gen** denotes speaker gender: **f** – female, **m** – male).

#	Dev. set	EER,%	C_{llr}^{min}	C_{llr}	Enr	Trl	Gen	Test set	EER,%	C_{llr}^{min}	C_{llr}
1	libri_dev	8.665	0.304	42.857	o	o	f	libri_test	7.664	0.183	26.793
2	libri_dev	50.280	0.997	146.010	o	a	f	libri_test	48.540	0.996	151.374
3	libri_dev	35.090	0.876	15.194	a	a	f	libri_test	29.740	0.797	13.999
4	libri_dev	1.242	0.034	14.250	o	o	m	libri_test	1.114	0.041	15.303
5	libri_dev	58.390	0.998	168.501	o	a	m	libri_test	53.230	0.999	167.136
6	libri_dev	29.660	0.806	20.081	a	a	m	libri_test	32.520	0.835	26.539
7	vctk_dev_com	2.616	0.088	0.868	o	o	f	vctk_test_com	2.890	0.091	0.866
8	vctk_dev_com	49.420	0.995	165.440	o	a	f	vctk_test_com	50.000	0.995	157.810
9	vctk_dev_com	25.290	0.740	7.961	a	a	f	vctk_test_com	30.920	0.829	9.493
10	vctk_dev_com	1.425	0.050	1.559	o	o	m	vctk_test_com	1.130	0.036	1.041
11	vctk_dev_com	56.980	1.000	191.899	o	a	m	vctk_test_com	55.930	1.000	189.236
12	vctk_dev_com	27.640	0.735	18.506	a	a	m	vctk_test_com	22.030	0.663	14.055
13	vctk_dev_dif	2.864	0.100	1.134	o	o	f	vctk_test_dif	4.887	0.169	1.495
14	vctk_dev_dif	50.030	0.988	162.908	o	a	f	vctk_test_dif	48.870	0.999	142.398
15	vctk_dev_dif	29.480	0.814	10.237	a	a	f	vctk_test_dif	34.210	0.884	12.333
16	vctk_dev_dif	1.439	0.052	1.158	o	o	m	vctk_test_dif	2.067	0.072	1.817
17	vctk_dev_dif	55.330	1.000	166.495	o	a	m	vctk_test_dif	53.730	1.000	165.621
18	vctk_dev_dif	26.100	0.756	18.811	a	a	m	vctk_test_dif	25.830	0.743	16.305

[13] In the baseline, the following parameter values are used: $N = 200$ and $N^* = 100$; and PLDA was used as the distance between x-vectors.

Table 9: ASR results for ***Baseline-1*** in terms of WER,% for development and test data (**o**-original, **a**-anonymized speech) for two trigram language models (LMs): $\mathbf{LM}_s$ – small, and $\mathbf{LM}_l$ – large LM.

#	Dev set	WER, %		Data	Test set	WER, %	
		$\mathbf{LM}_s$	$\mathbf{LM}_l$			$\mathbf{LM}_s$	$\mathbf{LM}_l$
1	libri_dev	5.25	3.83	o	libri_test	5.55	4.14
2	libri_dev	9.15	6.50	a	libri_test	9.06	6.77
3	vctk_dev	14.00	10.79	o	vctk_test	16.39	12.81
4	vctk_dev	19.06	15.50	a	vctk_test	19.24	15.53

6.2 Baseline-2: Anonymization using McAdams coefficient

A secondary, alternative baseline, in contrast to the primary baseline, does not require any training data and is based upon simple signal processing techniques. It employs the McAdams coefficient [18] to achieve anonymisation by shifting the pole positions derived from linear predictive coding (LPC) analysis of speech signals.

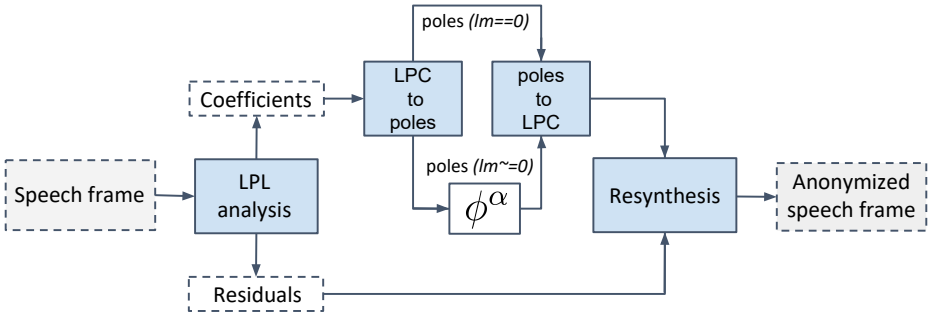

Figure 4: Pipeline of the application of the proposed McAdams coefficient-based approach to anonymisation on a speech frame basis. The angle ϕ of poles with a non-zero imaginary part are raised to the power of the McAdams coefficient α to provoke a shift in frequency of its associated formant.

The process is depicted in Figure 4. It starts with the application of frame-by-frame LPC source-filter analysis to derive the LPC coefficients and residuals for a speech signal. Residuals are set aside and retained for later resynthesis. LPC coefficients are converted into a set of pole positions. The McAdams transformation is then applied to the angles of each pole (with respect to the origin in the z-plane), each one of which corresponds to a peak in the spectrum (loosely resembling formant positions). While real-valued poles are left unmodified, the angles ϕ of the poles with a non-zero imaginary part (with values between 0 and π radians) are raised to the power of the McAdams coefficient α so that a transformed pole has new angle ϕ^α. For angles $\phi < 1$ (in radians) and $\alpha > 1$, ϕ^α results in a negative shift in angle, whereas for $\alpha < 1$, ϕ^α results in a positive shift. For $\phi > 1$, the shift is positive for $\alpha > 1$ and negative for $\alpha < 1$. The effect of such manipulation upon a set of arbitrary pole positions is illustrated in Figure 6 for values of $\alpha = \{0.9, 1.1\}$. The operation results in the contraction or expansion of the pole positions around $\phi = 1$. The effect upon the corresponding magnitude spectrum is illustrated in Figure 5. For a sampling rate of 16kHz, i.e. for data used in the challenge, $\phi = 1$ corresponds to approximately 2.5kHz which is the approximate mean formant position [19]. Corresponding complex conjugate poles are similarly and shifted in the opposite direction to their counterparts.

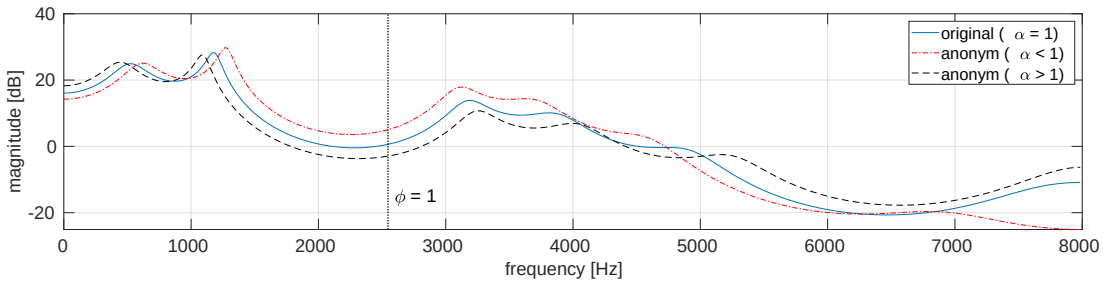

Figure 5: Example of the spectral envelope of a speech frame for both the original formants and the two anonymised versions.

The new, full set of poles, including original real-valued poles and shifted complex poles are then converted back to LPC coefficients. Finally, LPC coefficients and residuals are used to resynthesise a new speech frame in the time domain. Objective evaluation results for ASV and ASR are illustrated in Tables 10 and Table 11 respectively. This technique is similar in nature to the VoiceMask method [20] that was applied to preserve privacy in speech [4]. While results are inferior to those for the primary baseline, it is stressed that the algorithm has not been optimised in any way.

#	Dev. set	EER,%	C_{llr}^{min}	C_{llr}	Enr	Trl	Gen	Test set	EER,%	C_{llr}^{min}	C_{llr}
1	libri_dev	8.807	0.305	42.903	o	o	f	libri_test	7.664	0.184	26.808
2	libri_dev	35.370	0.821	116.892	o	a	f	libri_test	26.090	0.685	115.571
3	libri_dev	23.440	0.621	11.726	a	a	f	libri_test	15.330	0.490	12.553
4	libri_dev	1.242	0.035	14.294	o	o	m	libri_test	1.114	0.041	15.342
5	libri_dev	17.860	0.526	105.715	o	a	m	libri_test	17.820	0.498	106.434
6	libri_dev	10.560	0.359	11.951	a	a	m	libri_test	8.241	0.263	15.376
7	vctk_dev_com	2.616	0.088	0.869	o	o	f	vctk_test_com	2.890	0.092	0.861
8	vctk_dev_com	34.300	0.877	85.902	o	a	f	vctk_test_com	30.640	0.807	93.967
9	vctk_dev_com	11.630	0.366	43.551	a	a	f	vctk_test_com	14.450	0.465	42.734
10	vctk_dev_com	1.425	0.050	1.555	o	o	m	vctk_test_com	1.130	0.036	1.042
11	vctk_dev_com	23.930	0.669	90.757	o	a	m	vctk_test_com	24.290	0.713	99.336
12	vctk_dev_com	10.540	0.316	24.986	a	a	m	vctk_test_com	11.860	0.349	28.225
13	vctk_dev_dif	2.920	0.101	1.135	o	o	f	vctk_test_dif	4.938	0.169	1.492
14	vctk_dev_dif	35.540	0.907	90.540	o	a	f	vctk_test_dif	30.040	0.794	93.211
15	vctk_dev_dif	15.830	0.503	39.811	a	a	f	vctk_test_dif	16.920	0.546	41.341
16	vctk_dev_dif	1.439	0.052	1.155	o	o	m	vctk_test_dif	2.067	0.072	1.816
17	vctk_dev_dif	28.240	0.741	98.419	o	a	m	vctk_test_dif	28.240	0.720	101.704
18	vctk_dev_dif	11.220	0.384	23.093	a	a	m	vctk_test_dif	12.230	0.397	25.064

Table 10: ASV results for ***Baseline-2*** for development and test data (**o** – original, **a** – anonymized speech data).

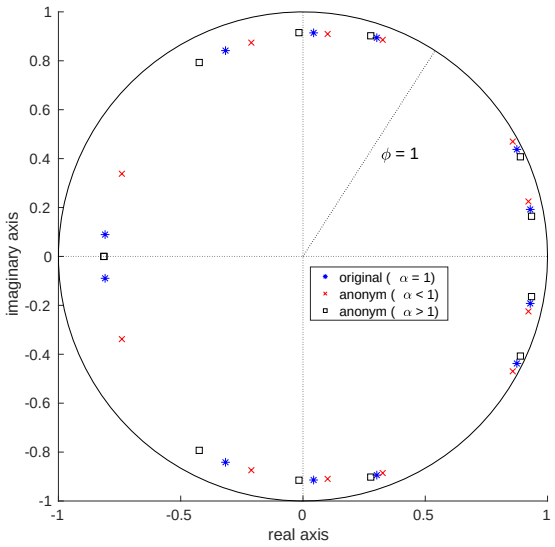

Figure 6: Example of pole-zero map as shown in Figure 5

#	Dev set	WER, %		Data	Test set	WER, %	
		$\mathbf{LM}_s$	$\mathbf{LM}_l$			$\mathbf{LM}_s$	$\mathbf{LM}_l$
1	libri_dev	5.24	3.84	o	libri_test	5.55	4.17
2	libri_dev	12.19	8.77	a	libri_test	11.77	8.88
3	vctk_dev	14.00	10.78	o	vctk_test	16.38	12.80
4	vctk_dev	30.10	25.56	a	vctk_test	33.25	28.22

Table 11: ASR results for ***Baseline-2*** in terms of WER,% for development and test data (**o**-original, **a**-anonymized speech) for two trigram LMs: $\mathbf{LM}_s$ – small, and $\mathbf{LM}_l$ – large LM.

7 Evaluation rules

- Participants are free to develop their own anonymization system, using parts of the baseline or not. They are also free to tune the compromise between speaker verifiability and speech quality/intelligibility according to their own preference.
- Participants can use only the training datasets and the enrollment datasets specified in Section 4.1 in order to train their system and tune its hyperparameters. Using additional speech data is not allowed.
- Speaker anonymization must be done in a speaker-to-speaker manner. All enrollment utterances from a given speaker must be converted into the same pseudo-speaker, and enrollment 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.
- Modifications of ASV_{eval} and ASR_{eval} (such as changing decoder parameters) are not allowed.
- For every tested system, participants should report all objective metrics (WER, EER, C_{llr} and

$C_{\text{llr}}^{\text{min}}$) on the two development datasets and the two evaluation datasets. Participants should compute these metrics themselves using the provided evaluation scripts. The organizers will be responsible for computing the subjective metrics and the post-evaluation metrics only.

- In the system description, participants should present their results in the same format (**without modification**) as used for Tables 8 and 9. In order to convert the result file generated by the evaluation system[14] to the latex tables in the required format, participants can use the following script: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2020/blob/master/baseline/local/results_to_latex.py.

8 Post-evaluation analysis

Following Deadline-2, the organizers will run additional evaluation experiments in order to further characterize the performance of the submitted systems and pave the way for the next Voice Privacy Challenge. To do so, we will ask volunteer participants to share with us the anonymized speech data obtained when running their anonymization system on the training/development datasets and on the evaluation dataset with different settings, and we will compute additional evaluation metrics on these data. The full details of the additional data to be provided by participants and the additional metrics which will be computed will be disclosed in due time.

Roughly speaking, this involves:

1. retraining the evaluation systems (ASR_{eval} and ASV_{eval}) on anonymized training data in order to evaluate the suitability of the proposed anonymization technique for ASR training;
2. computing subjective evaluation metrics (as described in Section 5.2) on selected subsets of evaluation data;
3. computing additional metrics to assess, e.g., the fulfillment of the goal that pseudo-speakers are sufficiently different from each other and the speaker verifiability performance in the presence of an attacker with additional knowledge.

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 2020'* as the subject line. Successful registrations are confirmed by return email. To post messages to the mailing list itself, emails should be addressed to:

2020@lists.voiceprivacychallenge.org

[14] Example: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2020/blob/master/baseline/RESULTS_baseline

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 2020 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 country; (iv) their status (academic/non-academic).

9.3 Submission of results

Each single submission should include:

1. Metric values for objective assessment (WER, EER, C_{llr} and $C_{\text{llr}}^{\text{min}}$) on the evaluation and development datasets according to the common protocol (participants should submit *result* file obtained by the provided evaluation scripts[15]).
2. Corresponding PLDA (LLR) scores in Kaldi format (for development and evaluation data) obtained with the provided scripts;
3. Corresponding anonymized speech data (wav files, 16kHz, with the same names as in the original corpus) generated from the evaluation and development datasets. These data will be used by the challenge organizers to verify the submitted scores, make post-evaluation analysis with other metrics and to run listening tests for subjective evaluation[16].

All the 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 shared among all registered participants, but will not be disclosed publicly. Further details concerning the submission procedure and URI will be published via the participants mailing list and via the VoicePrivacy Challenge website.

10 Special session at Interspeech 2020

A special session on the VoicePrivacy 2020 Challenge will be held at **Interspeech 2020** in Shanghai, China.

All participants are encouraged (but not obliged) to submit their papers, dedicated to early version of their challenge entry, to the special session of the conference by **8th May 2020** (which corresponds to the *Deadline-1* in the Challenge schedule, see Section 11). The review process for papers submitted to the special session is the same as for regular papers. Scientific novelty of the proposed ideas should be clearly stated and evaluated experimentally. Subjective evaluation results will be obtained only after *Deadline-2* and can be included (optionally) in the camera-ready version of the paper. Other participants may submit papers to the challenge only.

Accepted Interspeech papers and other Challenge submissions will both be presented at the special session in the form of posters. These two types of presentations will be clearly differentiated in the program.

[15] Example of the *result* file for the baseline system: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2020/blob/master/baseline/RESULTS_baseline

[16] Please note that subjective evaluation will be performed for *Deadline-2* submissions only, not for *Deadline-1*

11 Schedule

There will be two submission deadlines (early *Deadline-1* and late *Deadline-2*) in the Challenge. Participants may submit to one or both deadlines. Interspeech paper submission is encouraged but optional. All participants will be invited to present their work at the VoicePrivacy session/event.

Table 12: Important dates

Release of training and development data	8th February 2020
Release of evaluation data	15th February 2020
Deadline-1 for participants to submit objective evaluation results	8th May 2020
Interspeechi 2020 paper submission deadline	8th May 2020
Deadline-2 for participants to submit objective evaluation results and data	16th June 2020
Submission of system descriptions	23rd June 2020
Submission of additional data (optional - see Section 8)	1st July 2020
Organizers return subjective evaluation results to participants	Early/mid August 2020
VoicePrivacy special session/event at Interspeech 2020	26th–29th October 2020
Journal special issue submission deadline	Early 2021

12 Acknowledgement

This work was supported in part by the French National Research Agency under projects HARPOCRATES (ANR-19-DATA-0008) and DEEP-PRIVACY (ANR-18- CE23-0018), by the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement No. 825081 COMPRISE (https://www.compriseh2020.eu/), and jointly by the French National Research Agency and the Japan Science and Technology Agency under project VoicePersonae. The authors would like to thank Md Sahidullah and Fuming Fang.

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, M. Todisco, M. A. Hmani, M. A. Mtibaa, A. Abdelraheem, A. Abad, F. Teixeira, M. Gomez-Barrero, D. Petrovska, N. Chollet, G. Evans, T. Schneider, J.-F. Bonastre, B. Raj, I. Trancoso, and C. Busch, "Preserving privacy in speaker and speech characterisation," *Computer Speech & Language*, vol. 58, pp. 441–480, 2019.

[3] 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.

[4] 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.

[5] 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.

[6] 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," *arXiv preprint arXiv:1904.02882*, 2019.

[7] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.

[8] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[9] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[10] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," Tech. Rep., 2011.

[11] 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.

[12] N. Brümmer and J. Du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[13] D. Ramos and J. Gonzalez-Rodriguez, "Cross-entropy analysis of the information in forensic speaker recognition," in *Odyssey 2008: The Speaker and Language Recognition Workshop*, 2008.

[14] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "The ASVspoof 2019 database," *arXiv preprint arXiv:1911.01601*, 2019.

[15] 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.

[16] 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.

[17] 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.

[18] S. McAdams, "Spectral fusion, spectral parsing and the formation of the auditory image," *Ph. D. Thesis, Stanford*, 1984.

[19] S. Ghorshi, S. Vaseghi, and Q. Yan, "Cross-entropic comparison of formants of british, australian and american english accents," vol. 50, pp. 564–579, 2008.

[20] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, and X.-Y. Li, "Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity," in *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, 2018, pp. 82–94.