The VoicePrivacy 2020 Challenge: Results and findings

Natalia Tomashenko, Xin Wang, Emmanuel Vincent, Jose Patino, Brij Mohan Lal Srivastava, Paul-Gauthier Noé, Andreas Nautsch, Nicholas Evans, Junichi Yamagishi, Benjamin O'Brien, et al.

The VoicePrivacy 2020 Challenge: Results and findings

Natalia Tomashenko[a,*], Xin Wang[b], Emmanuel Vincent[c], Jose Patino[d], Brij Mohan Lal Srivastava[f], Paul-Gauthier Noé[a], Andreas Nautsch[d], Nicholas Evans[d], Junichi Yamagishi[b,e], Benjamin O'Brien[g], Anaïs Chanclu[a], Jean-François Bonastre[a], Massimiliano Todisco[d], Mohamed Maouche[f]

[a] *LIA, University of Avignon, Avignon, France*
[b] *National Institute of Informatics (NII), Tokyo, Japan*
[c] *Université de Lorraine, CNRS, Inria, LORIA, France*
[d] *Audio Security and Privacy Group, EURECOM, France*
[e] *University of Edinburgh, UK*
[f] *Inria, France*
[g] *LPL, Aix-Marseille University, France*

Abstract

This paper presents the results and analyses stemming from the first VoicePrivacy 2020 Challenge which focuses on developing anonymization solutions for speech technology. We provide a systematic overview of the challenge design with an analysis of submitted systems and evaluation results. In particular, we describe the voice anonymization task and datasets used for system development and evaluation. Also, we present different attack models and the associated objective and subjective evaluation metrics. We introduce two anonymization baselines and provide a summary description of the anonymization systems developed by the challenge participants. We report objective and subjective evaluation results for baseline and submitted systems. In addition, we present experimental results for alternative privacy metrics and attack models developed as a part of the post-evaluation analysis. Finally, we summarise our insights and observations that will influence the design of the next VoicePrivacy challenge edition and some directions for future voice anonymization research.

Keywords: privacy, anonymization, speech synthesis, voice conversion, speaker verification, automatic speech recognition, attack model, metrics, utility

1. Introduction

Due to the growing demand for privacy preservation in the recent years, privacy-preserving data processing has become an active research area. One reason for this is the European general data protection regulation (GDPR) in the European Union (EU) law and similar regulations in national laws of many

*Corresponding author
Email address: natalia.tomashenko@univ-avignon.fr (Natalia Tomashenko)

countries outside the EU concerning the implementation of the data protection principles when treating, transferring or storing personal data.

Although a legal definition of privacy is missing (Nautsch et al., 2019a), speech data contains a lot of personal information that can be disclosed by listening or by automated systems (Nautsch et al., 2019b). This includes, e.g., age, gender, ethnic origin, geographical background, health or emotional state, political orientations, and religious beliefs. Speaker recognition systems can also reveal the speaker's identity. Therefore, the increased interest in developing privacy preservation solutions for speech technology is not surprising. This motivated the launching of the VoicePrivacy initiative (Tomashenko et al., 2020b). This initiative aims to bring together a new community of researchers, engineers and privacy professionals in order to formulate the tasks of interest, develop evaluation methodologies, and benchmark new solutions through a series of challenges. The first VoicePrivacy challenge[1] was organized as a part of this initiative (Tomashenko et al., 2020b,a).

Existing approaches to privacy preservation for speech can be broadly classified into: obfuscation, encryption, distributed learning, or anonymization. Obfuscation methods (Cohen-Hadria et al., 2019; Gontier et al., 2020) suppress or modify the speech signal to the point where no information about it can be recovered. Encryption methods (Pathak et al., 2013; Brasser et al., 2018; Zhang et al., 2019) support computation upon data in the encrypted domain, however they significantly increase the computational complexity. Decentralized or federated learning methods learn models from distributed data without accessing it directly (Leroy et al., 2019), however the derived data used for learning (e.g., model gradients) may still leak information about the original data (Tomashenko et al., 2021a; Mdhaffar et al., 2021). Note also that the latter two categories of approaches are incompatible with using the data for supervised machine learning purposes, which requires third-party annotators to access the data in non-encrypted form.

Anonymization refers to the goal of suppressing personally identifiable information in the speech signal, leaving other attributes intact. In contrast to the above approaches, it allows the data to be used for supervised machine learning purposes and it can easily be integrated within existing systems. 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 other traits and states, words in the spoken contents, and sounds in the background which, when considered in combination with each other and possibly with external data, may reveal the speaker's identity.

As a first step towards this goal, the VoicePrivacy 2020 Challenge focuses on *voice anonymization*, that is the task of altering the speaker's voice to hide their identity to the greatest possible extent, while leaving all other speech attributes (traits, states, and spoken contents) intact. Approaches to voice anonymization

[1] `https://www.voiceprivacychallenge.org/`

include noise addition (Hashimoto et al., 2016), speech transformation (Qian et al., 2017; Patino et al., 2021), voice conversion (Fang et al., 2019; Han et al., 2020a; Srivastava et al., 2020a), and disentangled representation learning (Srivastava et al., 2019; Aloufi et al., 2020).

Despite the appeal of voice anonymization, the level of privacy protection offered by these solutions is unclear and not meaningful because there is no formal definition of the task and no formal attack model, and there are no common datasets, protocols and metrics. The VoicePrivacy 2020 Challenge aims to address all of these concerns.

The paper is structured as follows. The challenge design, including the description of the anonymization task, attack models, datasets, objective and subjective evaluation methodologies with the corresponding privacy and utility metrics, is presented in Section 2. The overview of the baseline and submitted systems is provided in Sections 3. Objective and subjective evaluation results and their comparison and analysis are presented in Section 4. We conclude and discuss future directions in Section 5.

2. Challenge design

In this section, we present an overview of the official challenge setup: anonymization task, corresponding attack models selected for the challenge, data and evaluation methodology. Also we present an additional attack model developed as part of the post-evaluation analysis (Tomashenko et al., 2020c).

2.1. Anonymization task and attack models

Privacy preservation is formulated as a game between *users* who share[2] some data and *attackers* who access this data or data derived from it and wish to infer information about the users (Qian et al., 2018b; Srivastava et al., 2020b; Tomashenko et al., 2020b). To protect their privacy, the users share data that contain as little personal information as possible while 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, spoken contents, 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 model, etc., (v) the attackers' prior knowledge: previously shared data, privacy preservation method applied, etc. Different specifications lead to different privacy preservation methods from the users' point of view and different attacks from the attackers' point of view.

[2]This data may be shared with selected individuals, with a company providing a service, with a public cloud provider, with the general public (open data), etc.. Attackers may include employees or subcontractors of these companies, hackers who get access to the cloud storage, or simply other individuals who browse the open data.

Here, we consider the scenario illustrated in Figure 1 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.

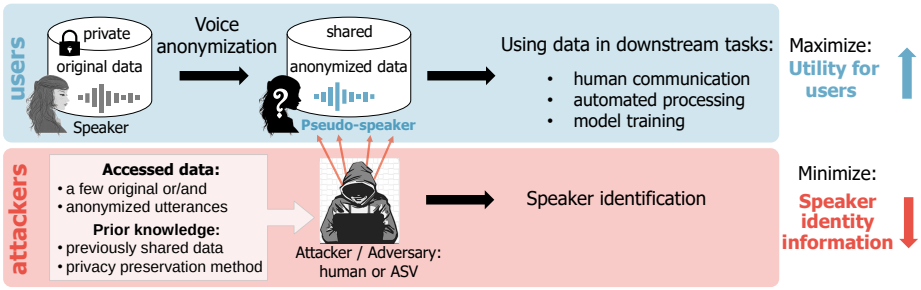


Figure 1: Example of a privacy preservation scenario as a game between *users* and *attackers* in the case where speaker identity is considered as personal information to be protected.

2.1.1. Anonymization task

The sentences shared by the users are called *trial* utterances.[3] 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 call *pseudo-speaker* since it may 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) hide speaker identity, (c) leave other speech characteristics 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.

The requirement (c) promotes the achievement of all possible downstream goals to the best possible extent. In practice, we restrict ourselves to a few goals corresponding to two use cases: ASR training and/or decoding, and multi-party human conversations. The requirement (d) corresponds to the latter goal and is motivated by the fact that, in a multi-party human conversation, the anonymized voices of all speakers must sound natural, be distinguishable from each other, and cannot change over time. The achievement of these goals is assessed via a range of *utility* metrics.

[3]The terms *trial* and *enrollment* are borrowed from the speaker verification literature, where they refer respectively to a speech signal uttered by a speaker willing to be authenticated and a speech signal (or a model) associated with the claimed identity. Although anonymization is a different task (there is no speaker willing to be authenticated), these terms are used here due to the high similarity between the evaluation protocols for these two tasks.

2.1.2. Attack models

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 called *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. The protection of identity information is assessed via *privacy* metrics, including objective speaker verifiability and subjective speaker verifiability and linkability. These metrics assume different attack models.

The objective speaker verifiability metrics (Section 2.3.1) assume that the attacker has access to a single anonymized trial utterance and several enrollment utterances. Two sets of metrics were computed, corresponding to the two attack models when the enrollment utterances are original or they have been anonymized by the user or the attacker. In the *post-evaluation* stage, we considered a stronger attack model where attackers also have access to anonymized training data and can retrain an automatic speaker verification system using this data.

For the subjective evaluation (Section 2.3.2), two situations are considered. The speaker verifiability metric assumes that the attacker has access to a single anonymized trial utterance and a single original enrollment utterance, while the speaker linkability metric assumes that the attacker has access to several original and anonymized trial utterances.

2.2. Datasets

Several publicly available corpora are used for the training, development and evaluation of voice anonymization systems.

Training set. The training set comprises the 2,800 h *VoxCeleb-1,2* corpus (Nagrani et al., 2017; Chung et al., 2018) and 600 h subsets of the *LibriSpeech* (Panayotov et al., 2015) and *LibriTTS* (Zen et al., 2019) corpora. These corpora are among the largest and the most widely used for speaker verification, ASR, and speech synthesis, respectively, hence they are natural choices for training voice anonymization systems which must extract speaker identity and phonetic information and resynthesize a speech signal which hides the former and preserves the latter. The selected subsets are detailed in Table 1 (top).

Development set. The development set involves *LibriSpeech dev-clean* and a subset of the VCTK corpus (Veaux et al., 2019), denoted *VCTK-dev* (see Table 1, middle). With the above attack models in mind, we split them into trial and enrollment subsets. For *LibriSpeech dev-clean*, the speakers in the enrollment set are a subset of those in the trial set. This corpus is meant for objective ASR performance evaluation. For *VCTK-dev*, we use the same speakers for enrollment and trial and we 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 is meant for subjective evaluation of speaker verifiability/linkability in a text-dependent manner. The enrollment and *different* subsets comprises distinct utterances for all speakers.

Table 1: Number of speakers and utterances in the VoicePrivacy 2020 training, development, and evaluation sets.

Subset			**Female**	**Male**	**Total**	**#Utter.**
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 (common)				695
		Trial (different)				10,677
Evaluation	LibriSpeech test-clean	Enrollment	16	13	29	438
		Trial	20	20	40	1,496
	VCTK-test	Enrollment	15	15	30	600
		Trial (common)				70
		Trial (different)				10,748

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

2.3. Utility and privacy metrics

We consider objective and subjective privacy metrics to assess speaker re-identification and linkability. We also propose objective and subjective utility metrics to assess the fulfillment of the user goals specified in Section 2.1.

2.3.1. Objective metrics

For objective evaluation of anonymization performance, two systems were 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 ASV_{eval}, is an automatic speaker verification (ASV) system based on x-vector speaker embeddings and probabilistic linear discriminant analysis (PLDA) (Snyder et al., 2018), which outputs a log-likelihood ratio (LLR) score. The second system, denoted ASR_{eval}, is an automatic speech recognition (ASR) system which outputs a word sequence. Both ASR_{eval} and ASV_{eval} were trained on the *LibriSpeech-train-clean-360* dataset using the Kaldi speech recognition toolkit (Povey et al., 2011). These two models were used in the official challenge setup (Tomashenko et al., 2020b). In addition, for post-evaluation analysis, we trained ASV and ASR systems on anonymized speech data. Both models, denoted $ASV_{\text{eval}}^{\text{anon}}$ and $ASR_{\text{eval}}^{\text{anon}}$, were trained in the same way as ASV_{eval} and ASR_{eval}, respectively.[4]

[4]Scripts for training ASR_{eval} and ASV_{eval} and for evaluation are provided at `https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2020`.

For objective utility evaluation, the official challenge setup relies on the ubiquitous word error rate (WER) metric. The post-evaluation phase also considers the *gain of voice distinctiveness* metric of Noé et al. (2020), which accounts for the additional requirement that the anonymized voices of all speakers must be distinguishable from each other.

For objective privacy evaluation, three well-established speaker verification metrics are computed in the official challenge setup: the *equal error rate (EER)* and the *log-likelihood ratio (LLR)* based costs C_{llr} and $C_{\text{llr}}^{\text{min}}$. As seen in Figure 2, these metrics are computed for 4 evaluation scenarios corresponding to different types of attacks depending on the amount of the attackers' knowledge. Following the terminology of Srivastava et al. (2021), we consider the following conditions.

1. *Unprotected*: no anonymization is performed by users; attackers have access to original trial and enrollment data.
2. *Ignorant attacker (oa)*: original enrollment and anonymized trial data are used for evaluation. We refer to this scenario as *(**o**riginal, **a**nonymized)* or *oa* in short. Users anonymize their trial data, but attackers are unaware of it, hence they use original data for enrollment.
3. *Lazy-informed (aa)* anonymized enrollment and anonymized trial data are used for evaluation. We refer to this scenario as *(**a**nonymized, **a**nonymized)* or *aa* in short. This scenario reflects the situation when the enrollment data are anonymized data produced by users, who are assumed to use the same anonymization system but different pseudo-speakers from their trial data.[5] While it is unlikely that attackers have access to anonymized data with explicit speaker identities, they may infer the identities of a subset of the data from the spoken contents and subsequently use this data as enrollment data. This scenario also reflects the alternative situation when attackers have access to original enrollment data and anonymize them using the same system (which is assumed to be publicly available) so that they become more similar to the anonymized trial data. Here again, the data is anonymized using a different pseudo-speaker, since attackers do not know which pseudo-speaker was picked by each user. Hence, both situations result in the same attack model.
4. *Semi-informed (aa with the model retrained on anonymized data)*: attackers have the same knowledge as in the previous case (the anonymization system, but not the pseudo-speaker picked by each speaker) and, in addition to this, they anonymize the training set for the ASV_{eval} model using the same anonymization system with different pseudo-speakers and retrain it on this data. These attackers are the strongest ones among the

[5] The scenario when enrollment data has been anonymized using the same pseudo-speakers as the trial data is of little practical interest. Indeed, prior work has shown that attackers who have access to such data can identify the speakers as well as if the data had not been anonymized (Srivastava et al., 2020b). Users are therefore required to pick a different, random pseudo-speaker for each conversation.

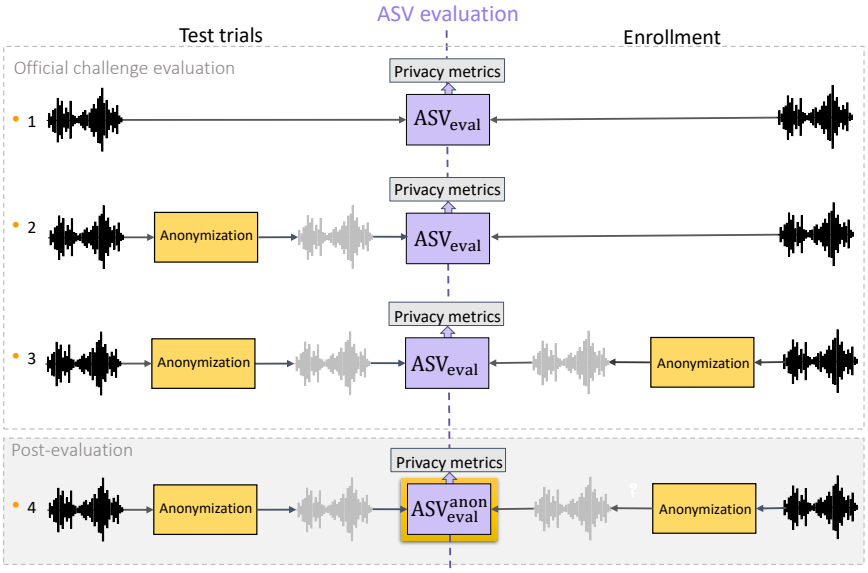


Figure 2: ASV evaluation for the official challenge setup using ASV_{eval} trained on original data is performed for three cases: (1) *Unprotected*: original enrollment and trial; (2) *Ignorant attacker (oa)*: original enrollment and anonymized trial; (3) *Lazy-informed attacker (aa)*: anonymized enrollment and trial. ASV evaluation for the post-evaluation analysis is performed using $ASV_{\text{eval}}^{\text{anon}}$ trained on anonymized data for case (4) *Semi-informed attacker (aa)*: anonymized enrollment and trial.

considered in this paper. This evaluation scenario is part of the post-evaluation stage.

The number of same-speaker and different-speaker trials in the development and evaluation datasets is given in Table 2. In addition to the EER, C_{llr}, and $C_{\text{llr}}^{\text{min}}$, the post-evaluation phase considers one more privacy metric, namely the *de-identification* metric of Noé et al. (2020) which assesses how different each pseudo-speaker is from the original speaker. Note that, although this metric provides useful additional information, it does not directly match the requirements set in Section 2.1. Indeed, the requirement that the original speaker cannot be identified from the anonymized signal does not imply that the pseudo-speaker's voice must be maximally different.

The objective evaluation metrics for privacy and utility are listed below.

Equal error rate (EER). Denoting by $P_{\text{fa}}(\theta)$ and $P_{\text{miss}}(\theta)$ the false alarm and miss rates at threshold θ, the 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_{llr}^{min}). C_{llr} is computed from PLDA scores as defined by Brümmer & Du Preez (2006) and Ramos & Gonzalez-Rodriguez (2008). It can be decomposed into a discrimination loss ($C_{\text{llr}}^{\text{min}}$) and

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

a calibration loss ($C_{\text{llr}} - C_{\text{llr}}^{\text{min}}$). $C_{\text{llr}}^{\text{min}}$ is estimated by optimal calibration using monotonic transformation of the scores to their empirical LLR values.

De-identification and gain of voice distinctiveness. To visualize anonymization performance across different speakers in a dataset, voice similarity matrices have been proposed by Noé et al. (2020). A voice similarity matrix $M = (M(i,j))_{1\le i\le N, 1\le j\le N}$ is defined for a set of N speakers using similarity values $M(i,j)$ computed for speakers i and j as follows:

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

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 n_i and n_j are the numbers of segments for these speakers. Three matrices are computed: M_{oo} on original data, M_{aa} on anonymized data, and M_{oa} on original and anonymized data. For computing the entries $M(i,j)$ of M_{oa}, we use original data for speaker i and anonymized data for speaker j.

Using voice similarity matrices, two additional metrics can be computed: de-identification (DeID) and gain of voice distinctiveness (G_{VD}) (Noé et al., 2020). They are computed based on the ratio of diagonal dominance for two pairs of matrices: $\{M_{\text{oa}}, M_{\text{oo}}\}$ or $\{M_{\text{oo}}, M_{\text{oo}}\}$, respectively. The diagonal dominance $D_{\text{diag}}(M)$ is defined as the absolute difference between the mean values of diagonal and off-diagonal elements:

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

The *de-identification* metric is defined as $\text{DeID} = 1 - D_{\text{diag}}(M_{\text{oa}})/D_{\text{diag}}(M_{\text{oo}})$ and it is expressed in percent. DeID = 100% means perfect de-identification,

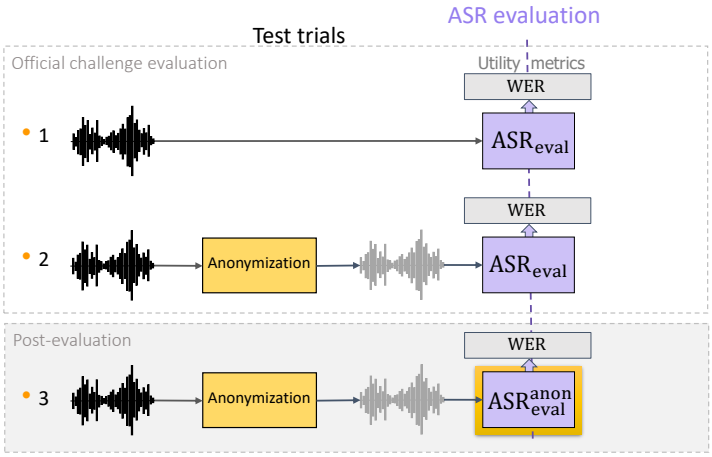


Figure 3: ASR evaluation for the official challenge setup using ASR_{eval} trained on original data is performed for two cases: (1) original trial data and (2) anonymized trial data. ASR evaluation for the post-evaluation analysis is performed using $ASR_{\text{eval}}^{\text{anon}}$ trained on anonymized data for case (3) anonymized trial data.

while DeID = 0% means no de-identification. *Gain of voice distinctiveness* is defined as $G_{\text{VD}} = 10\log_{10}\left(D_{\text{diag}}(M_{\text{aa}})/D_{\text{diag}}(M_{\text{oo}})\right)$, where 0 means that the voice distinctiveness remains globally the same after anonymization, and a gain above or below 0 corresponds respectively to a global increase or a loss of voice distinctiveness.

Word error rate (WER). ASR performance is assessed using ASR_{eval} which is based on the adapted Kaldi recipe for LibriSpeech involving an acoustic model with a factorized time delay neural network (TDNN-F) architecture (Povey et al., 2018; Peddinti et al., 2015), trained on the *LibriSpeech-train-clean-360* dataset, and a trigram language model. As shown in Figure 3, the (1) original and (2) anonymized trial data is decoded using the pretrained ASR_{eval} model and the WERs are calculated. For the post-evaluation analysis, we also perform decoding of anonymized trial data using the $ASR_{\text{eval}}^{\text{anon}}$ model trained on anonymized data (Figure 3, case 3).

2.3.2. Subjective metrics

We consider two subjective privacy metrics (*speaker verifiability* and *speaker linkability*), and two subjective utility metrics (*speech naturalness* and *speech intelligibility*). The speaker verifiability and speech intelligibility metrics are subjective counterparts to the EER/C_{llr}/$C_{\text{llr}}^{\text{min}}$ and WER metrics, and aim to assess how human perception differs from objective evaluation. The speaker linkability metric provides a closer account of the way humans perceive voice characteristics and distinguish voices as belonging to certain speakers. Finally, the speech intelligibility metric is motivated by the requirement that the anonymized voices should sound natural, for which no established objective metric

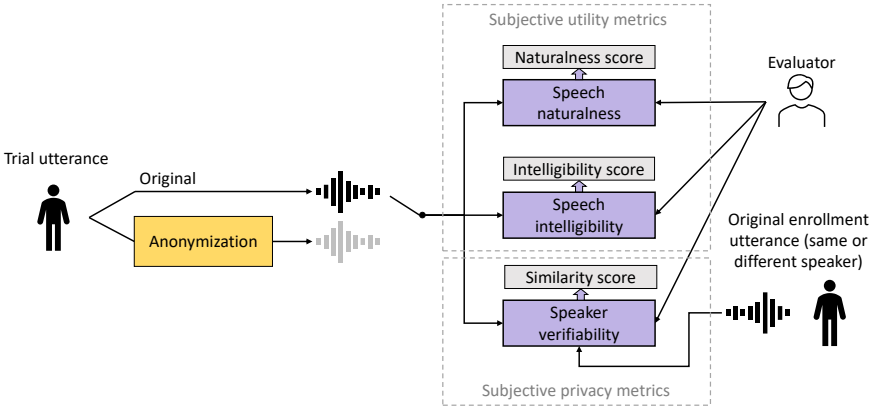


Figure 4: Subjective evaluation test for speech naturalness, intelligibility, and speaker verifiability.

exists.

Subjective speech naturalness, intelligibility, and speaker verifiability. These three metrics were evaluated using the unified subjective evaluation test illustrated in Figure 4. Each evaluator was asked to rate one original or anonymized test set trial at a time. For naturalness, the evaluator assigned a score from 1 ('totally unnatural') to 10 ('totally natural'). For intelligibility, the evaluator assigned a score from 1 ('totally unintelligible') to 10 ('totally intelligible'). For speaker verifiability, the evaluator was required to listen to one original enrollment utterance from the same or a different speaker and rate the similarity between the trial and enrollment voices using a scale of 1 to 10, where 1 denotes 'different speakers' and 10 denotes 'the same speaker' with highest confidence. The evaluator was instructed to assign the scores through a role-playing game.[6]

Every evaluator was required to evaluate 36 trials in one session, following the procedures in Figure 4. He or she could also evaluate more than one session. The trials were randomly sampled from the speakers in the three test sets. The ratio of anonymized vs. original trials was roughly 1:1. So was the ratio of enrollment-trial pairs from the same vs. different speakers. Among the anonymized trials, the proportion of trials from each submitted anonymization system was also balanced. 47 native English speakers participated in the evaluation and evaluated $16,200$ trials. The decomposed numbers of trials over the three test sets are listed in Table 3.

To reduce the perceptual bias of each evaluator, the scores were subject to normalized-rank normalization (Rosenberg & Ramabhadran, 2017). The normalized scores are real-valued numbers in $[0, 1]$. The Mann-Whiteney-U test (Rosenberg & Ramabhadran, 2017) was used to assess statistical significance.

[6]Details are given by Tomashenko et al. (2021b, Section 4.1).

Table 3: Number of trials for the subjective evaluation of speech naturalness, intelligibility, and speaker verifiability. The anonymized trials are from 9 anonymization systems (2 baselines and 7 primary participants' systems). The number of speakers is 30 (15 male and 15 female) in each dataset, i.e., with respect to Table 1, 2 male speakers were re-sampled and 1 female speaker was discarded for *LibriSpeech*.

Test set	**Trials**	**Female**	**Male**	**Total**
LibriSpeech test-clean	Original	1,330	1,330	2,660
	Anonymized	1,330	1,330	2,660
VCTK-test (common)	Original	1,380	1,380	2,760
	Anonymized	1,380	1,380	2,760
VCTK-test (different)	Original	1,340	1,340	2,680
	Anonymized	1,340	1,340	2,680

Perception of speaker identity and speaker linkability. Evaluating the perception of speaker identity by humans is not simple. The subjective verifiability and intelligibility scores described above closely mimic the corresponding objective metrics. Yet, the question whether they suffer from perceptual biases like the memorisation bias (the evaluator recalls hearing the same voice previously) or the well-known priming effect (exposure to a stimulus inconsciently influences the response to a subsequent stimulus) remains open. In order both to assess speaker linkability (i.e., the ability to cluster utterances into speakers) and to decrease as much as possible the influence of such biases, we designed a clustering-based perceptual experiment and the corresponding metrics. We developed a specific software tool for this purpose (O'Brien et al., 2021).[7]

Due to the time-consuming nature of this experiment, only the two baseline anonymization systems were evaluated. 74 evaluators were recruited: 29 are native English speakers and the others are either bilingual or hold a high level of English. Each evaluator did only one session composed of three panels, resulting in a total of 222 panels. Each panel includes 16 utterances from 3 reference speakers (2 to 6 utterances each) and 1 distractor speaker (1 utterance only). Including a distractor helps to verify that the evaluators focus on speaker specificities and are not disturbed by other acoustic differences. The anonymized distractor speaker was used to examine whether anonymization systems affect speaker discrimination performance, e.g., the evaluator either correctly identified the speaker as unique or incorrectly included it in a reference cluster.

For each panel, the evaluators were asked to group the 16 utterances into 1 to 4 clusters according to subjective speaker voice similarity. In order to avoid perceptual biases as much as possible, during a given session, each speaker was encountered in only 1 panel, and all speakers were of the same gender. For the control panel, original speech was used for all utterances; for the two other panels, half of the utterances were anonymized using the same anonymization system. The data used in the speaker clustering task come from the *VCTK-*

[7] https://demo-lia.univ-avignon.fr/voiceprivacy/instructions

test (common) corpus. Unlike all other experiments, only the first 3 s of each utterance were used. The motivation for this length restriction was to provide evaluators with excerpts that were short enough to not induce complex cognitive processes that involve complex syntactic, semantic, and pragmatic analysis. If the evaluators were provided longer excerpts, they could become distracted by attempting to complete and understand text narratives. In addition, limiting the duration of the excerpts reduces the risk of evaluator fatigue.

As a primary metric, we use the macro-average *F-measure* (F_1), a classical metric for such a task. We also use a secondary metric called *clustering purity*. *Clustering purity* associates each cluster with a unique ground truth speaker and focuses only on precision, while F_1 allows two clusters to correspond to the same ground truth speaker and is the harmonic mean of precision and recall. Clustering *purity* is defined as

$$purity(C) = \max_{s \in S} \frac{1}{N} \sum_{c \in C} |c \cap s_c|, \tag{3}$$

where C is the set of estimated clusters, c is an individual cluster in C, S is the set of all possible combinations of unique speakers assigned to each cluster, s_c is the speaker label assigned to cluster c in combination s, and N is the number of utterances in the panel. In addition, we consider a *clustering change* (CC) metric, that is the number of times an evaluator (re-)assigns an utterance to a cluster.

3. Anonymization systems

We now describe the two baseline systems provided by the challenge organizers as well as those prepared by challenge participants.

3.1. Baseline systems

Two different anonymization systems were provided as challenge baselines[8] to help the participants tackle this relatively new task and explore a wide range of solutions. The first baseline offers more flexibility in the choice of the pseudo-speaker and provides state-of-the-art objective privacy and utility, but it requires significant development efforts and big computational resources. In contrast, the second baseline is simpler and provides good subjective speech naturalness and intelligibility, but it results in weaker privacy preservation.

The *primary* baseline, denoted **B1**, is shown in Figure 5. It is inspired from Fang et al. (2019) and performs anonymization using x-vectors (Snyder et al., 2018) and neural speech synthesis. It comprises three steps: (1) x-vector, pitch (F0) and bottleneck (BN) feature extraction; (2) x-vector anonymization; (3) speech synthesis (SS) using the anonymized x-vector and the

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

original F0 and BN features. Step (1) encodes the spoken content by 256-dimensional BN features extracted using a TDNN-F ASR AM trained on the *LibriSpeech train-clean-100* and *train-other-500* datasets and speaker information by a 512-dimensional x-vector extracted using a TDNN trained on the *VoxCeleb-1,2* dataset. Both extractors are implemented with the Kaldi toolkit. Step (2) computes an anonymized x-vector for every original x-vector. It is generated by averaging a set of N^* x-vectors selected at random from a larger set of N x-vectors, itself composed of the N farthest x-vectors in the *LibriTTS train-other-500* dataset, according to PLDA distance.[9] Step (3) uses a SS AM to generate Mel-filterbank features from the anonymized x-vector and the original F0 and BN features, and a neural source-filter (NSF) waveform model (Wang & Yamagishi, 2019) to synthesize a speech signal from the anonymized x-vector and the F0 and Mel-filterbank features. The SS AM and NSF models are both trained on the *LibriTTS train-clean-100* dataset. With respect to the work by Fang et al. (2019), the differences in baseline B1 include using PLDA distance instead of cosine distance and using a different x-vector selection strategy. Also, the model architectures for each step and the training datasets differ. Full details are provided by Tomashenko et al. (2020a). Srivastava et al. (2020a) evaluate these design choices against other possible choices.

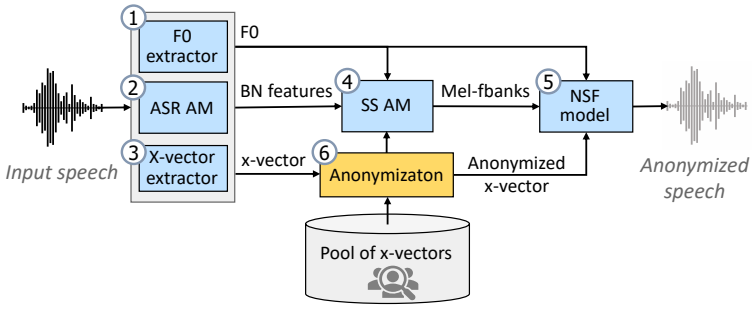


Figure 5: Primary baseline anonymization system (**B1**).

In contrast to the primary baseline, the *secondary* baseline, denoted **B2**, does not require any training data and is based upon traditional signal processing techniques (Patino et al., 2021). It employs the McAdams' coefficient (McAdams, 1984) to achieve anonymization by shifting the pole positions derived from the linear predictive coding (LPC) analysis of speech signals. The process is depicted in Figure 6. 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 LPC coefficients are converted into pole positions by polynomial root-finding. The McAdams' transformation is then applied to the angles of the poles (with respect to the origin in the z-plane), each one of which corresponds to a peak in the spectrum (resembling

[9] In the baseline, we use $N = 200$ and $N^* = 100$.

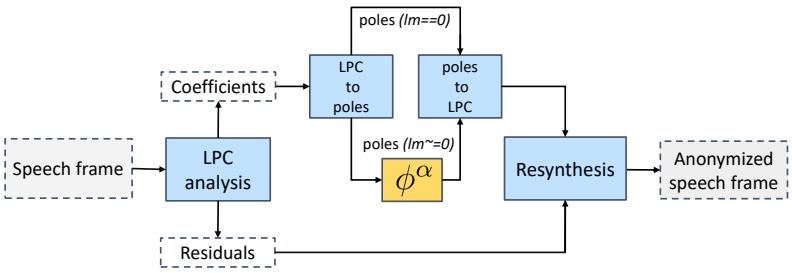


Figure 6: Secondary baseline anonymization system (**B2**).

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 the transformed pole has new, shifted angle ϕ^α. The value of α implies a contraction or expansion of the pole positions around $\phi = 1$. For a sampling rate of 16 kHz, i.e. for data used in the challenge, $\phi = 1$ corresponds to approximately 2.5 kHz which is the approximate mean formant position (Ghorshi et al., 2008). Corresponding complex conjugate poles are similarly shifted in the opposite direction and the new set of poles, including original real-valued 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. This technique shares some similarities with the frequency warping based methods previously explored by Qian et al. (2018a) and Srivastava et al. (2019) except that, for the sake of simplicity, it modifies only the spectral envelope (not the pitch). Full details are provided by Patino et al. (2021).

3.2. Submitted systems

The VoicePrivacy Challenge attracted 45 participants from both academic and industrial organizations and 13 countries, representing 25 teams. Among the 5 allowed submissions by each team, participants were required to designate one as their primary system with any others being designated as contrastive systems. With full descriptions available elsewhere, we provide only brief descriptions of the 16 successful, eligible submissions, a summary of which is provided in Table 4 which shows system identifiers (referred to below) in column 3. Most systems submitted to the VoicePrivacy 2020 challenge were inspired by the primary baseline (see Section 3.2.1). One submission is based upon the secondary baseline (see Section 3.2.2) whereas two others are not related to either (see Section 3.2.3).[10]

[10]There is also one non-challenge entry work related to the challenge (Huang, 2020). This team worked on the development of stronger attack models for ASV evaluation.

Table 4: Teams, organizations, and submitted systems. The submission identifier (ID) for each system in the last column comprises: <team id: first letter of the team name><submission deadline[11]: 1 or 2><c, if the system is contrastive><index of the contrastive system>. The symbol ⋆ in the first column indicates that the team submitted the anonymized training data for post-evaluation analysis. The colors (1) and (2) indicate systems that were developed from **B1** or **B2**, respectively, while ◯ indicates other systems.

Team (Reference)	**Organization(s)**	**Sys.**
AIS-lab JAIST (Mawalim et al., 2020) (A)	•Japan Advanced Institute of Science and Technology, Japan •NECTEC, National Science and Technology Development Agency, Thailand	A1 A2
DA-IICT Speech Group (Gupta et al., 2020) (D)	•Dhirubhai Ambani Institute of Information and Communication Technology, India	D1
Idiap-NKI (Dubagunta et al., 2020) (I)	•Idiap Research Institute, Martigny, Switzerland •École Polytechnique Fédérale de Lausanne (EPFL), Switzerland •Netherlands Cancer Institute (NKI), Amsterdam, Netherlands	I1
Kyoto Team (Han et al., 2020b) (K) ⋆	•Kyoto University, Kyoto, Japan •National Institute of Information and Communications Technology, Kyoto, Japan	K2
MultiSpeech (Champion et al., 2020a) (M) ⋆	•Université de Lorraine, CNRS, Inria, LORIA, Nancy, France •Le Mans Université, LIUM, France	M1 M1c1 M1c2 M1c3 M1c4
Oxford System Security Lab (Turner et al., 2020) (O) ⋆	•University of Oxford, UK	O1 O1c1
Sigma Technologies SLU (Espinoza-Cuadros et al., 2020a) (S) ⋆	•Sigma Technologies S.L.U., Madrid, Spain •Universidad Politecnica de Madrid, Spain	S1 S1c1 S2 S2c1
PingAn (Huang, 2020)	•PAII Inc., Palo Alto, CA, USA	-

3.2.1. Submissions derived from Baseline-1

Teams **A**, **M**, **O** and **S** (see identifiers in column 3 of Table 4 and column 1 of Table 5) submitted systems derived from the primary baseline. Table 5 provides an overview of the modifications made by each team to the baseline modules shown in Figure 5. None of the teams modified the x-vector extraction module (#3 in Table 5), whereas two systems modified the x-vector anonymization module (#6). Details of specific modifications are described in the following. We focus first on differences made to specific modules, then on specific system attributes.

[11] deadline-1: 8th May 2020; deadline-2: 16th June 2020.

Table 5: Summary of the challenge submissions derived from **B1**. ✓ and blue color indicate the components and speaker pool data that were modified w.r.t. **B1**.

Sys.	Description of modifications	1	2	3	4	5	6	Data for speaker pool
		F0	ASR	X-vect.	SS	NSF	Anon.	
A2	using singular value modification						✓	LibriTTS: train-other-500
A1	different F0 extractor[12]; x-vector anonymization using variability-driven ensemble regression modeling	✓					✓	LibriTTS: train-clean-100
M1	End-to-end ASR AM		✓		✓	✓		
M1c1	End-to-end ASR AM; semi-adversarial training to learn linguistic features while masking speaker information		✓		✓	✓		
M1c2	copy-synthesis (original x-vectors)						✓	
M1c3	x-vectors provided to SS AM are anonymized, x-vectors provided to NSF are original						✓	
M1c4	x-vectors provided to SS AM are original, x-vectors provided to NSF are anonymized						✓	
O1	keeping original distribution of cosine distances between speaker x-vectors; GMM for sampling speaker vectors in a PCA-reduced space followed by projection to the original x-vector dimension						✓	LibriTTS: train-other-500
O1c1	**O1** with forced dissimilarity between original and generated x-vectors						✓	VoxCeleb - 1,2
S1	**S1c1** applied on the top of the **B1** x-vector anonymization						✓	
S1c1	domain-adversarial training; autoencoders: using gender, accent, speaker id outputs corresponding to adversarial branches in ANN for x-vector reconstruction						✓	
S2	**S2c1** applied on the top of the **B1** x-vector anonymization						✓	
S2c1	**S1c1** with parameter optimization						✓	

F0:. Only team **A** (Mawalim et al., 2020) modified the pitch extractor. They replaced the baseline F0 extractor with WORLD (Morise et al., 2016) and by SPTK[13] alternatives. While no significant impact upon ASR performance was observed, SPTK F0 estimation was found to have some impact, albeit inconsistent, upon the ASV EER. Consequently, the final system used the baseline F0 extractor. Post-evaluation work conducted by Champion et al. (2020b) showed improved anonymization performance when F0 statistics of the original speaker are replaced with those of a pseudo-speaker, without significant impact upon the ASR performance.

ASR AM, speech synthesis AM and NSF model:. Instead of the baseline hybrid TDNN-F ASR acoustic model, systems **M1** and **M1c1** (Champion et al., 2020a) used an end-to-end model with a hybrid connectionist temporal classification (CTC) and attention architecture (Watanabe et al., 2017) for BN

[12]Different F0 extractors were used in experiments, but the baseline F0 in the final **A1**.
[13]Speech Signal Processing Toolkit (SPTK): `http://sp-tk.sourceforge.net/`

feature extraction. The SS AM and NSF models were then re-trained using the new BN features. In addition, the **M1c1** contrastive system relied on semi-adversarial training of the ASR AM to learn linguistic features while masking speaker information.

X-vector anonymization:. All teams explored different approaches to x-vector anonymization. They are described in the following:

∘***A2***. *Singular value modification* (Mawalim et al., 2020). The singular value decomposition (SVD) of the matrix constructed from the utterance-level speaker x-vectors was used for anonymization. The target x-vector was obtained from the least similar centroid using x-vector clustering. Anonymization was performed through modification of the matrix singular values. A singular value threshold parameter determines the dimensionality reduction used in the modification and determines the percentage of the kept non-zero singular values.

∘***A1***. *Variability-driven decomposition with regression models* (Mawalim et al., 2020). The speaker x-vector was decomposed into high- and low-variability components which were separately modified using two different regression models. It was argued that speaker-specific information is mostly contained in the low-variability component, which is hence the component upon which the anonymisation must focus.

∘***O1***. *Distribution-preserving x-vector generation* (Turner et al., 2020). Baseline **B1** performs anonymization through x-vector averaging. As a result, the anonymized voices are less diverse than the original voices and the resulting differences in the distribution of original vs. anonymized x-vectors leaves the anonymization system vulnerable to inversion. Turner et al. (2020) investigated the use of GMMs to sample x-vectors in a PCA-reduced space in a way that retains the original distribution of cosine distances between speaker x-vectors, thereby improving robustness to inversion.

∘***O1c1***. *Forced dissimilarity between original and anonymized x-vectors* (Turner et al., 2020). In a slight variation to the **O1** system, the **O1c1** contrastive system generates a new x-vector in the case when the original and anonymized x-vectors are not sufficiently dissimilar.

∘***S1c1 & S2c1***. *Domain-adversarial training* (Espinoza-Cuadros et al., 2020a). Domain adversarial training was used to generate x-vectors with separate gender, accent, and speaker adversarial branches in an autoencoder adversarial network. For system **S2c1**, the parameters of the adversarial branches were tuned to optimise the trade-off between the autoencoder and the adversarial objectives.

∘***S1 & S2***. *Domain-adversarial training on top of B1* (Espinoza-Cuadros et al., 2020a). The primary systems **S1** and **S2** are based upon the application of the contrastive systems **S1c1** and **S2c1** to the anonymized x-vectors generated by baseline **B1**.

∘***M1c2***. *Copy-synthesis* (Champion et al., 2020a). This contrastive system is essentially the **B1** baseline, but without *explicit* x-vector anonymization, It provides some insights into the added benefit of the latter, beyond simple copy-synthesis.

∘***M1c3***. *Original x-vectors for NSF.* Another contrastive system for which the NSF model receives original x-vectors while the SS AM receives anonymized x-vectors.

∘***M1c4***. *Original x-vectors for SS AM.* A variation on the above contrastive systems whereby the SS AM receives original x-vectors but the NSF model receives anonymised x-vectors.

∘***A and O***. *Speaker pool augmentation.* In addition to their respective modifications made to x-vector anonymization, some teams also investigated the augmentation of the x-vector pool using additional datasets, namely *LibriTTS-train-clean-100* (team **A**) and *VoxCeleb-1,2* (team **O**).

3.2.2. Submission derived from Baseline-2

∘***D1***. *Modifications of the pole radius* (Gupta et al., 2020). Team **D** investigated modifications of the pole radius (distance from the origin) in addition to the shift in phase operated by baseline **B2**. This approach further distorts the spectral envelope. Pole radii were reduced to 0.975 of the original values whereas the McAdams' coefficient was set to 0.8 as in baseline **B2**.

3.2.3. Other submissions

∘***K2***. *Anonymization using x-vectors, SS models and a voice-indistinguishability metric* (Han et al., 2020b). Similar to the primary baseline **B1**, system **K2** is also based on x-vector anonymization, but the anonymization process and SS models (and corresponding input features) are quite different from those of baseline **B1**. Other differences include using the test dataset for creating the speaker pool. The speech synthesis framework uses two modules: (1) an end-to-end AM implemented with ESPnet[14] which produces a Mel-spectrogram from filterbank features and speaker x-vectors; (2) a waveform vocoder based on the Griffin-Lim algorithm (Griffin & Lim, 1984) which produces a speech waveform from the Mel-spectrogram after conversion to a linear scale spectrogram. A voice indistinguishability metric (Han et al., 2020a) inspired by differential privacy concepts (Dwork, 2009) was applied during x-vector perturbation to select target speaker x-vectors.

∘***I1***. *Modifications to formants, F0 and speaking rate* (Dubagunta et al., 2020). The **I1** system is based upon a signal-processing technique inspired from van Son (2020). The playback speed was adjusted to linearly shift formant frequencies. Individual formants were then shifted to specific target values chosen from

[14] `https://github.com/espnet/espnet/tree/master/egs/librispeech/tts1`

a set of randomly chosen speakers in the *LibriSpeech-train-other-500* dataset. The F0 and the speaking rate were also adjusted using a pitch-synchronous overlap-and-add method (Moulines & Charpentier, 1990). Additional processing includes exchanging the F4 and F5 bands using a Hann filter method and adding modulated pink noise to the speaker F6–F9 bands for formant masking.

4. Results

In this section we report the evaluation results for the systems described in Section 3. The results obtained as part of the challenge and those obtained as part of the post-evaluation analysis are both presented without distinction.

4.1. Objective evaluation results

We first present and discuss the objective evaluation results.

4.1.1. Privacy: objective speaker verifiability

Speaker verifiability results are shown in Figure 7 in terms of EER averaged across all test datasets for the *ignorant* (**oa**) and *lazy-informed* (**aa**) attack models described in Section 2.1.2. Without anonymization, the EER is 3.29%. Anonymization is expected to increase the EER.

When only trial data is anonymized (**oa** condition, light bars in Figure 7), the EER increases for all anonymization systems: from 22.56% for **M1c4** to 53.37% for **M1c1**. Better anonymization is achieved by using x-vector based anonymization systems (**K2, A*, S*, M*, B1, O***) than signal processing based ones (**B2, D1, I1**). Systems **M1c2** and **M1c4** perform worst as expected, because they provide non-anonymized x-vectors to the speech synthesis AM, but they still result in an increased EER compared to original speech due to the acoustic mismatch between original and synthesised speech. Systems **K2, A*, M1c1, M1, B1** all produce EERs above 50%, indicating that the anonymization requirement against ignorant attackers is fully met.

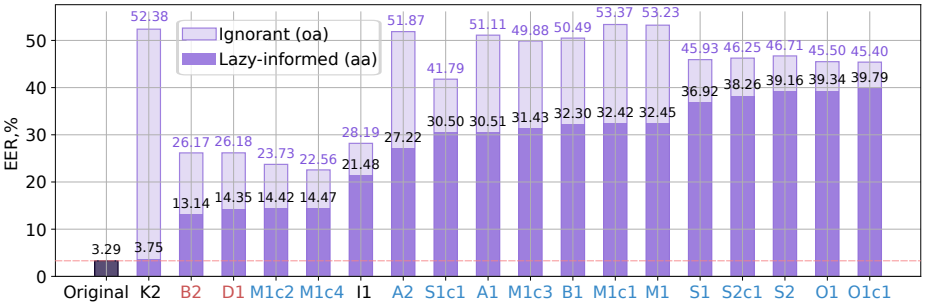


Figure 7: Average EER over all test datasets for all anonymization systems and for original data, against ignorant (**oa**) or lazy-informed (**aa**) attackers. Blue and red colors in the system IDs indicate systems developed from **B1** or **B2**, respectively. Higher EER corresponds to better privacy.

.

Anonymization of both enrollment and trial utterances (**aa** condition, darker bars in Figure 7) results in universally lower EERs for all systems. While system rankings are different for each attack model, the general trend is consistent: **B1** based systems outperform others. Some results are of particular interest. The EER of 3.75% for system **K2** is only marginally above the original EER of 3.29%, despite it being the 3rd best performing system for the **oa** condition. Even the best **O1c1** system achieves an EER of only 37.79%, which is far away from the 50% which indicates successful anonymization. These results highlight the importance of designing anonymization systems under the assumption of a strong attack model. Without it, results may provide a false sense of protection.

Overall, taking confidence intervals (not shown in the figure) into account, baseline **B1** is outperformed by systems **A1**, **A2**, **M1**, **M1c1**, and **K2** in the **oa** condition and by systems **S2**, **S2c1**, **O1**, and **O1c1** in the **aa** condition. These two sets of systems do not intersect and no single system works best in both conditions. This highlights the difficulty of designing and optimising an anonymization system that works well under different attack scenarios. The results for system **K2** are also of note. This system achieves a very high anonymization performance in the **oa** condition due to the fact that anonymized utterances are acoustically very different from the original ones. At the same time, it achieves a very poor performance in the **aa** condition since, instead of generating anonymized x-vectors from a dataset with many speakers (relative to the evaluation dataset), it generates them from the evaluation dataset itself. This results in distinct confusions between some speakers, however the number of such confusions is very low, especially for some test sets (see, for example, the speaker similarity matrix M_{aa} for female speakers on the *LibriSpeech-test* set in Figure 13h).

The results for other privacy metrics are consistent with those for the EER. See, for example, Figure 8 which illustrates EER vs. $C_{\text{llr}}^{\text{min}}$ results for ignorant and lazy-informed attack models for different datasets and systems. Due to space constrains, we therefore focus on the EER in the following. Results for other metrics are reported by Tomashenko et al. (2021b, Section 3).

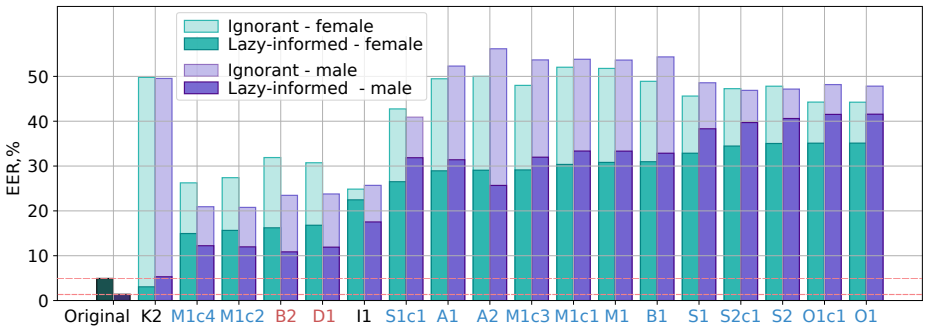
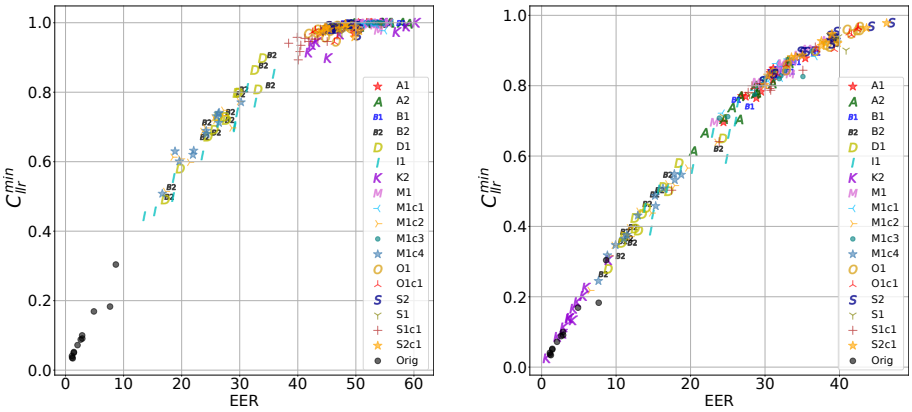


Figure 9: Average EER over all test datasets for all anonymization systems and for original data, depending on the attack model and the original speaker's gender.

.



(a) Ignorant attacker (oa)

(b) Lazy-informed attacker (aa)

Figure 8: EER vs. $\boldsymbol{C}_{\text{llr}}^{\text{min}}$ results for the ignorant and lazy-informed attack models. Each point corresponds to the result of one anonymization system on one dataset among the 12 development and evaluation datasets. Higher EER and $C_{\text{llr}}^{\text{min}}$ correspond to better privacy.

Anonymization performance differs at the gender level. Gender-dependent results averaged over all datasets in both evaluation conditions are given in Figure 9. On the original data, the EER is lower for male speakers than for female speakers. With only few exceptions (e.g., **A2**), the opposite is observed after x-vector based anonymization. Systems **M1c2** and **M1c4**, for which the SS AM is fed with original x-vectors, are two of the exceptions, indicating that gender-dependent differences are the result of x-vector anonymization rather than any extraneous influence, e.g., acoustic mismatch between original and synthesized data. In contrast, signal processing based approaches show the same gender-dependent trend as the original data.

4.1.2. Utility: speech recognition error

Figure 10 shows the ASR performance in terms of the WER. Since we observed substantial disparities in the results, they are illustrated separately for the *LibriSpeech-test* and *VCTK-test* datasets. The WER on original data (no anonymization) is 4.14% and 12.81%, respectively. This difference in the WER is explained as follows: with the ASR system being trained on the *LibriSpeech-train-clean-360* dataset, performance is better on the matched *LibriSpeech-test* dataset than on the mismatched *VCTK-test* set.

All anonymization systems degrade the WER. In other words, any improvement in privacy comes at the expense of lower utility. The relative WER increase is more substantial on the *LibriSpeech-test* dataset (40–217%) than on the *VCTK-test* dataset (14–120%).

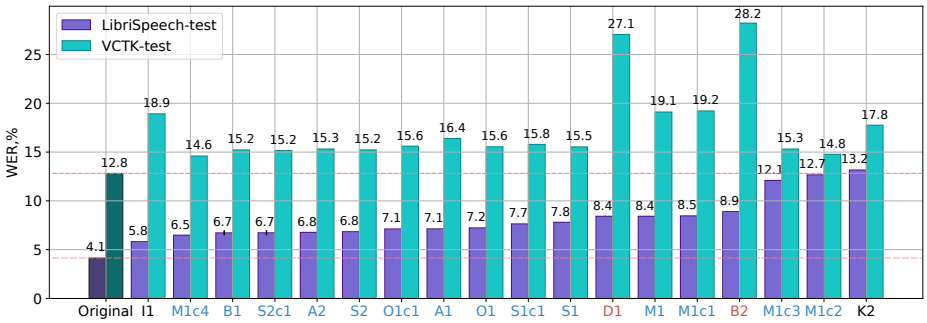


Figure 10: WER on *LibriSpeech-test* and *VCTK-test* for all anonymization systems and for original data. Lower WER corresponds to better utility.

After anonymization, the best WER of 5.83% on the *LibriSpeech* dataset is obtained by the signal processing based system **I1**. Compared to other systems, however, it performs poorly on the *VCTK-test* dataset. Other signal processing based systems based upon baseline **B2** fair even worse on this dataset. On average, on both test sets, x-vector based anonymization techniques related to the primary baseline (**B1**, **S2c1**, **A2**, **S2**) obtain better results than other systems (and very close to each other).

Of note is the high WER of system **M1c2**, which retains the original x-vectors, on the *LibriSpeech-test* dataset. Systems **M1c3** and **M1c4**, which partially retain the original x-vectors, yield a higher WER than original data on that dataset too. This suggests that resynthesis by itself significantly degrades ASR performance. The results for systems **M1** and **M1c1** (vs. **B1**) indicate that using an end-to-end ASR AM for BN feature extraction degrades ASV performance on both datasets. For signal processing based techniques (**I1**, **D1**, **B2**) the relative WER degradation is similar across the datasets, while for x-vector based techniques it is much larger on in-domain data with respect to the data used to train the ASR model (*LibriSpeech*) than on out-of-domain data.

4.1.3. Using anonymized speech data to assess privacy

The results reported in Section 4.1.1 were obtained using an ASV system trained on original data. We now report evaluation results using ASV systems trained on anonymized data, according to the semi-informed attacker scenario in Section 2.3.1. Four teams submitted anonymized *LibriSpeech-train-clean-360* training data for their primary systems **O1**, **M1**, **S2**, and **K2**, and we trained four new corresponding $ASV_{\text{eval}}^{\text{anon}}$ models on this data. In addition, we trained two $ASV_{\text{eval}}^{\text{anon}}$ models on the training data anonymized by the baseline systems **B1** and **B2**. Models were trained in the same way as before, and have the same topology as the ASV_{eval} model trained on original data.

Figure 11 compares the average EERs obtained for the semi-informed (dark, lower bars), lazy-informed, and ignorant attack models. For all anonymization systems, training the ASV evaluation model on anonymized data significantly

decreases the EER: EERs are substantially lower against semi-informed than ignorant or lazy-informed attackers. Thus, assessing the performance of anonymization systems using an ASV system trained on original data leads to a false impression of protection; if the ASV system is retrained on similarly anonymized data, the level of protection becomes closer to (but still better than) that for original, unprotected data.

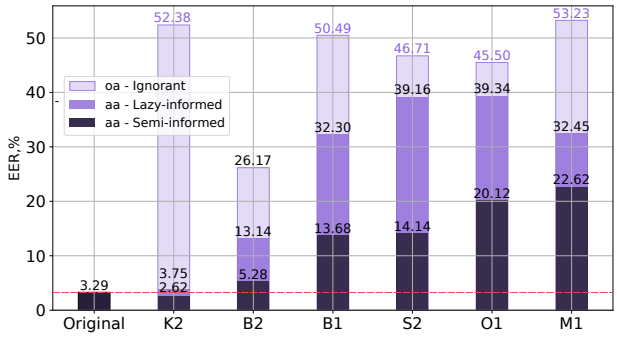


Figure 11: Average EER over all test datasets for a subset of anonymization systems and for original data, against the three attack models.

4.1.4. Using anonymized speech data to assess utility

Similarly, Figure 12 compares the WERs obtained by ASR systems trained on anonymized speech data ($ASR_{\text{eval}}^{\text{anon}}$) with those obtained by the ASR system trained on original data (ASR_{eval}). The WERs for $ASR_{\text{eval}}^{\text{anon}}$ (dark, lower bars, **a**) are consistently lower than for ASR_{eval} (light, upper bars, **o**). In some cases, the WER decreases to a level close to that of ASR_{eval} on original data. This implies that degradations to utility can be offset simply by retraining using similarly anonymized data. This substantially improves the trade-off between privacy and utility; there is potential to protect privacy with only modest impacts upon utility.

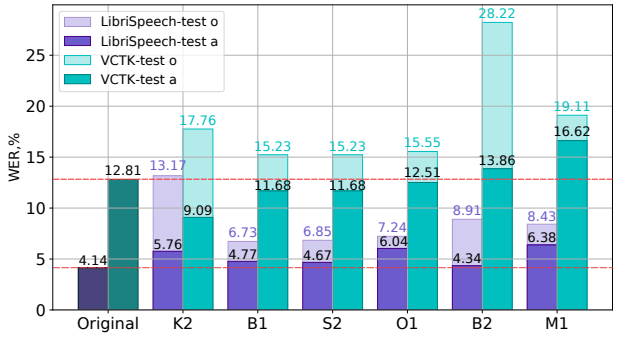
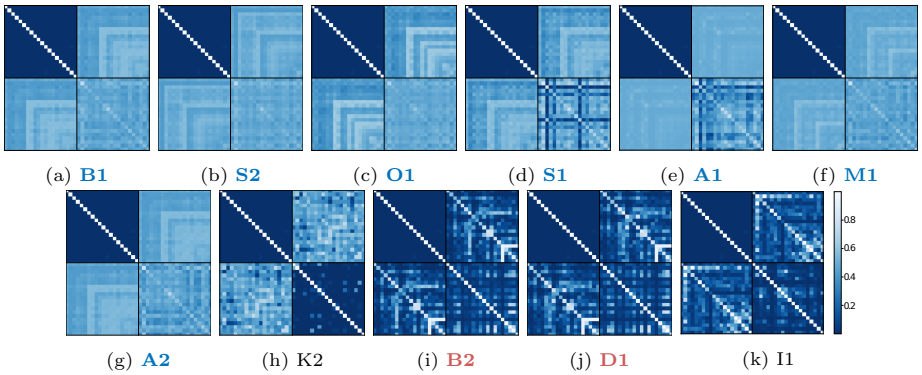


Figure 12: WER on *LibriSpeech-test* and *VCTK-test* for a subset of anonymization systems and for original data, evaluated using ASR_{eval} (**o**) or $ASR_{\text{eval}}^{\text{anon}}$ (**a**).

4.1.5. De-identification and gain of voice distinctiveness

Figure 13 illustrates the voice similarity matrices obtained for all primary systems. The distinct diagonal in M_{oo} (top left submatrix of each matrix M) points out the speaker discrimination ability in the original data. The two other submatrices, M_{oa} (top right) and M_{aa} (bottom right), show substantial differences across the systems. In M_{oa} the diagonal disappears if the pseudo-speakers differ from the original speakers, while in M_{aa} the diagonal emerges if the pseudo-speakers can be distinguished from each other (Noé et al., 2020). The matrices for signal processing based systems and for system **K2** exhibit a distinct diagonal in M_{aa}, indicating that voices remain distinguishable after anonymization. For x-vector based systems, this diagonal is much weaker.



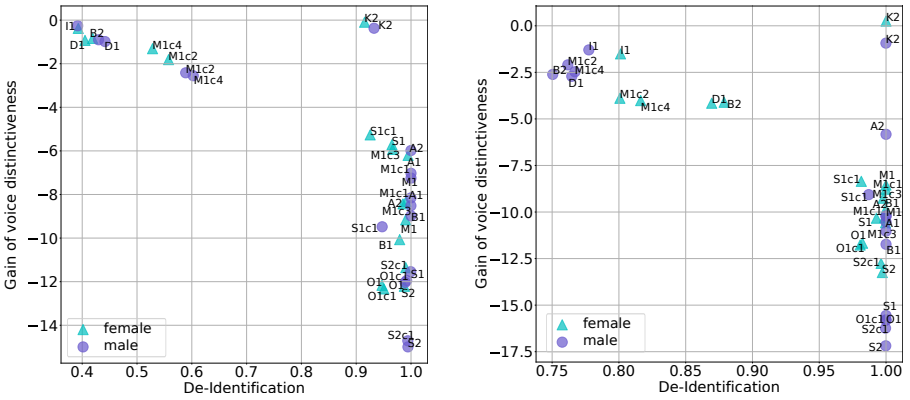
(a) **B1** (b) **S2** (c) **O1** (d) **S1** (e) **A1** (f) **M1**

(g) **A2** (h) K2 (i) **B2** (j) **D1** (k) I1

Figure 13: Voice similarity matrices for all primary systems on the female speakers of the *LibriSpeech-test* dataset. The global matrix M for each system is composed of the three submatrices M_{oo}, M_{oa} and M_{aa} defined in Section 2.3.1 as $M = \begin{pmatrix} M_{\text{oo}} & M_{\text{oa}} \\ M_{\text{oa}} & M_{\text{aa}} \end{pmatrix}$.

The scatter plots in Figure 14 show the *gain of voice distinctiveness* (G_{VD}) against *de-identification* performance (DeID) for the *LibriSpeech-test* (left) and *VCTK-test* (right) datasets.[15] The results show that systems based upon baseline **B1** provide close to perfect de-identification, while signal processing based solutions tend to better preserve voice distinctiveness. For the latter, de-identification performance varies across the datasets. Only system **K2** achieves high de-identification with only modest degradation to voice distinctiveness. The results for systems **M1c4** and **M1c2** which use original x-vectors show that copy-synthesis alone also degrades voice distinctiveness. Interestingly, de-identification performance for both systems is comparable to that for signal-processing based methods. These observations are consistent with EER and $C_{\text{llr}}^{\text{min}}$ results.

[15] For more details, see Tomashenko et al. (2021b, Sections 3.4 and 3.5)



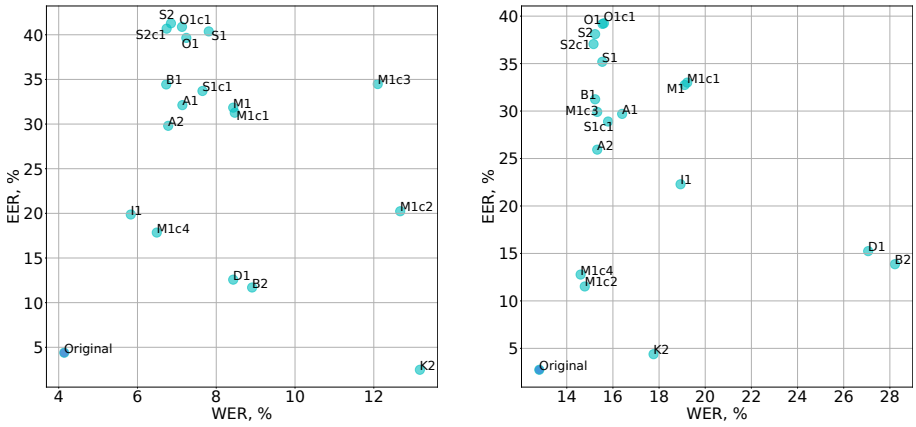
(a) LibriSpeech-test (b) VCTK-test (different)

Figure 14: De-identification (DeID) vs. gain of voice distinctiveness (G_{VD}) for all anonymization systems. Higher DeID corresponds to better privacy, higher G_{VD} to better distinctiveness of anonymized voices.

The results in Figure 14 also show that different systems lead to differences in voice distinctiveness for different genders. In particular, systems **S2** and **S2c1** better preserve distinctiveness for female speakers, while system **A2** better preserves distinctiveness for male speakers.

4.1.6. Relation between privacy and utility metrics

As we observed above, all anonymization systems reduce the utility of speech data. Therefore, it is important to consider the trade-off between privacy and utility. Figure 15 demonstrates the relation between objective privacy and utility in the form of scatter plots with WER and EER values for all anonymization systems, evaluated using ASR_{eval} and ASV_{eval} systems trained on original data. The best anonymization system should have maximum EER and minimum WER, i.e., be close to the top-left corner. We can see that there is no system which provides the best results for both metrics. On the *LibriSpeech-test* dataset, the best anonymization is achieved using x-vector based systems, while the lowest WER corresponds to system **I1** which is a signal processing based system. However on *VCTK-test*, the results for this system are different and better results for both metrics are obtained using x-vector based systems.



(a) LibriSpeech-test (b) VCTK-test

Figure 15: WER vs. EER on *LibriSpeech-test* and *VCTK-test* for all anonymization systems and for original data, evaluated using ASR_{eval} and the lazy-informed attack model.

4.2. Subjective evaluation results

This section presents subjective evaluation results for speech naturalness, intelligibility, and speaker verifiability (Sections 4.2.1 and 4.2.2), and speaker linkability (Section 4.2.3).

4.2.1. Distribution of naturalness, intelligibility, and verifiability scores

The distributions of normalized naturalness, intelligibility, and speaker similarity scores obtained from the unified subjective test are displayed in Figure 16 as violin plots (Hintze & Nelson, 1998).[16] The similarity scores for same-speaker and different-speaker pairs are plotted separately, since they are expected to be different.

The results for naturalness and intelligibility are as expected. Anonymized samples from all systems are inferior to the original data, and the differences are statistically significant at $p \ll 0.01$. This performance gap exists in both methods based on the primary baseline (**B1**, **O1**, **M1**, **S2**, and **A2**) and the secondary baseline (**B2**, **D1**). While **I1** outperforms the other anonymization systems in terms of naturalness, it is still far from perfect in terms of both naturalness and intelligibility. More efforts are necessary to address the degradation caused by existing anonymization methods.

Concerning speaker similarity, the anonymized trial data from a given speaker are perceptually much less similar to the original enrollment data of that speaker than the original trial data of that speaker. This indicates that all systems achieve a good degree of anonymization according to human perception.

[16] Statistical significance test results are reported by Tomashenko et al. (2021b, Tables 16 and 17).

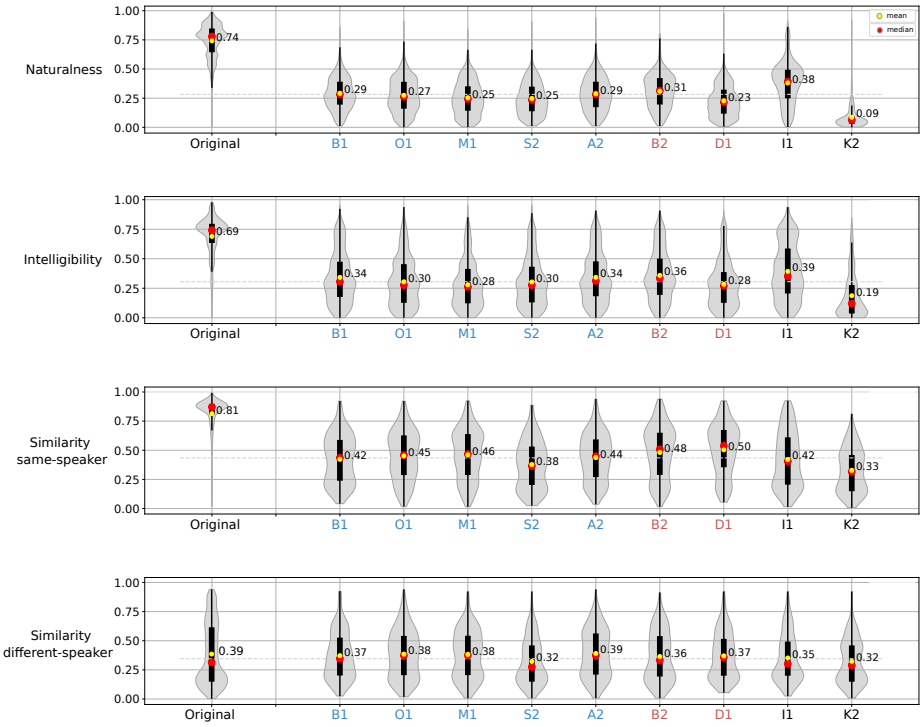


Figure 16: Violin plots of normalized **subjective speech naturalness, intelligibility**, and **speaker similarity** scores pooled over *LibriSpeech-test* and *VCTK-test*. The dotted line indicates the median for **B1**. Numbers indicate mean values. Higher naturalness and intelligibility scores correspond to better utility, and lower similarity scores to better privacy.

4.2.2. Naturalness, intelligibility, and verifiability DET curves

To further investigate the difference across systems, we plot detection error trade-off (DET) curves (Martin et al., 1997). 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. Once again, the DET curves for same-speaker and different-speaker pairs are plotted separately, since they are expected to be different.

The four types of DET curves are plotted in Figure 17.[17] Concerning natu-

[17] For separate results over *LibriSpeech-test* and *VCTK-test*, see Tomashenko et al. (2021b, Figure 26).

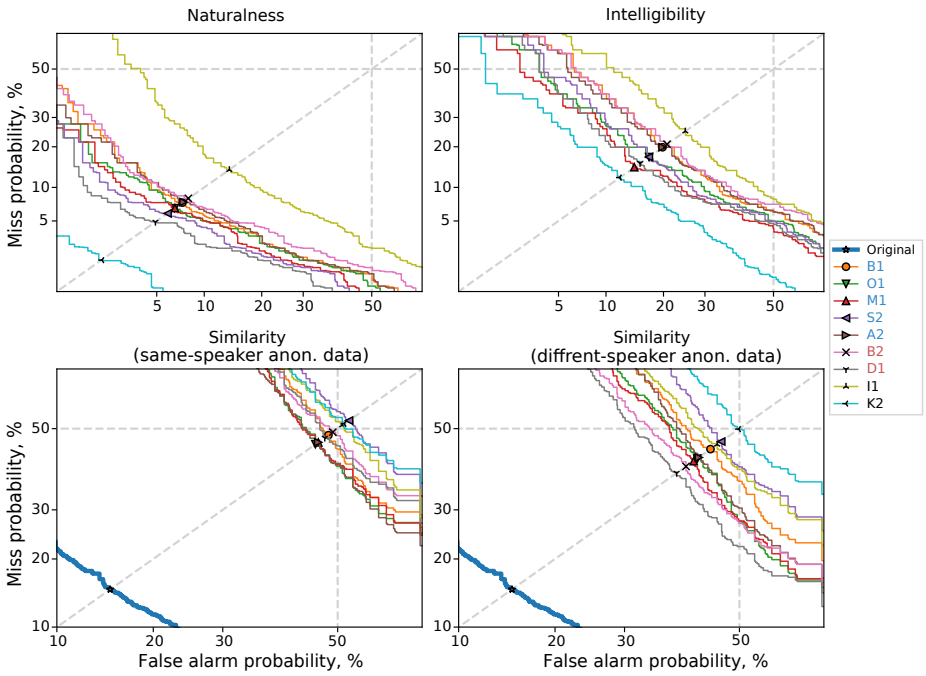


Figure 17: DET curves based on subjective evaluation scores pooled over *LibriSpeech-test* and *VCTK-test*.

ralness and intelligibility, the DET curves for anonymized data are far from the top-right corner, suggesting that anonymized data are inferior to original data in terms of naturalness and intelligibility. The naturalness DET curves of **I1** and **K2** seem to deviate from the other anonymization systems. While other systems are based on either **B1** or **B2**, **I1** uses a different signal processing based approach, and **K2** uses a different deep learning method. As such, **I1** avoids several errors such as ASR AM errors in **B1**, which may contribute to its naturalness. However, it is interesting to note how different signal processing algorithms result in different perceptual naturalness and intelligibility. Also note that none of the systems except **I1** outperforms **B2**.

Concerning speaker similarity, both in the same-speaker and different-speaker cases, the DET curves of original data are close to the bottom-left corner while those of anonymized data are close to the top-right corner. In other words, anonymization of the trial utterances makes it difficult to decide whether the original enrollment utterance comes from the same speaker or not. The similarity DET curves of **K2**, **S2**, and **I1** in the same-speaker case are closer to the top-right corner than others. However, these three systems behave quite differently in terms of naturalness and intelligibility, with **I1** and **K2** achieving the highest and lowest median score, respectively. This implies that an anonymized trial may sound like the voice of a different speaker simply because of the severe distortion caused by anonymization.

To sum up, all the submitted anonymization systems can conceal the perceived speaker identity to some degree. However, none of them can produce anonymized speech that is as natural and intelligible as original speech. One signal processing based anonymization method (**I1**) degrades the naturalness and intelligibility less severely, but it still degrades them to some extent.

4.2.3. Perception of speaker identity and speaker linkability

We report speaker linkability results for the two baseline systems in terms of the *F-measure* (F_1), *clustering change (CC)*, and *clustering purity* metrics. To measure the effects of anonymization for each evaluator, we calculated the *difference* between the values of the F_1 and CC metrics on the control panel (original data only) and their average values over the two other panels (half of the data anonymized by **B1** or **B2**).

We observed a main effect on the mean F_1 difference of the evaluator's native language $F_{1,64} = 6.5$, $p < 0.05$, $\eta_p^2 = 0.09$, but no effects of the anonymization system nor the original speaker's gender, $p > 0.05$. **B1** evaluators exhibited a greater mean F_1 difference (0.24 ± 0.02) than **B2** evaluators (0.21 ± 0.02). Post-hoc t-tests showed that non-native English speaking evaluators were more affected by linking natural and anonymized utterances (0.26 ± 0.02) than native English speaking evaluators (0.19 ± 0.022) (Figure 18a).

For the mean CC difference, we found a main effect of the original speaker's gender $F_{1,64} = 4.45$, $p < 0.05$, $\eta_p^2 = 0.06$, and interactions for anonymization system $\times$ language $F_{1,64} = 4.26$, $p < 0.05$, $\eta_p^2 = 0.06$ and anonymization system $\times$ language $\times$ original gender $F_{1,64} = 8.75$, $p < 0.01$, $\eta_p^2 = 0.11$. Post-hoc t-tests revealed that evaluators showed a greater mean CC difference when presented male utterances (0.07 ± 0.03) in comparison to female (-0.03 ± 0.04) (Figure 18b). Native English speaking evaluators also exhibited a greater mean CC difference than **B2** evaluators (Figure 18c). These results suggest that the evaluators were able to use the anonymized utterances to aid their performance when grouping female utterances, whereas performance diminished when they listened to anonymized male utterances. Non-native English speaking evaluators achieved a lower accuracy when presented with anonymized stimuli from either system. Overall, the above results suggest that the perceptual effectiveness of an anonymization system can depend on the users as well as on the attacker (here, the evaluator).

The distribution of clustering purity for the three panels is displayed in Figure 19a. The Mann-Whitney test shows an effect of the panel (control vs. other) on the purity: $\chi^2 = 82{,}688$ ($p < 0.001$) for female speakers and $\chi^2 = 41{,}344$ ($p < 0.001$) for male speakers, which indicates that the distributions for the original and the anonymized panels are different. As expected, the evaluators achieve a higher average purity (86.40%) on the original panel than on the two other panels (61.68% and 62.58%). These results indicate that linking an anonymized voice to its original counterpart is not as easy as clustering original voices. The distribution of the clustering purity is similar to that of F_1 for all panel types (see Figure 19b). No significant difference between the two baselines is noticed for both metrics.

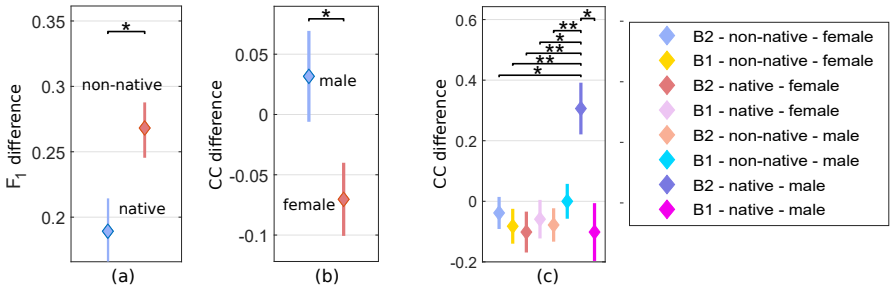


Figure 18: Diamonds and vertical lines represent the means and standard errors, respectively. (a) Mean F_1 difference for native and non-native English speaking evaluators. {*} signifies $p < 0.05$. (b) Mean CC difference depending on the original speaker's gender. (c) Mean CC difference for anonymization system×language×original gender interactions. {*, **} signify $p < \{0.05, 0.01\}$.

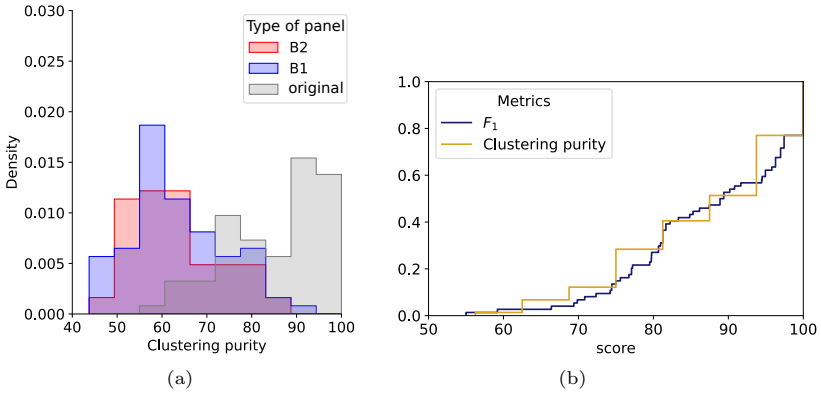


Figure 19: (a) density distributions for clustering purity; (b) cumulative density for clustering purity and F_1 on the original (control) trials.

4.3. Comparison of objective and subjective evaluation results

In this section, we compare objective and subjective evaluation results. Figure 20 plots the EER against the median subjective speaker verifiability (similarity) score for all primary anonymization systems and for original data (blue star) on the three test sets. The results indicate that anonymizing the trial increases the EER and decreases the same-speaker subjective similarity score, while it leaves the different-speaker similarity score roughly unchanged. The precise impact depends on the anonymization system and the test set. This suggests that the considered anonymization systems can hide the speaker identity to some degree from both ASV system and human ears. This is an encouraging message from the challenge. Similar results can be observed for other privacy metrics, as shown by Tomashenko et al. (2021b, Section 5).

Figure 21 plots the WER against the median subjective naturalness and intelligibility scores, averaged over all test datasets. The results reinforce the

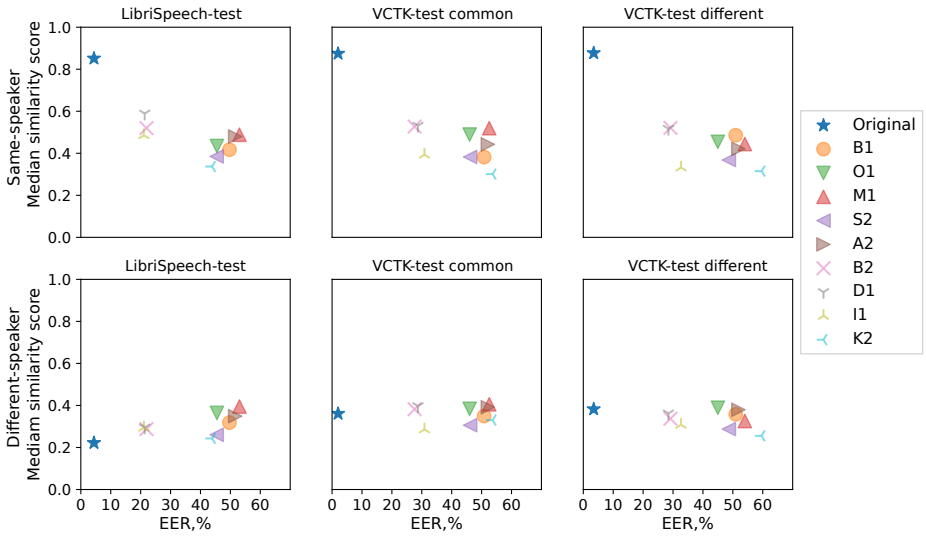


Figure 20: Objective EER (ignorant attack model) vs. subjective same- or different-speaker speaker similarity scores on *LibriSpeech-test* and the two subsets of *VCTK-test*.

observation made earlier that all anonymization systems degrade the objective and subjective utility metrics. On *LibriSpeech-test*, the best results for all utility metrics is achieved by the signal processing based system **I1**, and the worst one by system **K2**. However, on *VCTK-test*, there is no system that performs best (or worst) for all metrics. This is mostly due to the fact that the WER is less consistent across datasets than the subjective naturalness and intelligibility scores.

5. Conclusions

The VoicePrivacy 2020 Challenge was conceived to promote private-by-design and private-by-default speech technology and is the first evaluation campaign in voice anonymization. The voice anonymization task is defined as a game between users and attackers, with three possible attack models each corresponding to adversaries with different knowledge of the applied anonymization methods. The paper describes a full evaluation framework for the benchmarking of different anonymization solutions, including datasets, experimental protocols and metrics, as well as two open-source baseline anonymization solutions in addition to the comprehensive objective and subjective evaluation of both baseline systems and those submitted by challenge participants. These indicate the potential for successful anonymization and serve as a platform for future work in what is now a burgeoning research field.

5.1. Summary and findings

The challenge attracted participants from both academia and industry, including experts already working on anonymization and people new to the field.

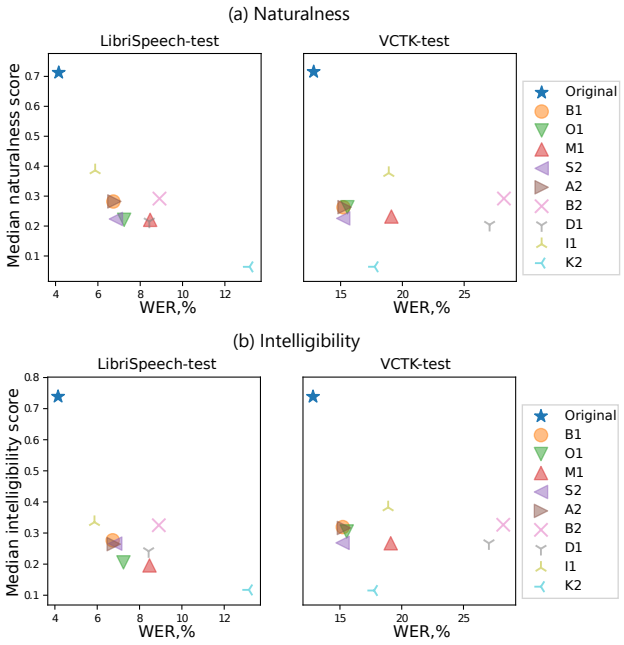


Figure 21: Objective WER vs. subjective naturalness and intelligibility scores averaged over *LibriSpeech-test* and *VCTK-test*.

The submitted anonymization systems can be broadly classified into two classes: x-vector based systems relying on speech synthesis (such as the primary baseline **B1**) and signal processing based systems (relating to the secondary baseline **B2** and system **I1**). X-vector based systems provide the best objective results on average.[18] In contrast, subjective evaluation shows that signal processing based systems tend to yield higher naturalness and intelligibility.

More consistent findings show that anonymization produced by all systems degrade naturalness and intelligibility, as well as the WER. Furthermore, the best systems in terms of WER are based on x-vector anonymization whereas the best system in terms of intelligibility is system **I1**.

Anonymization is also achieved only partially and always at the cost of utility; no single system gives the best performance for all metrics and each system offers a different trade-off between privacy and utility, whether judged objectively or subjectively. This finding holds irrespective of the attack model. While for the ignorant attack model, many systems achieve EERs above 50%, the best results are in the range of $33-43\%$ for the lazy-informed attack model, and in the range of $16-26\%$ for the semi-informed attack model. System rankings are

[18]There are some exceptions, related to the WER results for system **I1** and the *LibriSpeech* dataset in particular.

also different in each case, demonstrating the difficulty of designing an anonymization system that performs well across the range of different VoicePrivacy attacks models.

Challenge participants investigated the proposed anonymization approaches and suggested improvements in some test-cases over the baseline anonymization solutions. They found out, that (1) resynthesis alone degrades utility, while also improving privacy; (2) there is potential for privacy leakage not only in x-vector embeddings, but also in phonetic features and pitch estimates (Champion et al., 2020a; Mawalim et al., 2020); (3) the distribution of anonymized x-vectors differs from that of original x-vectors (Turner et al., 2020). Recent work shows the potential to reduce privacy leakage in pitch estimates while also protecting utility (Champion et al., 2020b; Srivastava et al., 2021). Other findings show that degradations to utility can be mitigated by retraining models used for downstream goals, such as ASR, using anonymized data. Lastly, we identified some differences or bias in performance across different datasets and for different speaker gender. The scale of these differences is one factor, among others discussed below, that warrants further attention in future research.

5.2. Open questions and future directions

A common understanding of VoicePrivacy is still in its infancy. For one, communicating the achievements in layperson terms remains a challenge to better integrate the larger speech community and for outreach to the public at large; for another, VoicePrivacy cannot remain at scratching the surface of privacy issues related to speech and language technology. While considering biometric identity as sensitive information in the first edition, there are other types of sensitive information encoded and transported through speech as a communication medium. Moreover, by constraining the first edition to the operability of speech recognition, linguistic features still allow for extracting biometric characteristics to identify authorship. Depending on the context, the settings of ASV and ASR systems, one might argue that for prompted speech in automated call centers, there is less subjective variability in what is said; let alone, the goal of VoicePrivacy as a community is speech technology as a whole.

Future editions of the VoicePrivacy Challenge will include stronger baseline solutions, possible extensions of the tasks, and re-visited evaluation protocols:

- *Improved anonymization methods for stronger baseline solutions.* For the primary baseline and related approaches, perspective improvements in x-vector based anonymization include adversarial learning (Espinoza-Cuadros et al., 2020a) and design strategies based on speaker space analysis, gender, distance metric, etc. (Srivastava et al., 2020a, 2021). Sensitive information can be further removed from prosodic and other features, in particular, from pitch (Srivastava et al., 2021; Champion et al., 2020b; Gaznepoglu & Peters, 2021) and phonetic (BN) features. Improved algorithms to use the speaker pool should take into account not only speaker characteristics before anonymization but also voice distinctiveness after anonymization. Moreover, the

quality of the synthesized speech using unseen x-vectors has room for improvement. For the secondary baseline, we will consider its extension using a stochastic choice of McAdams' coefficient (Patino et al., 2021).

- *Stronger and more realistic attack models.* Development and investigation of stronger attack models is another potential direction. A knowledgeable and experienced adversary will improve the ASV system and adapt it to make better decisions, i.e., to yield better class discrimination alongside accurate forecasts. Contrary to the conventional experimental validation based on error rates, an adversary actually needs to put a specific threshold and might want to change this threshold, depending on the settings of the ASV systems. In other words, priors and costs that determine the decision policy of an adversary need to be highly adaptable.

- *Alternative privacy and utility metrics and datasets.* The ongoing work on privacy preservation assessment is focusing on the development of new evaluation frameworks, anonymization metrics, and investigation of their correlation and complementarity. This includes the ZEBRA framework (Nautsch et al., 2020; Noé et al., 2022), and objective and subjective linkability metrics (Maouche et al., 2020). Also one may be interested in evaluation that is close to real industry applications and tasks, for example, speaker labeling for diarization, analysis of time and quality required for annotation of real vs. anonymized speech (Espinoza-Cuadros et al., 2020b). The metrics considered in the challenge do not evaluate fully the requirement that all characteristics in the speech signal except the speaker identity should be intact. Relevant utility metrics depend on the user's downstream goals, and for additional downstream goals other utility metrics should be considered. This will require additional datasets for which these goals have been annotated. Datasets collected in real usage conditions should also be considered to assess the impact of acoustic conditions (reverberation, noise, overlapping speech) and full conversations.

- *Attributes.* Besides the speaker identity information, speech also conveys other attributes that can be considered as sensitive, such as emotional state, age, gender, accent, etc. Selective suppression of such attributes is a possible task extension. Except for age and gender which are available in *LibriSpeech*, this will require additional datasets for which these attributes have been annotated.

- *Privacy vs utility trade-off.* The privacy is often achieved at the expense of utility, and an important question is how to set up a proper threshold between privacy and utility (Li & Li, 2009). When developing anonymization methods, a joint optimization of utility gain and privacy loss can be performed by incorporating them into the criterion for training anonymization models (Kai et al., 2021).

- *Integrated approach to voice privacy and security.* In the bigger picture, security and privacy need to be thought of together and not as opposing forces:

positive-sum solutions (Cavoukian, 2017) need to be sought to design technology for better products and services. In other words, while one might draw inspiration from machine learning, forensic sciences, and biometrics, integrated privacy designs for speech and language technology must sacrifice neither security, business interests, nor privacy. Developing of adequate VoicePrivacy safeguards demands future directions that empower capacity for their credible and adequate use in integrated privacy designs which beyond technology include organisational measures.

Acknowledgment

VoicePrivacy was born at the crossroads of projects VoicePersonae, COMPRISE (`https://www.compriseh2020.eu/`), and DEEP-PRIVACY. Project HARPOCRATES was designed specifically to support it. The authors acknowledge support by ANR, JST (21K17775), and the European Union's Horizon 2020 Research and Innovation Program, and they would like to thank Christine Meunier.

References

Aloufi, R., Haddadi, H., & Boyle, D. (2020). Privacy-preserving voice analysis via disentangled representations. In *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop* (pp. 1–14).

Brasser, F., Frassetto, T., Riedhammer, K., Sadeghi, A.-R., Schneider, T., & Weinert, C. (2018). VoiceGuard: Secure and private speech processing. In *Interspeech* (pp. 1303–1307).

Brümmer, N., & Du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech and Language*, *20*, 230–275.

Cavoukian, A. (2017). Global privacy and security, by design: Turning the "privacy vs. security" paradigm on its head. *Health and Technology, Privacy and Security of Medical Information*, *7*, 329–333.

Champion, P., Jouvet, D., & Larcher, A. (2020a). Speaker information modification in the VoicePrivacy 2020 toolchain. `https://hal.archives-ouvertes.fr/hal-02995855`.

Champion, P., Jouvet, D., & Larcher, A. (2020b). A study of F0 modification for x-vector based speech pseudo-anonymization across gender. `https://hal.archives-ouvertes.fr/hal-02995862`.

Chung, J. S., Nagrani, A., & Zisserman, A. (2018). VoxCeleb2: Deep speaker recognition. In *Interspeech* (pp. 1086–1090).

Cohen-Hadria, A., Cartwright, M., McFee, B., & Bello, J. P. (2019). Voice anonymization in urban sound recordings. In *2019 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1–6).

Dubagunta, S. P., van Son, R. J., & Doss, M. M. (2020). Adjustable deterministic pseudonymisation of speech: Idiap-NKI's submission to VoicePrivacy 2020 challenge. `https://www.voiceprivacychallenge.org/docs/Idiap-NKI.pdf`.

Dwork, C. (2009). The differential privacy frontier. In *Theory of Cryptography Conference* (pp. 496–502).

Espinoza-Cuadros, F. M., Perero-Codosero, J. M., Antón-Martín, J., & Hernández-Gómez, L. A. (2020a). Speaker de-identification system using autoencoders and adversarial training. *arXiv preprint arXiv:2011.04696*, .

Espinoza-Cuadros, F. M., Perero-Codosero, J. M., Antón-Martín, J., & Hernández-Gómez, L. A. (2020b). Speaker de-identification system using autoencoders and adversarial training. `https://youtu.be/wCvIh4G3fFM`.

Fang, F., Wang, X., Yamagishi, J., Echizen, I., Todisco, M., Evans, N., & Bonastre, J.-F. (2019). Speaker anonymization using x-vector and neural waveform models. In *Speech Synthesis Workshop* (pp. 155–160).

Gaznepoglu, Ü. E., & Peters, N. (2021). Exploring the importance of f0 trajectories for speaker anonymization using x-vectors and neural waveform models. *arXiv preprint arXiv:2110.06887*, .

Ghorshi, S., Vaseghi, S., & Yan, Q. (2008). Cross-entropic comparison of formants of British, Australian and American English accents. *Speech Communication*, *50*, 564–579.

Gontier, F., Lagrange, M., Lavandier, C., & Petiot, J.-F. (2020). Privacy aware acoustic scene synthesis using deep spectral feature inversion. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 886–890).

Griffin, D., & Lim, J. (1984). Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *32*, 236–243.

Gupta, P., Prajapati, G. P., Singh, S., Kamble, M. R., & Patil, H. A. (2020). Design of voice privacy system using linear prediction. `https://www.voiceprivacychallenge.org/docs/DA-IICT-Speech-Group.pdf`.

Han, Y., Li, S., Cao, Y., Ma, Q., & Yoshikawa, M. (2020a). Voice-indistinguishability: Protecting voiceprint in privacy-preserving speech data release. *arXiv preprint arXiv:2004.07442*, .

Han, Y., Li, S., Cao, Y., & Yoshikawa, M. (2020b). System description for Voice Privacy Challenge. Kyoto team. `https://www.voiceprivacychallenge.org/docs/Kyoto.pdf`.

Hashimoto, K., Yamagishi, J., & Echizen, I. (2016). Privacy-preserving sound to degrade automatic speaker verification performance. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5500–5504).

Hintze, J. L., & Nelson, R. D. (1998). Violin plots: a box plot-density trace synergism. *The American Statistician*, *52*, 181–184.

Huang, C.-L. (2020). Analysis of PingAn submission in the VoicePrivacy 2020 Challenge. `https://www.voiceprivacychallenge.org/docs/PingAn.pdf`.

Kai, H., Takamichi, S., Shiota, S., & Kiya, H. (2021). Lightweight voice anonymization based on data-driven optimization of cascaded voice modification modules. In *2021 IEEE Spoken Language Technology Workshop (SLT)* (pp. 560–566).

Leroy, D., Coucke, A., Lavril, T., Gisselbrecht, T., & Dureau, J. (2019). Federated learning for keyword spotting. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6341–6345).

Li, T., & Li, N. (2009). On the tradeoff between privacy and utility in data publishing. In *15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 517–526).

Maouche, M., Srivastava, B. M. L., Vauquier, N., Bellet, A., Tommasi, M., & Vincent, E. (2020). A comparative study of speech anonymization metrics. In *Interspeech* (pp. 1708–1712).

Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997). *The DET curve in assessment of detection task performance*. Technical Report National Inst of Standards and Technology Gaithersburg MD.

Mawalim, C. O., Galajit, K., Karnjana, J., & Unoki, M. (2020). X-vector singular value modification and statistical-based decomposition with ensemble regression modeling for speaker anonymization system. In *Interspeech* (pp. 1703–1707).

McAdams, S. (1984). *Spectral fusion, spectral parsing and the formation of the auditory image*. Ph.D. thesis Stanford University.

Mdhaffar, S., Bonastre, J.-F., Tommasi, M., Tomashenko, N., & Estève, Y. (2021). Retrieving speaker information from personalized acoustic models for speech recognition. *arXiv preprint arXiv:2111.04194*, .

Morise, M., Yokomori, F., & Ozawa, K. (2016). World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, *99*, 1877–1884.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*, 453–467.

Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: a large-scale speaker identification dataset. In *Interspeech* (pp. 2616–2620).

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

Nautsch, A., Jimenez, A., Treiber, A., Kolberg, J., Jasserand, C., Kindt, E., Delgado, H. et al. (2019b). Preserving privacy in speaker and speech characterisation. *Computer Speech and Language*, *58*, 441–480.

Nautsch, A., Patino, J., Tomashenko, N., Yamagishi, J., Noé, P.-G., Bonastre, J.-F., Todisco, M., & Evans, N. (2020). The Privacy ZEBRA: Zero evidence biometric recognition assessment. In *Interspeech* (pp. 1698–1702).

Noé, P.-G., Bonastre, J.-F., Matrouf, D., Tomashenko, N., Nautsch, A., & Evans, N. (2020). Speech pseudonymisation assessment using voice similarity matrices. In *Interspeech* (pp. 1718–1722).

Noé, P.-G., Nautsch, A., Evans, N., Patino, J., Bonastre, J.-F., Tomashenko, N., & Matrouf, D. (2022). Towards a unified assessment framework of speech pseudonymisation. *Computer Speech & Language*, *72*, 101299.

O'Brien, B., Tomashenko, N., Chanclu, A., & Bonastre, J.-F. (2021). Anonymous speaker clusters: Making distinctions between anonymised speech recordings with clustering interface. In *Interspeech* (pp. 3580–3584).

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

Pathak, M. A., Raj, B., Rane, S. D., & Smaragdis, P. (2013). Privacy-preserving speech processing: cryptographic and string-matching frameworks show promise. *IEEE Signal Processing Magazine*, *30*, 62–74.

Patino, J., Tomashenko, N., Todisco, M., Nautsch, A., & Evans, N. (2021). Speaker anonymisation using the McAdams coefficient. In *Interspeech* (pp. 1099–1103).

Peddinti, V., Povey, D., & Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Interspeech* (pp. 3214–3218).

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

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N. et al. (2011). The Kaldi speech recognition toolkit. In *ASRU*.

Qian, J., Du, H., Hou, J., Chen, L., Jung, T., & Li, X.-Y. (2018a). Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity. In *16th ACM Conference on Embedded Networked Sensor Systems* (pp. 82–94).

Qian, J., Du, H., Hou, J., Chen, L., Jung, T., Li, X.-Y., Wang, Y., & Deng, Y. (2017). Voicemask: Anonymize and sanitize voice input on mobile devices. *arXiv preprint arXiv:1711.11460*, .

Qian, J., Han, F., Hou, J., Zhang, C., Wang, Y., & Li, X.-Y. (2018b). Towards privacy-preserving speech data publishing. In *2018 IEEE Conference on Computer Communications (INFOCOM)* (pp. 1079–1087).

Ramos, D., & Gonzalez-Rodriguez, J. (2008). Cross-entropy analysis of the information in forensic speaker recognition. In *Odyssey*.

Rosenberg, A., & Ramabhadran, B. (2017). Bias and statistical significance in evaluating speech synthesis with mean opinion scores. In *Interspeech* (pp. 3976–3980).

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

van Son, R. (2020). Adjustable deterministic pseudonymization of speech listening experiment, Report of listening experiments. `https://doi.org/10.5281/zenodo.3773931`.

Srivastava, B. M. L., Bellet, A., Tommasi, M., & Vincent, E. (2019). Privacy-preserving adversarial representation learning in ASR: Reality or illusion? In *Interspeech* (pp. 3700–3704).

Srivastava, B. M. L., Maouche, M., Sahidullah, M., Vincent, E., Bellet, A., Tommasi, M., Tomashenko, N., Wang, X., & Yamagishi, J. (2021). Privacy and utility of x-vector based speaker anonymization. *submitted*, .

Srivastava, B. M. L., Tomashenko, N., Wang, X., Vincent, E., Yamagishi, J., Maouche, M., Bellet, A., & Tommasi, M. (2020a). Design choices for x-vector based speaker anonymization. In *Interspeech* (pp. 1713–1717).

Srivastava, B. M. L., Vauquier, N., Sahidullah, M., Bellet, A., Tommasi, M., & Vincent, E. (2020b). Evaluating voice conversion-based privacy protection against informed attackers. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2802–2806).

Tomashenko, N., Mdhaffar, S., Tommasi, M., Estève, Y., & Bonastre, J.-F. (2021a). Privacy attacks for automatic speech recognition acoustic models in a federated learning framework. *arXiv preprint arXiv:2111.03777*, .

Tomashenko, N., Srivastava, B. M. L., Wang, X., Vincent, E., Nautsch, A., Yamagishi, J., Evans, N., Patino, J., Bonastre, J.-F., Noé, P.-G., & Todisco, M. (2020a). The VoicePrivacy 2020 Challenge evaluation plan. `https://www.voiceprivacychallenge.org/docs/VoicePrivacy_2020_Eval_Plan_v1_3.pdf`.

Tomashenko, N., Srivastava, B. M. L., Wang, X., Vincent, E., Nautsch, A., Yamagishi, J., Evans, N., Patino, J., Bonastre, J.-F., Noé, P.-G., & Todisco, M. (2020b). Introducing the VoicePrivacy initiative. In *Interspeech* (pp. 1693–1697).

Tomashenko, N., Srivastava, B. M. L., Wang, X., Vincent, E., Nautsch, A., Yamagishi, J., Evans, N. et al. (2020c). 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`.

Tomashenko, N., Wang, X., Vincent, E., Patino, J., Srivastava, B. M. L., Noé, P.-G., Nautsch, A., Evans, N., Yamagishi, J., O'Brien, B., Chanclu, A., Bonastre, J.-F., Todisco, M., & Maouche, M. (2021b). Supplementary material to the paper. The VoicePrivacy 2020 Challenge: Results and findings. `https://hal.archives-ouvertes.fr/hal-03335126`.

Turner, H., Lovisotto, G., & Martinovic, I. (2020). Speaker anonymization with distribution-preserving x-vector generation for the VoicePrivacy Challenge 2020. *arXiv preprint arXiv:2010.13457*, .

Veaux, C., Yamagishi, J., & MacDonald, K. (2019). CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92). `https://datashare.is.ed.ac.uk/handle/10283/3443`.

Wang, X., & Yamagishi, J. (2019). Neural harmonic-plus-noise waveform model with trainable maximum voice frequency for text-to-speech synthesis. In *Speech Synthesis Workshop* (pp. 1–6).

Watanabe, S., Hori, T., Kim, S., Hershey, J. R., & Hayashi, T. (2017). Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, *11*, 1240–1253.

Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z., & Wu, Y. (2019). LibriTTS: A corpus derived from LibriSpeech for text-to-speech. In *Interspeech* (pp. 1526–1530).

Zhang, S.-X., Gong, Y., & Yu, D. (2019). Encrypted speech recognition using deep polynomial networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5691–5695).