The VoicePrivacy 2022 Challenge: Progress and Perspectives in Voice Anonymisation

Michele Panariello,* *Student Member, IEEE,* Natalia Tomashenko,* *Member, IEEE,* Xin Wang,* *Member, IEEE,* Xiaoxiao Miao, *Member, IEEE,* Pierre Champion, Hubert Nourtel, Massimiliano Todisco, *Member, IEEE,* Nicholas Evans, *Member, IEEE,* Emmanuel Vincent, *Fellow, IEEE,* Junichi Yamagishi, *Senior Member, IEEE*

***Abstract*—The VoicePrivacy Challenge promotes the development of voice anonymisation solutions for speech technology. In this paper we present a systematic overview and analysis of the second edition held in 2022. We describe the voice anonymisation task and datasets used for system development and evaluation, present the different attack models used for evaluation, and the associated objective and subjective metrics. We describe three anonymisation baselines, provide a summary description of the anonymisation systems developed by challenge participants, and report objective and subjective evaluation results for all. In addition, we describe post-evaluation analyses and a summary of related work reported in the open literature. Results show that solutions based on voice conversion better preserve utility, that an alternative which combines automatic speech recognition with synthesis achieves greater privacy, and that a privacy-utility trade-off remains inherent to current anonymisation solutions. Finally, we present our ideas and priorities for future VoicePrivacy Challenge editions.**

***Index Terms*—anonymisation, pseudonymisation, voice privacy, speech synthesis, voice conversion, attack model**

I. Introduction

SPEECH data contain extensive personal, sensitive information which goes far beyond the spoken message. The speaker identity, health and emotional condition, socio-economic status, geographical origin, among a host of other attributes, can all be estimated from speech recordings [1], [2]. Without safeguards, all such information is potentially disclosed as soon as speech signals are shared. Even when consent is given to the use of speech data for a specific voice service, e.g. those provided by a smart speaker, there is no guarantee that it will not also be used for other purposes.

Privacy can be preserved by sanitising a speech recording of specific personal, sensitive information before it is shared. While the community has made inroads in recent years to develop approaches to disentangle and suppress different sources of such information, effective and comprehensive solutions have yet to be developed. One branch of research in privacy preservation in which progress has been rapid in recent years involves *anonymisation*, namely the suppression of personally identifiable information (PII) or cues which can be used by human listeners and/or automatic systems to infer identity.

PII includes information carried by the pitch or fundamental frequency, the timber or spectral envelope, distinctive spoken content (e.g. the speaker's name or social security number), para-linguistic traits and background sounds, among other sources. While a comprehensive approach calls for *all* such sources of personal, sensitive information (or a selected subset) to be suppressed or masked, thus far the community has focused predominantly upon *voice anonymisation*[1] [3]. It refers to the substitution of a speaker's own, natural voice with that of another, pseudo-speaker, while leaving all other attributes (e.g. linguistic content and para-linguistic attributes) intact [3]. Privacy is assumed to be preserved if any remaining sensitive content can no longer be linked to the original speaker through the use of traditional voice-related cues. Note that voice anonymisation focuses exclusively on concealing the *acoustic* characteristics of the speaker voice and is not concerned with other sources of PII which can also be inferred (e.g. the spoken content of the speech).

The first VoicePrivacy Challenge held in 2020 showed that, while voice anonymisation can improve privacy, the technology at the time was far away from the goal of delivering full anonymisation. Results showed that the voice identity can still be revealed, albeit with greater difficulty, under a semi-informed attack model, and that more robust anonymisation cannot be achieved without degrading utility. The second challenge edition was organised in 2022, with a view to bolstering research effort in the field and to fostering progress in the development of more reliable, effective voice anonymisation solutions. The VoicePrivacy 2022 workshop was held in conjunction with the 2nd Symposium on Security and Privacy in Speech Communication (SPSC),[2] a joint event co-located with Interspeech 2022. Reported in this paper are the principal findings from the challenge results and discussions during the workshop. We describe the privacy preservation scenario addressed with VoicePrivacy, the evaluation methodology, a summary of the baselines, competing systems and results, together with a treatment of post-evaluation work and a roadmap for future research directions and priorities. The challenge results show that no system excelled in all evaluation metrics and that there is currently no silver bullet solution to voice anonymisation.

II. Challenge design

We present the scenario and requirements, the anonymisation task, and the attack models, protocol and datasets.

*The first three authors contributed equally to this work.

[1] Note that, in the legal community, the term "anonymisation" means that this goal has been achieved. Here, it refers to the task to be addressed, even when the method being evaluated has failed.

[2] https://symposium2022.spsc-sig.org/

A. Scenario and requirements

The scenario includes two actors – a user who wishes to preserve privacy using an anonymisation safeguard, and an adversary who wishes to undermine the same safeguard. The user wishes to post online or otherwise share an audio recording. It contains speech in his or her voice (or perhaps that of some other individual[3]) and is shared in order to accomplish some downstream task, e.g. to share personal content via social media or to access a voice interactive information service. The individual wishes to preserve privacy and guard against the misuse of recordings by fraudsters for malicious purposes, e.g. the training of voice conversion or speech synthesis models which would allow the generation of artificial speech recordings (spoofs/deepfakes) in their voice.

Before sharing, the recordings are anonymised so as to ensure, to the extent possible, that they cannot be linked to the speaker whose voice they contain. The speech in an anonymised recording should hence contain the voice of a different individual referred to as a *pseudo-speaker*. The pseudo-speaker might, for instance, have an artificial voice which does not correspond to any real speaker. Nonetheless, the voice of the same pseudo-speaker should always be used for recordings of the same, original speaker. This requirement stems from the eventual use of anonymisation in multi-speaker scenarios in which the voice of different individuals should remain distinctive.

While the voice identity should be masked, other speech characteristics should be preserved. These include both linguistic and para-linguistic attributes. The preservation of linguistic content – the spoken words – is paramount to almost any conceivable downstream task and hence intelligibility should be preserved. The requirement to preserve other, para-linguistic attributes is more dependent upon the specific downstream task (e.g. automatic speech recognition, speaker recognition or diarization, emotion analysis, etc.). To promote the development of anonymisation solutions for reasonably diverse downstream tasks, prosody (intonation, stress, rhythm etc.) should also be preserved.

The privacy adversary seeks to undermine the safeguard and to *re-identify* the original speaker whose voice is contained in an anonymised recording. To do so, the attacker makes comparisons between recordings which contain genuine, unprotected voices and recordings containing anonymised voices. An effective anonymisation system should make re-identification – the linking of anonymised, pseudo-voices to genuine, unprotected voices – as difficult as possible. Note that the potential for an attacker to re-identify the speaker is gauged based solely on the use of acoustic voice characteristics, and does not take into account any additional sources of PII such as that contained in the spoken content.

B. Anonymisation task

The task of VoicePrivacy challenge participants is to develop anonymisation systems which fulfil the requirements outlined above. They should:

(a) output a speech waveform;
(b) conceal the speaker identity;
(c) preserve linguistic and para-linguistic attributes;
(d) ensure that all utterances corresponding to one speaker are anonymised so that they contain the voice of the same pseudo-speaker, while utterances corresponding to different speakers are anonymised so that they contain the voice of different pseudo-speakers.

We define the latter condition as *speaker-level* anonymisation. This is the default requirement for VoicePrivacy challenges. It is different to *utterance-level* anonymisation for which each utterance is anonymised using *different* pseudo-voices, even when they are produced by the same speaker.

C. Attack models

The resources and information that are available to the privacy adversary, and the efforts to which the adversary goes in order to undermine the anonymisation safeguard are defined in the form of an *attack model*. The adversary is assumed to be anonymisation-aware and will adapt in order to increase the chances of re-identifying the speaker. The adversary is furthermore assumed to have access to one or more recordings which contain the genuine, unprotected voice of an individual whose voice is suspected to correspond to the anonymised recording. The adversary can hence make comparisons between the unprotected recordings and the anonymised/protected recordings to infer identity.

The adversary may decide to anonymise the unprotected recordings to reduce domain mismatch during the comparison. From hereon, and in the vein and terminology of automatic speaker verification (ASV), recordings that are anonymised by users who wish to protect their identity are referred to as *trial utterances*, whereas recordings used by privacy adversaries to undermine or reverse the anonymisation are referred to as *enrolment utterances*. The attacker uses an ASV system to verify whether or not the voices in trial and enrolment utterances correspond to that of the same individual.

The attacker is assumed to have access to the same anonymisation system as the user, but not their specific configuration or system parameters. This is a reasonable assumption when the anonymisation system is available to many different users and embraces a worse-case scenario for assessment, i.e. a *strong* attack model. To improve the potential of re-identifying (or not) the original speaker of a trial utterance, the attacker can use the system to anonymise the enrolment utterance in order to reduce the domain mismatch between it and the anonymised trial utterance. This is achieved using a large set of similarly anonymised data which the adversary can use to train a new ASV system, or adapt an existing system, optimised to operate upon anonymised data.

D. Protocols and datasets

A set of protocols was designed and made available to VoicePrivacy 2022 participants so that solutions developed using the same data resources can be meaningfully compared.

[3] Each recording is assumed to contain the speech of a single individual. The anonymisation of recordings containing multiple voices can be also be achieved, e.g. using speaker diarization and the separate application of anonymisation to the set of segments corresponding to each speaker/voice.

Table I: Number of speakers and utterances in the training, development, and evaluation sets [2].

Subset			Female	Male	Total	#Utterances
Training	VoxCeleb-1,2		2,912	4,451	7,363	1,281,762
	LibriSpeech train-clean-100		125	126	251	28,539
	LibriSpeech train-other-500		564	602	1,166	148,688
	LibriTTS train-clean-100		123	124	247	33,236
	LibriTTS train-other-500		560	600	1,160	205,044
Dev.	LibriSpeech dev-clean	Enrollment	15	14	29	343
		Trial	20	20	40	1,978
	VCTK-dev	Enrollment	15	15	30	600
		Trial				11,372
Eval.	LibriSpeech test-clean	Enrollment	16	13	29	438
		Trial	20	20	40	1,496
	VCTK-test	Enrollment	15	15	30	600
		Trial				11,448

A set of publicly-available datasets are used for training, development and evaluation and are the same as those used for the inaugural 2020 challenge edition [3]. They comprise the distinct, non-overlapping sets presented in Table I, all of which are composed of multiple corpora as follows.

a) Training set: The training set comprises the 2,800 h *VoxCeleb-1,2* speaker verification corpora [4], [5] and 600 h subsets of the *LibriSpeech* [6] and *LibriTTS* [7] corpora. See Table I for more details. Challenge participants were permitted to use this data only for the training of an anonymisation system.

b) Development set: The development set comprises the *LibriSpeech dev-clean* dataset and a subset of the *VCTK* dataset [8] denoted *VCTK-dev*. Both are split into trial and enrolment subsets which are used in the manner described in Section II-C above. Two development datasets are used in order to test anonymisation performance under matched and mis-matched conditions; the training partition contains data sourced from the *LibriSpeech* dataset only. Full details can be found in the VoicePrivacy 2022 evaluation plan [9]. Details of the development protocol are shown in the middle of Table I.

c) Evaluation set: The evaluation set comprises the *LibriSpeech test-clean* dataset and another subset of the *VCTK* dataset denoted *VCTK-test*. Details are shown to the bottom of Table I. Both development and evaluation sets contain exclusively utterances in English. The majority of the training set is also in English, however VoxCeleb1-2 also contain speech from non-English speakers [4], [5].

III. Metrics

A set of objective and subjective *privacy* and *utility* metrics were adopted for the VoicePrivacy challenge series to assess voice anonymisation performance and the fulfilment of user goals or downstream tasks respectively. Example downstream tasks are described in Section II-A. A suite of evaluation tools and scripts is freely available.[4]

A. Primary objective assessment

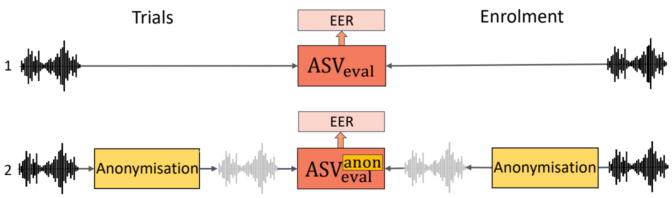


Figure 1: ASV evaluation (1) *unprotected*: original trial and enrollment data, ASV_{eval} trained on original data; (2) *semi-informed* attacker: *speaker-level* anonymised trial and enrollment data with different pseudo-speakers, $ASV^{\mathrm{anon}}_{\mathrm{eval}}$ trained on *utterance-level* anonymised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

Table II: Number of speaker verification trials.

Subset		Trials	Female	Male	Total
Dev.	LibriSpeech dev-clean	Same-speaker	704	644	1,348
		Different-speaker	14,566	12,796	27,362
	VCTK-dev	Same-speaker	2,125	2,366	4,491
		Different-speaker	18,029	17,896	35,925
Eval.	LibriSpeech test-clean	Same-speaker	548	449	997
		Different-speaker	11,196	9,457	20,653
	VCTK-test	Same-speaker	2,290	2,096	4,386
		Different-speaker	17,894	18,210	36,104

1) Privacy – equal error rate (EER): Anonymisation performance is assessed objectively using an ASV system based on x-vector speaker embeddings and probabilistic linear discriminant analysis (PLDA) [10]. As shown in Figure 1, the EER is computed for a pair of evaluation scenarios [3], [11]:

- *Unprotected* — No anonymisation (top of Figure 1). Users share unprotected utterances (trial data). The attacker attempts to identify the speaker using shared, unprotected trial data, unprotected enrolment data and an ASV system (denoted ASV_{eval}) trained using the original, unprotected *LibriSpeech-train-clean-360* dataset.
- *Semi-informed* [12] — Users and the attacker apply anonymisation (bottom of Figure 1) to trial and enrolment data respectively. Users share anonymised trial data, to which the attacker also has access, in addition to unprotected enrollment data. The attacker applies *speaker-level* anonymisation to the latter using the same anonymisation system. Trial and enrollment utterances are, however, anonymised using different pseudo-speakers, since the attacker does not know the pseudo-speaker chosen by each user. The attacker applies *utterance-level* anonymisation[5] to the *LibriSpeech-train-clean-360* dataset and then trains a new ASV system, denoted $ASV^{\mathrm{anon}}_{\mathrm{eval}}$. Using this system, the attacker attempts to re-identify the original speaker of each trial utterance.

The EER for the unprotected scenario forms the baseline. The higher the EER for the semi-informed attack scenario, the better the privacy preservation. The number of same-speaker and different-speaker trials in the development and evaluation datasets is given in Table II. For a given speaker and for both evaluation scenarios, all available enrolment utterances are used to compute an average enrolment x-vector.

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

[5] Previous work [13] has shown that anonymising the training data at the utterance-level rather than the speaker-level results in a stronger attack.

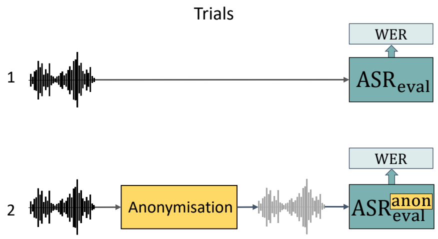


Figure 2: ASR evaluation (1) original data decoded with ASR_{eval} trained on original data; (2) *speaker-level* anonymised data decoded with $ASR^{\text{anon}}_{\text{eval}}$ trained on *utterance-level* anonymised data. The WER is computed on the *trial* utterances of the development and evaluation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

2) Utility metric: word error rate (WER): The preservation of linguistic information is assessed objectively using an ASR system based on the Kaldi toolkit [14]. We adapted the Kaldi recipe for *LibriSpeech* to use an acoustic model with a factorised time delay neural network (TDNN-F) architecture [15] and a *large* trigram language model. As shown in Figure 2, and as for the approach to gauge privacy, we consider two ASR evaluation scenarios:

- *Unprotected* — No anonymisation (top of Figure 2). Unprotected trial data is decoded using an ASR model (denoted ASR_{eval}) trained using the original *LibriSpeech-train-clean-360* dataset.
- *Anonymised* — Anonymised trial data is decoded using an ASR model (denoted $ASR^{\text{anon}}_{\text{eval}}$) trained using the *LibriSpeech-train-clean-360* dataset after it is treated with *utterance-level* anonymisation, using the same anonymisation system as the trial data (bottom of Figure 2).

The first scenario again serves as a baseline. The lower the WER for the second, the better the utility preservation.

3) Privacy-utility tradeoff: New to the 2022 edition of VoicePrivacy was the use of multiple evaluation conditions. These were introduced in recognition of the practical demand for different privacy-utility trade-offs and solutions which can be configured to operate at different operating points, as well as to provide *common* optimisation criteria; for the 2020 challenge, participants had to select an appropriate privacy-utility trade-off themselves, resulting in each team essentially choosing *different* optimisation criteria. Evaluation conditions take the form of increasingly demanding minimum privacy requirements. For each condition, systems which meet the corresponding minimum privacy condition are then ranked according to utility preservation. The primary privacy and utility metrics (EER and WER) are used for this purpose.

To stimulate progress, the 4 evaluation conditions are specified by a range of modest-to-ambitious minimum target EERs: 15%, 20%, 25% and 30%. Participants were encouraged to submit solutions to as many conditions as possible, with submissions to any one condition being required to achieve a weighted average EER for the evaluation set greater than the minimum target. EERs and WERs are equally-weighted averages computed from those for the *LibriSpeech-test-clean* and *VCTK-test* datasets (refer to [9] for full details).

B. Secondary objective metrics

Also new to VoicePrivacy 2022 was the introduction of a pair of secondary utility metrics, namely estimates of the pitch correlation $\boldsymbol{\rho}^{F_0}$ and the gain of voice distinctiveness G_{VD}.

1) Pitch correlation $\boldsymbol{\rho}^{F_0}$: Estimates of pitch correlation are used to approximate the degree to which an anonymisation system preserves intonation. Following [16], the pitch correlation metric $\boldsymbol{\rho}^{F_0}$ is the Pearson correlation between the pitch contours of original and anonymised utterances. The shortest of the two sequences is linearly interpolated so that its length matches that of the longest sequence. The temporal lag between original and anonymised utterances is then adjusted in order to maximise the Pearson cross-correlation when estimated using only segments during which both original and anonymised utterances are voiced. Estimates of $\boldsymbol{\rho}^{F_0}$ are averaged across the full set of utterances in a given data set.

While a secondary metric, all submissions were required to achieve an average pitch correlation of $\boldsymbol{\rho}^{F_0} > 0.3$ for each dataset and for each evaluation condition to which a submission was made. This threshold was set according to anonymisation results for baseline systems (described in Section IV-A).

2) Gain of voice distinctiveness G_{VD}: The gain of voice distinctiveness was adopted for VoicePrivacy 2022 to help observe the consistency in pseudo-voices for speaker-level anonymisation (see Section II-B). G_{VD} is estimated using voice similarity matrices [17], [18]. A voice similarity matrix $M = (M(i,j))_{1\leq i\leq N, 1\leq j\leq N}$ is defined for a set of N speakers. $M(i,j)$ reflects the similarity between the voices of speakers i and j:

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

where $\text{LLR}(x_k^{(i)}, x_l^{(j)})$ is the log-likelihood-ratio obtained by comparing the k-th utterance from the i-th speaker with the l-th utterance from the j-th speaker, and where n_i and n_j are the numbers of utterance for each speaker. LLRs are estimated using the ASV_{eval} model trained using unprotected data. Two matrices are computed: M_{oo}, computed using unprotected utterances; M_{aa}, computed using anonymised utterances. The diagonal dominance $D_{\text{diag}}(M)$ is then computed for both. $D_{\text{diag}}(M)$ is the absolute difference between the mean values of diagonal and off-diagonal elements:

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

G_{VD} is then computed as the ratio of diagonal dominance for each of the two matrices [17]:

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

A gain of $G_{\text{VD}} = 0$ implies the preservation of voice distinctiveness. Positive and negative gains correspond respectively to an average increase or decrease in voice distinctiveness.

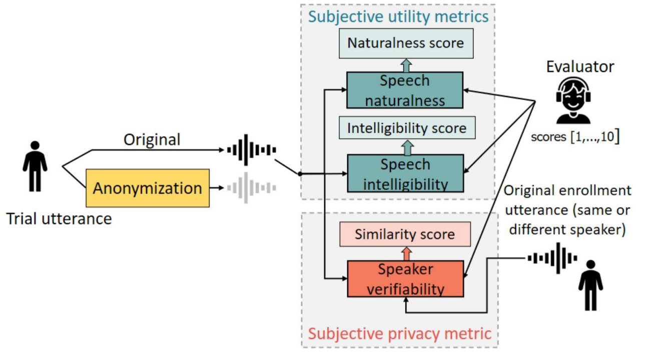


Figure 3: The approach to subjective assessment for speech naturalness, intelligibility, and speaker verifiability.

C. Subjective assessment

As illustrated in Figure 3, subjective metrics include estimates of speaker verifiability, speech intelligibility and speech naturalness. Subjective evaluation tests were conducted by the challenge organisers. For naturalness and intelligibility assessments, evaluators were asked to rate a *single* original or anonymised trial utterance at a time. For naturalness, evaluators assigned a score from 1 (‘totally unnatural’) to 10 (‘totally natural’). For intelligibility assessments, evaluators assigned a score from 1 (‘totally unintelligible’) to 10 (‘totally intelligible’). Both naturalness and intelligibility scores were normalised to within a range between 0 and 1 using rank normalisation [19], with 0 representing the value of lowest naturalness/intelligibility, and 1 representing the highest. Assessments of speaker verifiability were performed with *pairs* of utterances, namely an unprotected enrolment utterance, and either an unprotected or anonymised trial utterance collected from the same or a different speaker. Evaluators assigned a speaker similarity score between 1 (‘the trial and enrolment speakers are surely different’) and 10 (‘the trial and enrolment speaker are surely the same’). Similarity scores were normalised in the same way as the naturalness and intelligibility scores.

The evaluation trials are taken from the *LibriSpeech-test-clean* dataset and include 1,352 unprotected utterances and 104 anonymised utterances per anonymisation system. Each subset of anonymised utterances contains 1 target trial and 1 non-target trial for each of 52 different speakers, evenly split between female and male. The evaluation is performed by 52 native English speakers aged 18 to 70, of whom 40 where male, 11 were female, and 1 of undisclosed gender. Each evaluator rated 52 trials, 26 of which were unprotected, with the remaining utterance being anonymised either by a baseline or by one of the submitted systems. With this configuration, all trial-enrolment pairs used for subjective evaluation were rated by at least one evaluator.

IV. Anonymisation systems

We describe the three VoicePrivacy 2022 baseline systems as well as those prepared by challenge participants. A summary of the description is presented in Table III.

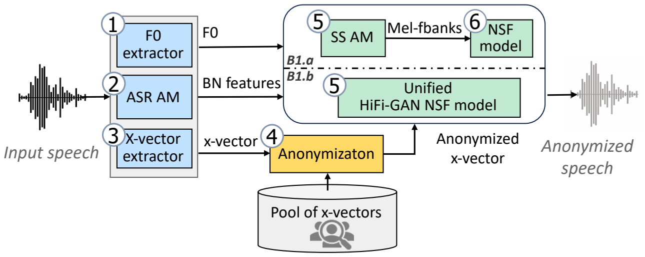


Figure 4: Baseline anonymisation systems `B1.a` and `B1.b`.

A. Baseline systems

Three different anonymisation systems were provided as challenge baselines, denoted `B1.a`, `B1.b`, and `B2`. Baselines `B1.a` and `B1.b` are shown in Figure 4. Inspired by [20], `B1.a` uses x-vectors and neural waveform models, and comprises three steps: first, x-vector [10], pitch (F0) and bottleneck (BN) features [15] which encode linguistic content are extracted from the input utterance (blocks ①, ②, ③); second, the x-vector is anonymised (block ④); third, speech is synthesised using the anonymised x-vector and the original F0 and BN features (blocks ⑤ and ⑥).

Pitch is estimated using YAAPT [21]. BN features are 256-dimensional vectors extracted using a TDNN-F ASR acoustic model (AM) [15]. Speaker encodings are 512-dimensional x-vectors extracted using a time-delay neural network (TDNN) [10]. The anonymisation function (yellow block in Figure 4) converts the original x-vector to an anonymised substitute. Anonymised x-vectors are generated by averaging a set of N^* x-vectors. The latter are selected from a larger set of the N farthest x-vectors from the original x-vector selected at random using a probabilistic linear discriminant analysis (PLDA) [22] distance metric. A speech synthesis (SS) AM generates Mel-filterbank features from the anonymised x-vector and F0+BN features. The speech synthesis module is a neural source-filter (NSF) waveform model [23]. Full details are available in [24]. Baseline `B1.b` is the same as `B1.a`, except that the SS AM is removed and the NSF waveform model is fed directly with BN features. Moreover, the NSF is trained with an additional discriminator loss inspired by the HiFi-GAN system [25].

Baseline `B2` is the technique presented in [26], and is based purely on signal processing techniques. The method utilises a coefficient α which is referred to as the *McAdams* coefficient. Each pseudo-speaker is associated to a value of α randomly sampled from a uniform distribution within the range $(\alpha_{\text{min}}, \alpha_{\text{max}})$. Linear predictive coding (LPC) is used to decompose the input utterance into a set of pole positions and an excitation signal. The poles positions are rotated within the complex plane to adjust the phase ϕ to ϕ^α, hence shifting the formant positions of the input signal. An anonymised utterance is then synthesised using the modified pole positions and the original excitation signal.

B. Submitted systems

The 2022 edition of the VoicePrivacy Challenge attracted submissions from 5 participating teams, all academic organisa-

Table III: A summary of the different systems, techniques and performance.

Team	Feature extraction	Anonymisation	Resynthesis	Results summary
T04	x-vectors and ECAPA vectors are concatenated to create speaker embeddings. Linguistic content is transcribed to phonemes. Removed F0 extraction.	Pseudo-speaker embeddings generated with GAN.	TTS model generates Mel-spectrograms that are converted to waveform by HiFi-GAN.	TTS-based approach provides excellent levels of privacy and utility, but barely passes the $\boldsymbol{\rho}^{F0} > 0.3$ requirement.
T11	Yingram for F0 extraction, U2++ for linguistic features. Speaker embeddings are one-hot representations of the speaker IDs encountered during training. One "pseudo-speaker ID" not corresponding to any real speaker is also stored.	Final pseudo-speaker embedding created by means of a weighted average between K random speaker embeddings and the one-hot representation of the "pseudo-speaker ID".	HiFi-GAN is used to synthesise waveforms from AM-generated Mel-spectrograms.	T11-p4 has one of the best results in terms of privacy and utility, but very low G_{VD}: all pseudo-speakers are "similar".
T18	As for B1.a and B1.b, except for their second system where x-vectors are replaced with Transformer-based ASR embeddings.	Two anonymisation strategies are proposed. The first uses adversarial noise to anonymise the speaker embedding. The second replaces the x-vector embedding with an ASR-based embedding.	As in B1.a.	Both approaches offer modest privacy improvement over B1.a and B1.b at the cost of reduced WER.
T40	F0 curve not extracted from the intput signal directly, but estimated from x-vector and BN features.	As in B1.a and B1.b.	As in B1.b.	Modest improvement over B1 in terms of privacy.
T32	Signal processing-based approach: pitch shift with TD-PSOLA and PV-TSM.			Performance mostly on par with B2 except for a higher $\boldsymbol{\rho}^{F0}$ and subjective intelligibility.

tions. For each evaluation condition, participants were required to designate one submission as their primary system with any others being designated as contrastive systems. Henceforth, submitted systems are named according to the following scheme: <team number>-<'p' if primary, 'c' if contrastive>-<incremental identifier>. With full descriptions of each system available in the literature cited below, we report only brief summaries and provide an overview of the trends.

Team T04 [27] proposed an ASR+TTS-like approach.[6] A connectionist-temporal-classification/attention hybrid ASR model [28] is used to transcribe input speech into phonemes, which are then converted to articulatory feature vectors [29]. Pseudo-speaker embeddings are created by means of a generative adversarial network (GAN) [30]. Articulatory feature vectors and pseudo-speaker embeddings are used to synthesise spectro-temporal representations of an anonymised speech signal using the FastSpeech 2 TTS engine [31]. A HiFi-GAN [25] is then used to synthesise speech signal outputs from the resulting representations. No pitch-related information from the original signal is used in the synthesis step. A key motivation behind this approach is use of textual transcriptions in place of F0 curves to improve the suppression of speaker-related information.

Team T11 [32] replaced the F0 trajectory with Yingrams [33]. The authors argue that Yingram F0 extraction is more reliable in the case of 'creaky' voices caused by irregular glottal pulse periodicity. BN features are extracted with the U2++ model and WeNet toolkit [34]. Unique to this system is the facility to configure the similarity between different pseudo-speaker voices, thereby controlling the privacy protection and voice distinctiveness trade-off. Voice identity is encoded as a one-hot vector of size $N + 1$, where N is the number of speakers in the combined *LibriTTS-train-clean-100* and *LibriTTS-train-other-500* datasets. The additional vector component is referred to as a 'pseudo-ID'. The pseudo-speaker representation is a weighted average between the one-hot vectors of K random speakers and that of the pseudo-ID. The weights act to control pseudo-speaker similarity. A higher weight assigned to the pseudo-ID causes greater convergence of the pseudo-speakers to a single voice. Using pitch, bottleneck and speaker-related features, a spectrogram is generated using an acoustic model based on Tacotron 2 [35], and converted to a waveform using a HiFi-GAN vocoder [25].

Team T18 [36] focused on enhancements to the anonymisation function of baseline B1.a. Two variations were investigated. The first uses adversarial noise to compute the pseudo-speaker embedding from the original speaker embedding. The second explored the substitution of speaker embeddings with features extracted from the encoder component of a speech Transformer [37], with the rationale that they can, to some extent, encode speaker-related information.

Team T40 [38] proposed a variation of baseline B1.b. The YAAPT F0 extractor was replaced with a F0 regressor which uses BN features and pseudo-speaker embeddings to synthesise an F0 contour which better matches the voice identity of the pseudo-speaker. The hypothesis is that this approach can suppress speaker information contained in the F0 contour and result in more natural anonymised speech.

Team T32 [39] was the only team to adopt a signal-processing based approach. Motivated by the comparatively lightweight computational demands and since it does not require training data, they adopted a pitch shifting-based solution. Pitch shifting is achieved by either downsampling or upsampling before one of two different approaches is used to readjust the duration to that of the original speech signal. The first is based upon a time-domain pitch synchronous overlap-add (TD-PSOLA) approach [40]. The second uses phase-vocoder-based time-scale modification (PV-TSM) [41].

[6] Code available at: https://github.com/DigitalPhonetics/speaker-anonymization/tree/phonetic_representations

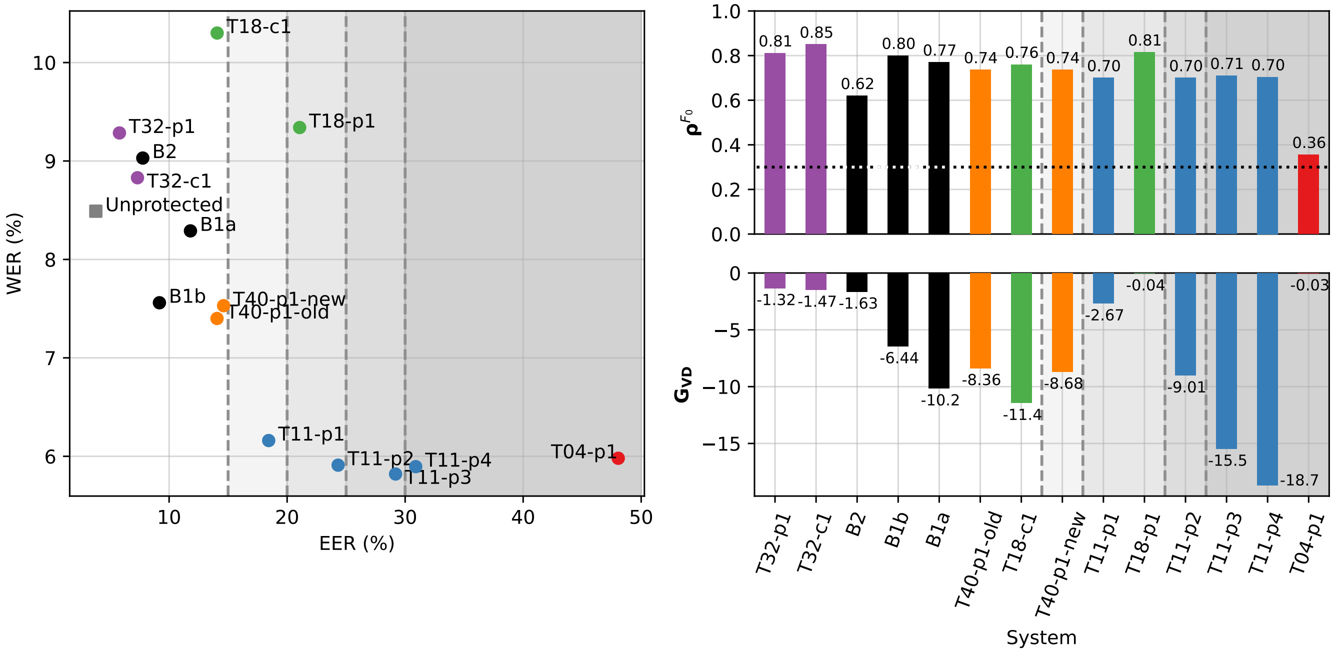


Figure 5: Objective evaluation results for the test set. Unprotected data was evaluated with ASV_{eval} and ASR_{eval}, while anonymised data was evaluated with $ASV^{\text{anon}}_{\text{eval}}$ and $ASR^{\text{anon}}_{\text{eval}}$. Vertical dashed lines indicate the separation between different evaluation conditions. The horizontal dotted line in the pitch correlation plot shows the minimum pitch correlation threshold $\boldsymbol{\rho}^{F_0} = 0.3$. Primary systems are denoted '-p' whereas the single contrastive system is denoted '-c'.

C. Trends

A summary of the submitted systems is illustrated in Table III. In the following we describe the common trends and principal differences between them.

Feature extraction – The majority of systems use the same three components used by the `B1.a` and `B1.b` baseline systems, namely a speaker embedding, linguistic embeddings and a pitch contour. While x-vectors remain popular, `T04` switched to the use of ECAPA-TDNN, as have a number of other works reported post-evaluation (see Section VII). Alternative approaches to the extraction of linguistic embeddings and F0 extraction include WeNet [34] and Yingram [33] respectively.

Anonymisation – Probably because, instinctively, it has a major bearing on anonymisation performance, teams invested notable effort in improving the anonymisation function. Common to the approaches of `T04`, `T11` and `T18` are alternatives to x-vector pooling using a GAN, a one-hot encoded speaker representation, or adversarial noise to generate anonymised speaker embeddings.

Resynthesis – There is comparatively little variation in the exploration of different approaches to resynthesis, perhaps indicating that most teams either found or expect the approach to synthesis to have comparatively less impact upon anonymisation performance. Most x-vector–based systems used a HiFi-GAN (or a variation thereof), as for baseline `B1.b`.

V. Results

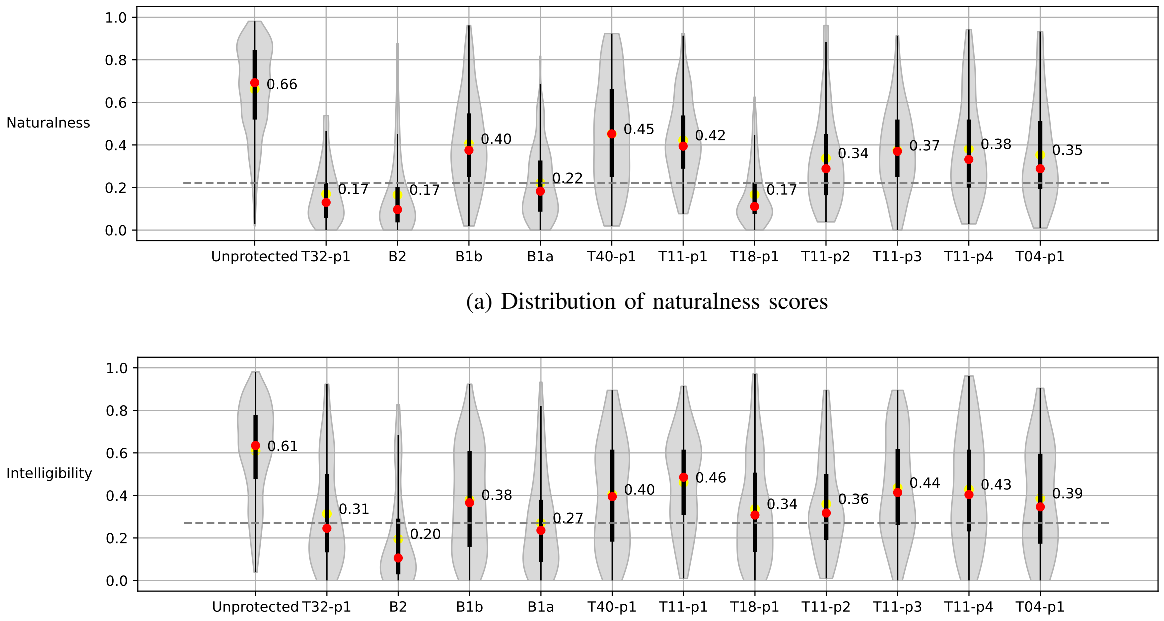
We report results for baseline and submitted systems in terms of both primary and secondary objective metrics and subjective assessment.

A. Objective evaluation results

Primary objective assessment results for baselines and all submitted systems for the evaluation set are illustrated in Figure 5. The plot to the left depicts the privacy-utility trade-off for unprotected data (grey points), for each baseline system (black points) and for each submission (coloured points). All systems submitted by a given team are depicted by points of the same colour. The vertical dashed bars depict the set of minimum target EERs of the four evaluation conditions defined in Section III-A3.

To the right of Figure 5 are results for secondary metrics $\boldsymbol{\rho}^{F_0}$ and G_{VD}. The set of systems is sorted according to the EER (low to high), and each bar has the same colour as corresponding points in the privacy-utility plots. The dashed bars again depict the juncture between evaluation conditions. The horizontal dotted line in the upper plot of pitch correlation results indicates the minimum threshold of $\boldsymbol{\rho}^{F_0} = 0.3$.

All systems produce EERs higher than that for unprotected data and the majority of systems also outperform the baselines. The `T11-p3` submission produces the lowest WER for the 15%, 20% and 25% minimum target EERs. The `T11-p4` system also achieves the lowest WER for the minimum target EER of 30%. While these systems yield high pitch correlation values $\boldsymbol{\rho}^{F_0}$, increased privacy (higher EERs) translates to a decrease in voice distinctiveness (lower G_{VD}), the lowest of all being for the `T11-p4` system. These results nonetheless show the merit in the approach of Team `T11` to weight the pseudo-speaker embedding extraction in order to control the level of privacy. Results show that greater privacy (higher EER) can be delivered with only negligible impact to utility (WER), albeit



(a) Distribution of naturalness scores

(b) Distribution of intelligibility scores

Figure 6: Subjective assessment results of baselines and submitted systems in terms of (a) naturalness and (b) intelligibility. Red and yellow dots represent respectively the median and the mean of each set of scores. The dashed gray line corresponds to the mean score of baseline B1.a, provided to facilitate the comparison with the other systems.

with notable degradation to voice distinctiveness (G_{VD}).

The EER for the T04-p1 submission is substantially higher, just shy of a 50% EER which would indicate perfect anonymisation. This can be attributed to the use of an ASR+TTS-like system. The gain in voice distinctiveness, at $G_{\text{VD}} = -0.03$ is the best of all. However, unsurprisingly for an ASR+TTS-like system, the pitch correlation is by far the lowest, with a value of $\boldsymbol{\rho}^{F_0} = 0.36$, albeit still above the threshold. We return to this result later in Section VI-B.

The efforts of Team T18 to improve the anonymisation function also appear to be successful. Unfortunately, while the EER increases, the modifications also cause an increase in the WER, though the pitch correlation of $\boldsymbol{\rho}^{F_0} = 0.81$ and gain in voice distinctiveness $G_{\text{VD}} = -0.04$ are among the highest.

Perhaps due to the focus on modifications only to F0 extraction, the WER for Team T40 submissions[7] is relatively unaffected compared to results for the B1.b baseline from which their system is derived. Increases to the EER are modest but the degradation to voice distinctiveness is relatively high. While improvements were obtained for the development set, neither of the two approaches to pitch shifting explored by Team T32 are successful in improving privacy in case of the evaluation set. The submissions of both Teams T18 and T32 show high pitch correlation and gain in voice distinctiveness.

B. Subjective evaluation results

Results of subjective naturalness assessment are shown in Figure 6a. The first observation is a universal and substantial degradation in naturalness stemming from anonymisation. Baseline B1.b and derived T40-p1 systems achieve among the highest scores and indicate that the NSF waveform model produces marginally more natural speech when fed directly with BN features instead of Mel-filterbank features. The relatively higher scores for Team T11 systems and lower scores for B1.a and derived T18-p1 systems suggest that adversarially trained vocoders produce more natural speech. The T04-p1 system, albeit ASR+TTS-like, is also competitive. Naturalness scores for the B2 baseline and T32-p1 systems, both signal-processing based, are among the lowest.

The trends for intelligibility scores shown in Figure 6b reflect those for naturalness; anonymisation also universally degrades intelligibility. The B1.b baseline, the T40-p1 and the set of T11 systems are all competitive, as is the ASR+TTS-like T04-p1 system. Scores for signal processing based solutions and the B1.a baseline are the lowest though, in contrast to results for objective utility assessment, the T18-p1 system fares better. We address the correlation between objective and subjective results later in Section VI-A.

Verifiability score histograms shown in Figure 7 depict the distribution of verifiability scores for target trials (the speaker of both enrolment and trial utterances is the same) and non-target trials (the speaker of both enrolment and trial utterances is different). The top-left-most histogram shows distributions for unprotected enrolment and trial utterances (no anonymisation) and that listeners can determine with reasonable reliability when the speaker of each utterance is the same or different.

[7] Team T40 submitted two versions their system: one before the challenge deadline, which contained an implementation error (named T40-p1-old), and one after the deadline, where the bug was fixed (T40-p1-new). While both systems are displayed in the objective metric results, only T40-p1-old was used in the subjective evaluation.

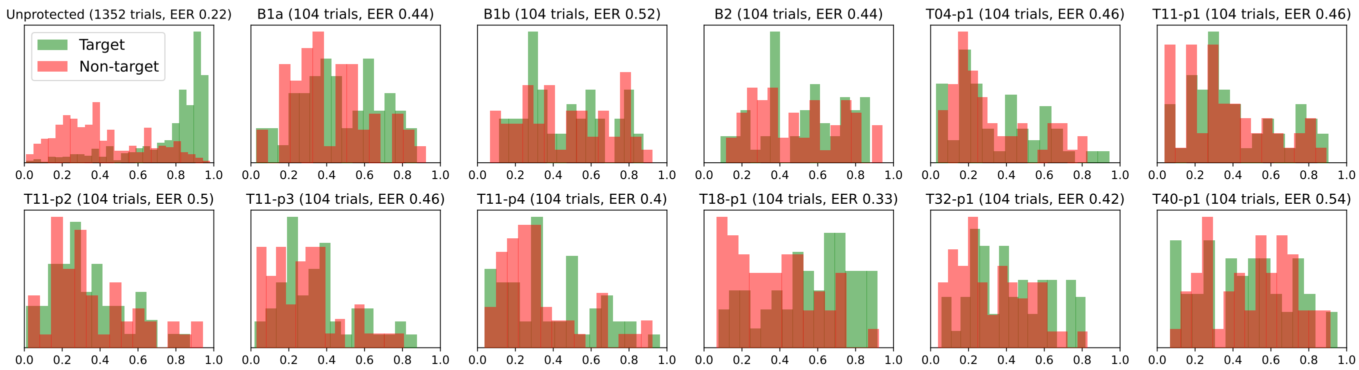


Figure 7: Distribution of target and non-target scores according to human listeners. For the top-left histogram, all data is unprotected. For all others, the title above each histogram identities the system used to anonymise trial utterances. Enrolment utterances remain unprotected.

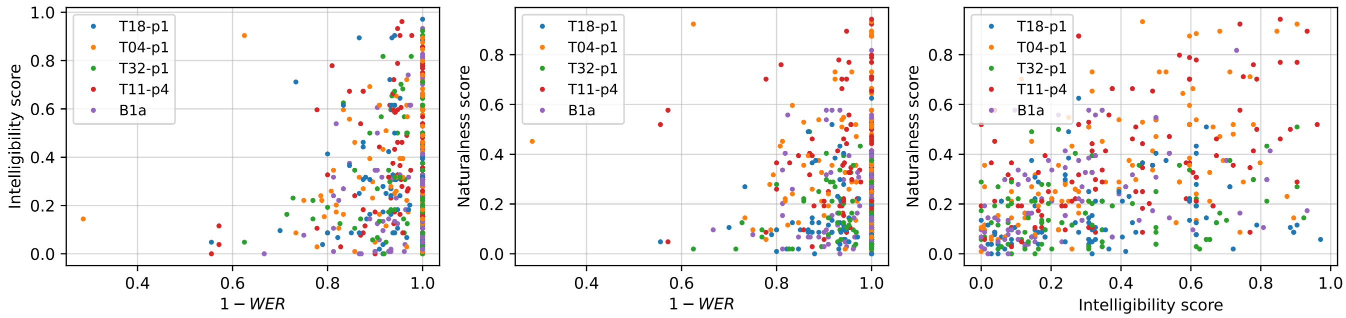


Figure 8: Scatterplots showing intelligibility score, naturalness score and the score corresponding to $1 - \text{WER}$ for utterances of different systems.

All other histograms depict distributions when trial utterances are anonymised with one of 11 anonymisation baselines and submitted systems. Enrolment utterances remain unprotected. The number of histogram bins is reduced compared to the plot for unprotected data because of the smaller number of trials. In almost all cases, the distributions for target and non-target trials are highly overlapping, indicating that listeners have greater difficulty to determine when the speaker of each utterance is the same or different. This includes distributions for the `B2` and `T32-p1` systems indicating that signal processing and deep learning based approaches to anonymisation are equally effective in the case of human listeners. EERs estimated from the scores provided by human evaluators are almost all above 40%. These estimates are, however, particularly noisy given the low number of trials, hence the reporting of histograms.

VI. Further analysis

In this section, we report additional analyses performed post evaluation.

A. Subjective versus objective estimates of utility

The results in Figure 5 show that WER estimates for some submissions are even lower than for unprotected data, implying that anonymisation actually *improves objective estimates* of intelligibility (see Section VI-D for a specific discussion of this issue). However, results in Figure 6 show that anonymisation *degrades subjective estimates* of intelligibility. It is hence of interest to explore these contradicting observations further.

Figure 8 shows a set of scatter plots which depict the correlation between *utterance-level* subjective intelligibility scores (left) and subjective naturalness scores (middle) against $1-\text{WER}$, for four different submissions and the baseline `B1.a` system. In both cases, and for all five anonymisation systems, the correlation is low; the Pearson correlation with $1-\text{WER}$ is 0.14 for intelligibility and 0.05 for naturalness. Curiously, for some utterances, $1-\text{WER} = 1$ (they are perfectly transcribed) but the corresponding subjective scores are near zero. These observations suggest that objective and subjective measures are *not functionally equivalent*, even though the WER is defined in [9] as a proxy for intelligibility. Based on this observation, it is clear that objective measures should no longer be considered as a proxy for subjective measures. Even so, both are still of interest; they are indicators of anonymisation performance for different use cases, one involving the automatic treatment of anonymised utterances by machines and, for the other, consumption by human listeners. As shown in the scatterplot to the right of Figure 8, there appears to be some degree of correlation between the two subjective measures, with their Pearson correlation coefficient being 0.58. This might indicate that, from a perceptual perspective, the concept of ‘naturalness’ is intrinsically linked to intelligibility. In the future, subjective metrics which better distinguish between the two

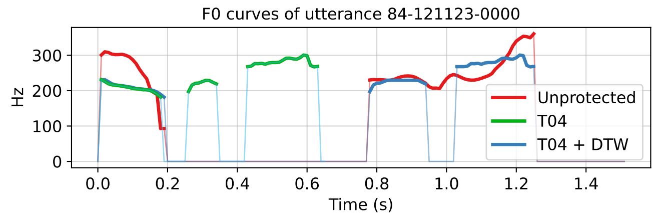


Figure 9: F0 contours for utterance 84-121123-0000 (from the *LibriSpeech-test-clean* set). Red – unprotected; green – anonymised with `T04` system; blue – anonymised with `T04` system and then aligned to the original utterance using dynamic time warping (DTW).

aspects should be considered, e.g. assessments of intelligibility might be made by asking listeners to transcribe the spoken content of both anonymized and unprocessed utterances, and comparing the results.

B. *ASR+TTS-based anonymisation*

Since its inception, VoicePrivacy was designed to encourage the development of anonymisation systems which preserve linguistic and para-linguistic speech attributes. Despite it being a design goal, a means to gauge the preservation of para-linguistic attributes was missing in 2020. This was an obvious weakness since ASR systems could be used to generate an intermediate transcription of the input before the application of TTS to generate perfectly voice-anonymised and intelligible utterances, albeit without the para-linguistic attributes of the input. While the submission of ASR+TTS systems was not prohibited, the pitch correlation metric $\boldsymbol{\rho}^{F_0}$ and minimum threshold were hence introduced for the VoicePrivacy 2022 edition in order to favour solutions (e.g. those based on voice conversion) which offer better potential for anonymisation while also preserving para-linguistic attributes.

The majority of teams pursued voice conversion-based solutions. Team `T04` was alone in exploring an ASR+TTS-like solution which, perhaps unsurprisingly, delivers near-to-perfect objective anonymisation results and among the lowest WERs for test data (see Figure 5), as well as competitive subjective naturalness and intelligibility assessment results (see Figure 6). The gain in voice distinctiveness G_{VD} is the best of all and, interestingly, the pitch correlation $\boldsymbol{\rho}^{F_0}$ also exceeds the minimum threshold. Whether ASR+TTS solutions are appropriate, whether the minimum pitch correlation threshold for VoicePrivacy 2022 was perhaps too low, or even whether pitch correlation is sufficient on its own as a measure of para-linguistic attribute preservation, are all matters of opinion. Ultimately, all are also dependent upon the specific use case scenario. Given the promising EER, WER and G_{VD}, ASR+TTS systems warrant continued attention, especially given that various techniques could be applied readily to boost the pitch correlation or the preservation of para-linguistic attributes in order to improve their competitiveness with voice conversion-based solutions. In the following, we present some initial work designed to gauge the potential.

	$\boldsymbol{\rho}^{F_0}$ **w/o DTW**	$\boldsymbol{\rho}^{F_0}$ **with DTW**
T04-p1	0.36	0.83
T11-p4	0.70	0.91
T18-p1	0.70	0.93
T32-p1	0.81	0.94
T40-p1 (old)	0.74	0.90
T40-p1 (new)	0.74	0.90

Table IV: Pitch correlation values of different anonymisation systems with and without the application of DTW.

C. *Pitch realignment*

We applied a trivial form of pitch realignment using dynamic time warping (DTW) to determine whether the pitch correlation for ASR+TTS anonymised utterances can be improved, and hence whether even higher minimum thresholds might also be met with ASR+TTS approaches. Whereas the pitch correlation $\boldsymbol{\rho}^{F_0}$ is estimated by compensating for any misalignment between anonymised and unprotected utterances using linear warping functions, DTW supports more flexible non-linear warping functions. This greater flexibility should result in improved pitch correlation. We set the DTW local continuity constraints to warp the pitch correlation of anonymised utterances onto those of corresponding unprotected utterances.

Figure 9 shows three pitch contours for the same expressive utterance *"Go! Do you hear?"*, contained within the *LibriSpeech* test set. They correspond to: the unprotected utterance (red); the `T04`-anonymised utterance (green); the corresponding DTW-aligned utterance (blue). Alignment is shown to improve greatly with the application of DTW.

We applied the same pitch realignment process to the full test set and utterances treated with one of the six systems shown in Table IV. Pitch correlation results, shown in all cases with and without the application of DTW alignment, improve markedly for all systems, and for the `T04-p1` system the most. The improvement in this case is substantial, giving values of $\boldsymbol{\rho}^{F_0}$ higher than those for all original systems without DTW alignment. Still, the result of $\boldsymbol{\rho}^{F_0} = 0.83$ is lower than that for all systems with DTW alignment. This result suggests that high values of $\boldsymbol{\rho}^{F_0}$ can be obtained with minimal additional effort, and either that the threshold of 0.3 is far too low, and/or the metric itself is deficient. We stress that we did not use the warped F0 contours to resynthesise speech signals, the naturalness and intelligibility of which would inevitably degrade. Resynthesis would require further work to warp-adjust articulatory or linguistic features, or an alternative approach to DTW, e.g. by operating upon spectral features.

D. *Utility increase with anonymisation*

The plot to the left in Figure 5 shows that some anonymisation systems lead to lower WERs than that for unprotected data. This would suggest that, contrary to intuition, anonymisation *improves* intelligibility. This would be a rather favourable and unrealistic interpretation.

The ASR_{eval} model is trained using unprotected data sourced from the *LibriSpeech-train-clean-360* dataset. Conversely, the $ASR_{\text{eval}}^{\text{anon}}$ model is trained using the anonymised version of the same dataset. The anonymisation system, however, is trained using a much larger amount of data, namely that sourced from the *LibriSpeech-train-clean-100*, *LibriSpeech-train-other-500*, and *VoxCeleb 1-2* datasets. This likely leads to the leaking of anonymisation training data into the $ASR_{\text{eval}}^{\text{anon}}$ model, implying that it is exposed to more data during training than the ASR_{eval} model. While comparisons made between two different anonymisation systems involve models trained under identical data conditions, comparisons in utility before and after anonymisation do not. Results should then be interpreted with appropriate caution. Fairer comparisons of utility before and after anonymisation might be made if the data used for the training of the the ASR_{eval} model were augmented with the same data used in the training of anonymisation systems. It should be noted, though, that this would lead to other undesirable issues concerning data overlap.

VII. Post-evaluation work and results

In this section we provide an overview of relevant, peer-reviewed work published in the open literature post evaluation.

Improvements to the **anonymisation function** based on *orthogonal Householder neural networks* (OHNNs) are reported in [42]. The layers of the OHNN consist in a linear transformation defined by an orthogonal matrix. The parameters of each layer are trained to maximise the distance between original and anonymised speaker embeddings while preserving voice distinctiveness and the overall distribution in the embedding space. Use of OHNNs results in greatly improved anonymisation performance to nearly 50% EER in a semi-informed attack scenario without loss to utility and is hence an attractive direction for future work. To the best of our knowledge, [42] is also the only post-evaluation work which investigates the effect of anonymising non-English speech using models trained on mostly English corpora.

Vocoder contributions to anonymisation, discussed in [43], can even outweigh those of the anonymisation function; the x-vector extracted from the anonymised utterance often differs substantially from the x-vector at the vocoder input. This phenomenon, referred to as *vocoder drift*, likely stems from the leakage and re-entanglement of speaker information contained in linguistic and prosodic features [44]. It is argued that poor control of the speaker embedding space can hinder the development of better anonymisation functions, implying that future work should also take vocoder effects into account.

Privacy leakage was reported earlier in [45] which shows that personally identifiable information contained in sources other than the speaker embedding can leak into anonymised speech signals. The authors show that leakage can be reduced through the application of vector quantisation to the linguistic features and argue that properly *disentangled* features are key to achieving more effective anonymisation. The EER of their baseline system sees an increase of EER from 8% to 16% when adding a quantization bottleneck on linguistic features, at the cost of a WER degradation from 7% to 10%. More elaborate disentanglement techniques, e.g. to isolate cues relating to the speaker sex [46], have been reported. These techniques will likely attract greater attention in the future; they might one day allow for multiple privacy-sensitive attributes (e.g. the voice identity and the speaker's sex and age) to be selectively and simultaneously suppressed or manipulated.

The **pitch correlation of ASR+TTS-like anonymisation** was explored further by the `T04` participants. They report [47] a technique to transfer prosodic information contained in the input utterance to the synthesised, anonymised output. They explored the extraction of phoneme duration, average pitch and average energy, not from phoneme transcriptions as in their original approach [27] but, instead, directly from the input utterance. While pitch correlation is shown to improve from 0.3 (for the original `T04` system) to 0.7, the experiments were conducted with a lazy-informed attack scenario instead of the VoicePrivacy semi-informed attack scenario.

Alternatives to x-vector speaker embeddings have also been investigated. A flow-based TTS model, adapted to perform voice conversion, is reported in [48]. The spectrogram of the original utterance is passed through an encoder and a direct flow model conditioned on the original speaker embedding. The resulting latent utterance representation is then fed into an inverse flow model conditioned on a pseudo-speaker embedding and an output waveform is resynthesised. A competitive EER of 22% is achieved with minimal utility degradation. An approach to speaker anonymisation based on auto-regressive modelling of neural audio codecs is reported in [49]. An input utterance is converted to a set of semantic and acoustic tokens using a semantic encoder and a neural audio codec respectively, and a transfomer model is then trained to predict the acoustic tokens from the semantic tokens. Voice conversion is then performed by conditioning the Transformer on acoustic tokens extracted from a different speaker. The method gives an EER of 32% and both pitch correlation and voice distinctiveness are relatively well preserved. Nonetheless, the methods in [48] and [49] both degrade utility.

The authors of system `T11` proposed a variation of baseline `B1.a` for which the x-vector speaker encoder is replaced by a formant-based identity representation [50]. BN features are computed in the same way as for the original `T11` system. The first five formants are extracted from the input utterance, and further processed by different neural-network based encoders to create a speaker representation that better preserves voice distinctiveness. F0 curves are extracted with PRAAT and the final waveform is synthesised with an acoustic model and a vocoder. The speaker identity is anonymised using a variety of techniques to scaling the F0 and formant position. The use of larger scaling values can trade stronger anonymisation for degraded utility, pitch correlation and, still, voice distinctiveness. The configuration which achieves the best voice distinctiveness ($G_{\text{VD}} \approx -4.5$) yields an EER of 21% with utility comparable to that of the original `T11` system.

VIII. Future directions

A third edition of the VoicePrivacy challenge is planned. In the following we discuss some of the priorities and likely directions.

The choice of **source data** has been queried within the community. *LibriSpeech* and *VCTK* datasets contain relatively high-quality speech data *read* from book chapters and newspaper text. Read speech lacks the spontaneity of speech collected in settings which might be more representative of practical use case scenarios. *VCTK* data was collected with a single microphone and from within a single, hemi-anechoic chamber. In contrast, it can be assumed reasonably that, for a given speaker, *LibriSpeech* data were likely collected using the same microphone and from within the same room. Its use for ASV experiments is hence potentially problematic since the use of cues related to the voice as well as the channel characteristics can be used for speaker verification. The impact of persisting channel characteristics and their potential use to infer speaker identity remains untested in the scope of VoicePrivacy and merits attention in the design of future challenges. There is also concern that the size of the datasets used for ASV training (being vastly smaller than more popular and current alternatives) is insufficient, that the resulting models are hence poorly trained, and that the use of larger datasets will influence anonymisation performance, leading to different findings.

Speaker-level anonymisation (see Section II-B) facilitates its application to multi-speaker conversations. Nonetheless, utterance-level anonymisation may be sufficient for other applications. Assessment is also arguably inconsistent with speaker-level anonymisation: performance is assessed at the utterance-level and does not take into account the advantage that an attacker can gain from knowing that anonymisation is applied at the speaker-level. Given that each utterance corresponding to the same speaker is anonymised with the same pseudo-voice, an attacker need only overcome the protection of one (perhaps weakly anonymised) utterance and then link utterances having the same pseudo-voice in order to gain an advantage in overcoming the protection of *every* other utterance produced by the same speaker. The use case and approach to assessment may hence need additional thought.

VoicePrivacy 2022 results show that **objective assessment** is **no substitute for subjective assessment**. The correlation in results for each is low; WER results for some utterances are near to one, while subjective intelligibility and naturalness scores for the same utterances are near to zero. Objective and subjective metrics are nonetheless both useful, but they reflect performance in different use cases. The work in [51] suggests that the retraining of an ASR system using anonymised data can result in mispronunciations still being correctly transcribed. WER estimates are then biased since they can mask mispronunciation errors. In the future we may have to consider distinct evaluation tracks or rankings which reflect anonymisation performance where the downstream task involves either automatic treatment of anonymised speech or consumption by human listeners (notwithstanding that some use cases may involve both). In case of the latter, it may be necessary to conduct human listening transcription tests so that mispronunciations are properly taken into account.

Challenge complexity, including that of the baseline systems and evaluation framework, may account for why the challenge did not attract greater participation. Ideas to simplify the baselines and evaluation framework involve the adoption of purely Python based solutions without reliance upon the integration of Kaldi components, and the clearer separation of baseline anonymisation components from those related to evaluation [52], [53]. With hope, these changes might attract participation from the voice conversion community.

Last, we have foreseen for some time the development of a parallel **VoicePrivacy attacker challenge**. Appropriate strength testing is essential if we are to have any confidence in anonymisation safeguards. Such an attacker challenge necessitates the collection and re-distribution of anonymised data, which would require the consent of VoicePrivacy challenge participants, in addition to the design of new evaluation protocols and, potentially, also metrics. A potential attacker challenge will likely follow the third edition of the traditional VoicePrivacy defender (anonymisation) challenge.

IX. Conclusions

VoicePrivacy 2022 attracted broad interest from the research community. Eleven valid submissions were received, all of which improved upon the challenge baselines to some degree, be it upon the primary privacy and utility measures, and/or upon other secondary measures such as voice distinctiveness or pitch correlation, etc. Given the primary motivation of fostering progress in voice anonymisation, VoicePrivacy 2022 was hence largely successful. We must nonetheless acknowledge that we remain far from achieving real, effective voice anonymisation and that, with mounting privacy regulation, the community must redouble its efforts in the future.

It is clear from the findings and our experience of VoicePrivacy 2022 that the organisation of challenges in voice anonymisation is itself *challenging*. The use cases for anonymisation parallel those for speech technology more generally (e.g. the defined *downstream* tasks). Consequently, it is difficult to design a single challenge to foster progress in voice anonymisation in a manner that is suited to them all. We must also acknowledge that voice anonymisation is but only one approach to preserve privacy in the use of speech technology and that, for some applications other, perhaps complementary techniques also warrant exploration.

With alternative approaches being extremely diverse (e.g. encryption, federated learning, differential privacy etc.), VoicePrivacy will, at least for the time being, remain focused upon voice anonymisation. In the penultimate section above, we outline our ideas for future challenge editions. Our priorities will be to develop the challenge in a way which embraces more practical, commercial or societal use cases but also to simplify the challenge where possible, from the protocols, to the metrics, evaluation conditions and baseline systems. By lowering the cost of entry into what is still an emerging and demanding field of research, we hope to encourage broader participation. Given the increase in privacy regulation worldwide, we expect interest to grow rapidly in the coming years and with future challenge editions.

References

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

[2] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "Introducing the VoicePrivacy Initiative," in *INTERSPEECH*, 2020, pp. 1693–1697.

[3] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O'Brien, A. Chanclu, J.-F. Bonastre, M. Todisco, and M. Maouche, "The VoicePrivacy 2020 Challenge: Results and findings," *Computer Speech and Language*, vol. 74, 2022.

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

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

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

[7] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," in *INTERSPEECH*, 2019, pp. 1526–1530.

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

[9] N. Tomashenko, X. Wang, X. Miao, H. Nourtel, P. Champion, M. Todisco, E. Vincent, N. Evans, J. Yamagishi, and J. F. Bonastre, "The VoicePrivacy 2022 Challenge evaluation plan," *arXiv preprint arXiv:2203.12468*, 2022.

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

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

[12] M. Maouche, B. M. L. Srivastava, N. Vauquier, A. Bellet, M. Tommasi, and E. Vincent, "Enhancing Speech Privacy with Slicing," in *INTERSPEECH*, 2022, pp. 5025–5029.

[13] A. S. Shamsabadi, B. M. L. Srivastava, A. Bellet, N. Vauquier, E. Vincent, M. Maouche, M. Tommasi, and N. Papernot, "Differentially private speaker anonymization," *Proceedings on Privacy Enhancing Technologies*, 2023.

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

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

[16] D. Hirst, "A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation. ICPhS XVI, Saabrücken," 2007.

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

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

[19] A. Rosenberg and B. Ramabhadran, "Bias and statistical significance in evaluating speech synthesis with mean opinion scores," in *INTERSPEECH*, 2017, pp. 3976–3980.

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

[21] K. Kasi and S. A. Zahorian, "Yet Another Algorithm for Pitch Tracking," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, May 2002, pp. I–361–I–364, iSSN: 1520-6149.

[22] S. Ioffe, "Probabilistic linear discriminant analysis," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 531–542.

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

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

[25] J. Kong, J. Kim, and J. Bae, "Hifi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

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

[27] S. Meyer, P. Tilli, F. Lux, P. Denisov, J. Koch, and N. T. Vu, "Cascade of phonetic speech recognition, speaker embeddings gan and multispeaker speech synthesis for the VoicePrivacy 2022 Challenge ," in *Proc. 2nd Symposium on Security and Privacy in Speech Communication*, 2022.

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

[29] F. Lux and T. Vu, "Language-agnostic meta-learning for low-resource text-to-speech with articulatory features," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6858–6868.

[30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.

[31] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *9th International Conference on Learning Representations, ICLR, Virtual Event, Austria, May 3-7*. OpenReview.net, 2021.

[32] J. Yao, Q. Wang, L. Zhang, P. Guo, Y. Liang, and L. Xie, "NWPU-ASLP System for the VoicePrivacy 2022 Challenge," in *Proc. 2nd Symposium on Security and Privacy in Speech Communication*, 2022.

[33] H.-S. Choi, J. Lee, W. Kim, J. Lee, H. Heo, and K. Lee, "Neural analysis and synthesis: Reconstructing speech from self-supervised representations," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 16 251–16 265.

[34] Z. Yao, D. Wu, X. Wang, B. Zhang, F. Yu, C. Yang, Z. Peng, X. Chen, L. Xie, and X. Lei, "WeNet: Production Oriented Streaming and Non-Streaming End-to-End Speech Recognition Toolkit," in *INTERSPEECH*, 2021, pp. 4054–4058.

[35] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[36] X. Chen, G. Li, H. Huang, W. Zhou, S. Li, Y. Cao, and Y. Zhao, "System description for Voice Privacy Challenge 2022 ," in *Proc. 2nd Symposium on Security and Privacy in Speech Communication*, 2022.

[37] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5884–5888.

[38] U. E. Gaznepoglu, A. Leschanowsky, and N. Peters, "VoicePrivacy 2022 system description: speaker anonymization with feature-matched f0 trajectories," in *Proc. 2nd Symposium on Security and Privacy in Speech Communication*, 2022.

[39] C. O. Mawalim, S. Okada, and M. Unoki, "Speaker anonymization by pitch shifting based on time-scale modification," in *Proc. 2nd Symposium on Security and Privacy in Speech Communication*, 2022, pp. 35–42.

[40] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech communication*, vol. 9, no. 5-6, pp. 453–467, 1990.

[41] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323–332, 1999.

[42] X. Miao, X. Wang, E. Cooper, J. Yamagishi, and N. Tomashenko, "Speaker anonymization using orthogonal householder neural network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3681–3695, 2023.

[43] M. Panariello, M. Todisco, and N. Evans, "Vocoder drift in x-vector–based speaker anonymization," in *INTERSPEECH*. ISCA, Aug. 2023, pp. 2863–2867.

[44] ——, "Vocoder drift compensation by x-vector alignment in speaker anonymisation," in *Proc. 3rd Symposium on Security and Privacy in Speech Communication*, 2023, pp. 16–20.

[45] P. Champion, A. Larcher, and D. Jouvet, "Are disentangled representations all you need to build speaker anonymization systems?" in *INTERSPEECH*. ISCA, Sept. 2022, pp. 2793–2797.

[46] P.-G. Noé, X. Miao, X. Wang, J. Yamagishi, J.-F. Bonastre, and D. Matrouf, "Hiding speaker's sex in speech using zero-evidence speaker representation in an analysis/synthesis pipeline," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[47] S. Meyer, F. Lux, J. Koch, P. Denisov, P. Tilli, and N. T. Vu, "Prosody Is Not Identity: A Speaker Anonymization Approach Using Prosody Cloning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2023, pp. 1–5.

[48] F. Nespoli, D. Barreda, J. Bitzer, and P. A. Naylor, "Two-Stage Voice Anonymization for Enhanced Privacy," in *INTERSPEECH*. ISCA, Aug. 2023, pp. 3854–3858.

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

[50] J. Yao, Q. Wang, Y. Lei, P. Guo, L. Xie, N. Wang, and J. Liu, "Distinguishable Speaker Anonymization Based on Formant and Fundamental Frequency Scaling," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2023, pp. 1–5.

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

[52] S. Meyer, X. Miao, and N. T. Vu, "Voicepat: An efficient open-source evaluation toolkit for voice privacy research," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 257–265, 2024.

[53] C. Franzreb, T. Polzehl, and S. Möller, "A Comprehensive Evaluation Framework for Speaker Anonymization Systems," in *Proc. 3rd Symposium on Security and Privacy in Speech Communication*, 2023, pp. 65–72.