Evaluating Voice Conversion-based Privacy Protection against Informed Attackers

Brij Mohan Lal Srivastava, Nathalie Vauquier, Md Sahidullah, Aurélien Bellet, Marc Tommasi, Emmanuel Vincent

EVALUATING VOICE CONVERSION-BASED PRIVACY PROTECTION AGAINST INFORMED ATTACKERS

Brij Mohan Lal Srivastava[1], *Nathalie Vauquier*[1], *Md Sahidullah*[3], *Aurélien Bellet*[1], *Marc Tommasi*[2], *Emmanuel Vincent*[3]

[1]INRIA, France [2]Université de Lille, France
[3]Université de Lorraine, CNRS, Inria, Loria, F-54000 Nancy, France

ABSTRACT

Speech data conveys sensitive speaker attributes like identity or accent. With a small amount of found data, such attributes can be inferred and exploited for malicious purposes: voice cloning, spoofing, etc. Anonymization aims to make the data *unlinkable*, i.e., ensure that no utterance can be linked to its original speaker. In this paper, we investigate anonymization methods based on voice conversion. In contrast to prior work, we argue that various *linkage* attacks can be designed depending on the attackers' knowledge about the anonymization scheme. We compare two frequency warping-based conversion methods and a deep learning based method in three attack scenarios. The utility of converted speech is measured via the word error rate achieved by automatic speech recognition, while privacy protection is assessed by the increase in equal error rate achieved by state-of-the-art i-vector or x-vector based speaker verification. Our results show that voice conversion schemes are unable to effectively protect against an attacker that has extensive knowledge of the type of conversion and how it has been applied, but may provide some protection against less knowledgeable attackers.

***Index Terms*—** privacy, voice conversion, speech recognition, speaker verification, linkage attack

1. INTRODUCTION

Speech is a behavioural biometric characteristic of human beings [1], which can produce distinguishing and repeatable biometric features. Dramatic improvements in speech synthesis [2], voice cloning [3, 4] and speaker recognition [5] that leverage "found data" pose severe privacy threats to the users of speech interfaces [6]. According to the ISO/IEC International Standard 24745 on biometric information protection [7], biometric references must be *irreversible* and *unlinkable* for full privacy protection. Anonymization or de-identification [8–10] refers to the task of concealing the speaker's identity while retaining the linguistic content, thereby making the data *unlinkable* [11]. In this work, we consider the following threat model: given a public dataset of (supposedly) anonymized speech, an attacker records/finds a sample of speech of a target user and attempts to guess which utterances in the anonymized dataset are spoken by the target user. A good anonymization scheme should prevent such *linkage attacks* from being successful, while preserving the perceived speech naturalness and intelligibility and/or the performance of downstream tasks such as automatic speech recognition (ASR).

Fang et al. [12] classify speaker anonymization methods into two categories: *physical* vs. *logical*. Physical methods perturb speech in the physical space by adding acoustic noise, while logical methods apply a transformation to the recorded signal. Among the latter, voice conversion (VC) methods have been traditionally exploited as a way to map the input voice (*source*) into that of another speaker (*target*) [13–15]. In contrast to feature-domain approaches [10], the output of VC remains a speech waveform and it may be used for listening or transcription purposes. The anonymized speech should thus sound as natural and intelligible as possible [16].

Crucially, all past studies assumed a weak attack scenario where the attacker is unaware that an anonymization method has been applied to the found data [12]. This raises the concern that the privacy protection may entirely rely on the secrecy of the design and implementation of the anonymization scheme, a principle known as "security by obscurity" [17] that has long been rejected by the security community. There is therefore a strong need to evaluate the robustness of the anonymization to the knowledge that the adversary may have about the transformation. In practice, such knowledge may for instance be acquired by inspecting the code embedded in the user's device or in an open-source implementation.

As opposed to past studies, we consider different linkage attacks depending on the attacker's knowledge of the anonymization method. At one end of the spectrum, an *Ignorant* attacker is unaware of the speech transformation being applied, while at the other end an *Informed* attacker can leverage complete knowledge of the transformation algorithm. A *Semi-Informed* attacker may know the voice transformation algorithm but not its parameter values. In our experiments, we evaluate three VC methods with different target speaker selection strategies in various attack scenarios to study unlinkability in the spirit of ISO/IEC 30136 standard [18]. In each scenario, we assess how well each method protects the speaker identity against attackers that leverage state-of-the-art speaker verification techniques based on i-vectors [19] or x-vectors [5] to design linkage attacks. We also report the *word error rate* (WER) achieved by a state-of-the-art end-to-end automatic speech recognizer [20]. While a formal listening test is beyond the scope of this paper, we make a few samples of converted speech available for informal comparison.[1]

In Section 2, we describe the three VC methods we evaluate in the context of anonymization. Section 3 introduces the target speaker selection strategies and the attack scenarios. Section 4 presents the experimental settings and the results. We conclude in Section 5.

This work was supported in part by the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement No. 825081 COMPRISE (https://www.compriseh2020.eu/) and by the French National Research Agency under project DEEP-PRIVACY (ANR-18-CE23-0018). Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

[1]https://github.com/brijmohan/adaptive_voice_conversion/tree/master/samples

2. VOICE CONVERSION METHODS

The criteria for selecting the VC methods in our study are that they must be **1) non-parallel**, i.e., do not require a parallel corpus of sentences uttered by both the source and target speakers for training — this is important from a privacy perspective since there exist few parallel corpora and selecting openly available targets would increase the risk of an inversion attack; **2) many-to-many**, i.e., allow conversion between arbitrary sources and targets so that any speaker in a large corpus can be selected as the target ; **3) source- and language-independent**, i.e., do not require enrollment sentences for the source speaker and do not rely on language-specific ASR or phoneme classification — this is important from a usability perspective as it frees the user from the burden of enrolling and it is applicable to any language (including under-resourced ones), and from a privacy perspective since enrollment translates into the storage of a voiceprint which poses even greater privacy threats.

The third criterion is quite strict: many VC methods, such as StarGAN-VC [21] or the ASR-based method in [12], do not satisfy it. We found that the vocal tract length normalization (VTLN) based methods in [15, 22] and the one-shot method in [23] satisfy all criteria. In this paper, we use models trained over English speech [24] but do not use any other linguistic resources such as transcriptions.

2.1. VoiceMask

VoiceMask is described in [15] as the frequency warping method based on the composition of a log-bilinear function, expressed as $f(\omega, \alpha) = |-i \ln \frac{z-\alpha}{1-\alpha z}|$, and a quadratic function, given by $g(\omega, \beta) = \omega + \beta(\frac{\omega}{\pi} - (\frac{\omega}{\pi})^2)$. Here $\omega \in [0, \pi]$ is the normalized frequency, $\alpha \in [-1, 1]$ is the warping factor for the bilinear function, $z = \mathrm{e}^{i\omega}$, and $\beta > 0$ is the warping factor for the quadratic function. Therefore, the warping function is of the form $g(f(\omega, \alpha), \beta)$. The two parameters, α and β, are chosen uniformly at random from a predefined range which is found to produce intelligible speech while perceptually concealing the speaker identity. In the following, we apply this transform to the spectral envelope rather than the pitch-synchronous spectrum as in the original paper. In addition, we apply logarithm Gaussian normalized pitch transformation (see [25]) so as to match the pitch statistics of a target speaker[2].

The authors claim that this transformation is difficult to inverse when the parameter values are unknown because they are randomly selected from a large interval. However, VoiceMask uses the same parameter values to warp the spectra at each time step of the utterance. This approach is quite limited to conceal the identity of the source speaker and to mimic the target speaker because it warps the entire frequency axis in a single direction.

2.2. VTLN-based voice conversion

VTLN-based VC [22] represents each speaker by a set of centroid spectra extracted using the CheapTrick [26] algorithm for k pseudo-phonetic classes. These classes are learned in an unsupervised fashion by clustering all speech frames of all utterances from this speaker. For each class of the source speaker, the procedure finds the class of the target speaker and the warping parameters that minimize the distance between the transformed source centroid spectrum and the target centroid spectrum. All speech frames in that class are then warped using a power function. Similarly to above, we apply this warping to the spectral envelope and also perform Gaussian normalized pitch transformation so as to match the pitch statistics of the target. Compared to VoiceMask, this approach warps the frequency axis in different directions over time. The parameters of this method include the number of classes k and the chosen target speaker.

[2]Strictly speaking, VoiceMask is a voice transformation method rather than a VC method: pitch is converted from the source speaker to a target speaker, but the spectral envelope is not related to a particular target speaker.

2.3. Disentangled representation based voice conversion

The third approach is based on disentangled representation of speech as proposed in [23, 27]. The core idea is that speaker information is statically present throughout the utterance but content information is dynamic. This approach is based on a neural network transformation and uses a *speaker encoder* and a *content encoder* to separate the factors of variation corresponding to speaker and content information. The only parameter of this method is the chosen target speaker.

3. TARGET SELECTION STRATEGIES AND ATTACKERS

In this study, we consider that the VC function and the sets of possible parameter values are known to all users. Each user records his/her voice on his/her device and applies a VC scheme locally before sending it to a public database. In the threat model we consider, an attacker then performs a linkage attack to try to identify which converted utterances in this public database are spoken by a particular user. To this end, we assume that the attacker has access to a small amount of found speech from this user (and potentially some additional public resources, such as benchmark speech processing datasets to train generic speaker models).

In the following, we define three parameter selection (a.k.a. target selection) strategies for the three VC methods above, which can be seen as key ingredients of a "private-by-design" speech processing system. We then describe the knowledge that an attacker trying to compromise the system could have about the VC function and the target selection strategy.

3.1. Target selection strategies

We consider three possible target selection strategies. In strategy *const*, the VC function is constant across all users and all utterances. This means choosing a unique target speaker and, in the case of VoiceMask, fixed values for α and β. In strategy *perm*, the conversion parameters are chosen at random once by each user. In other words, when a user downloads the VC module on his/her device, he/she selects a personal target speaker and, in the case of VoiceMask, personal random values for α and β. Finally, in the *random* strategy, each time a user applies VC to an utterance, a random set of parameters is drawn, i.e., a random target speaker is selected and, in the case of VoiceMask, random values are drawn for α and β.

3.2. Attackers' knowledge

We define the types of attackers based on the extent of their knowledge about the VC function and its parameters. An *Ignorant* attacker is not aware that VC has been applied at all. In contrast, an *Informed* attacker knows the VC method and its exact parameter values (i.e., the chosen target speaker and the values of α and β). One may argue that an *Informed* attacker is not very realistic (except for the *const* strategy), while an *Ignorant* attacker is very weak. Between these two extreme cases, various types of attackers can be defined. For instance, we consider a *Semi-Informed* attacker who knows the chosen VC method (VoiceMask, VTLN, or disentangled representation) and the target selection strategy (*const*, *perm*, or *random*), but not the

actual target (i.e., the actual target speaker or the value of α and β). This is arguably more realistic since the VC algorithm and the target selection strategy may be open-source, while (except for the *const* strategy) the target chosen by the user is much less easily accessible.

It is important to note that many concrete instances of attackers of the above types can be designed, and finding out the "best" attacker of a particular type is a hard problem. In the experiments section, we propose attackers exploiting these different levels of knowledge based on the assumptions defined above. A more exhaustive investigation of the design of attackers is left for future work.

4. EXPERIMENTS AND RESULTS

4.1. Data and evaluation setup

All experiments are performed on the LibriSpeech corpus [24]. We use the 460 h clean training set (*train-clean-100* + *train-clean-360*), which contains 1,172 speakers, to train the disentanglement transform. Out of the *test-clean* set, we create an *enrollment* set (438 utterances) and a *trial* set (1,496 utterances) with different utterances from the same 29 speakers (13 male and 16 female, not in the training set) considered as source speakers. The target speakers for all three VC methods are randomly picked from the training and *test-clean* sets. See [10] for more details.

For each VC method and target selection strategy, all utterances in the trial set are mapped to possibly different target speakers in the training or trial set. The converted trial set serves as the public database that attackers want to de-anonymize by designing a linkage attack. To this end, attackers have access to the enrollment set which serves as the found data used to model the speakers in the trial set.

The attackers also have access to the 460 h training set to train state-of-the-art speaker verification methods based on x-vectors [5] and i-vectors, which are stronger than the Gaussian mixture model-universal background model (GMM-UBM) based method used in the seminal work of [16]. We adapt the *sre16* Kaldi recipe for training x-vectors and i-vectors to LibriSpeech[3]. We use a smaller network architecture for x-vector computation than the original recipe. Specifically, compared to the architecture in [5, Table 1], we remove the *frame4*, *frame5* and *segment7* layers, thereby also reducing the *stats pooling* layer to $512T\times1024$ and the *segment6* layer to 1024×512. Here T refers to the utterance-level context. This reduced architecture performs slightly better on LibriSpeech than the architecture in the original recipe. We give more details on the different attackers in Section 4.3.

Finally, we evaluate the utility of each VC method in terms of the resulting ASR performance on the converted data. We use a hybrid connectionist temporal classification (CTC) and attention based encoder-decoder [20] trained on the converted 460 h training set using the standard recipe for LibriSpeech provided in ESPnet[4].

4.2. Voice conversion settings

VoiceMask. Pitch, aperiodicity and spectral envelope are extracted using the pyworld vocoder[5]. We follow strategy *random* only. We sample α uniformly such that $|\alpha| \in [0.08, 0.10]$ then β in $[-2, 2]$ such that $0.32 \leq dist_{f_{\alpha,\beta}} \leq 0.40$ where $dist_{f_{\alpha,\beta}} = \int_0^\pi |f_{\alpha,\beta}(\omega) - \omega|$ is the distortion strength of the warping function. These ranges are provided by VoiceMask's authors in [15] since they produce most intelligible output. A subset of 100 target speakers is randomly selected and, for every utterance, pitch is transformed so as to match a random speaker within that subset. Other target selections strategies have not been applied because fixed values for α and β (whether speaker-dependent or not) are prone to inversion attacks.

VTLN-based VC. Pitch, aperiodicity and spectral envelope are extracted using the pyworld vocoder. For each speaker, we collect speech frames using energy-based voice activity detection (VAD) with a threshold of 0.06, and we cluster their spectral envelopes via k-means with $k = 8$. In strategy *const*, only one target speaker is selected. In *perm*, we draw a random subset of 100 target speakers and, for each source speaker, we select a random target within it. In *random*, we draw a random subset of 100 target speakers and, for each source utterance, we select a random target within it.

Disentangled representation based VC. We use a publicly available implementation of this method[6]. As per the authors' suggestion in the preprocessing script, we train the disentanglement models (speaker encoder, content encoder, decoder) over the *train-clean-100* subset of the LibriTTS corpus (itself a subset of the 460 h training set of LibriSpeech), with a batch size of 128 and learning rate of 0.0005 for 500,000 iterations. All three target selection strategies are applied similarly to VTLN-based VC except that only the source utterance and one random utterance from the target speaker are used as inputs to the content and speaker encoders, respectively. Other utterances from the source and targets speakers are unused.

4.3. Attackers

We have implemented several attackers depending on the choice of the VC algorithm and the target selection strategy as well as the extent of the attacker's knowledge (*Informed*, *Semi-Informed* or *Ignorant*). Our *Ignorant* attacker is unaware of the VC step: he/she simply trains x-vector/i-vector models on the untransformed training set, and applies them to the untransformed enrollment set. Our *Semi-Informed* attacker knows the VC algorithm and the target selection strategy (*const*, *random* or *perm*) but not the particular choices of targets. He/she applies this strategy to the training and enrollment sets by drawing random target speakers from the subset of 100 target speakers used by the VC method (we assume that the value of k in VTLN is known to the attacker). As a result, the training and enrollment data are converted in a similar way as the trial data, but the target speaker associated with every speaker in the enrollment set is typically different from that associated with the same speaker in the converted trial set. Finally, our *Informed* attacker has access to the actual VC models and target choices used to anonymize the trial set, so it converts the training and enrollment sets accordingly.

In our preliminary experiments, we also considered attackers who convert the enrollment set only and use x-vector/i-vector models trained on the untransformed training set. Unsurprisingly, we found that this leads to significantly larger *equal error rates* (EER) than retraining the x-vector/i-vector model (which can easily be done by the attacker using public benchmark data). Therefore, we do not report results for such attackers below.

4.4. Results and discussion

We first train and apply the ASR and speaker verification systems on the original (untransformed) data for baseline performance. We

[3] https://github.com/brijmohan/kaldi/tree/master/egs/librispeech_spkv/v2

[4] https://espnet.github.io/espnet/

[5] https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder

[6] https://github.com/jjery2243542/adaptive_voice_conversion

Table 1. EER (%) achieved using x-vector based speaker verification.

	VoiceMask	**VTLN-based VC**			**Disentangl.-based VC**		
Attackers ↓ / Strategies →	*random*	*const*	*perm*	*random*	*const*	*perm*	*random*
Informed	5.01	4.71	3.91	6.32	4.71	0.20	5.52
Semi-Informed	-	12.84	23.37	6.32	13.64	43.03	5.42
Ignorant	28.69	24.27	30.99	27.38	27.68	32.20	30.59

Table 2. EER (%) achieved using i-vector based speaker verification.

	VoiceMask	**VTLN-based VC**			**Disentangl.-based VC**		
Attackers ↓ / Strategies →	*random*	*const*	*perm*	*random*	*const*	*perm*	*random*
Informed	8.22	6.22	10.23	9.84	4.71	0.20	11.03
Semi-Informed	-	18.25	31.49	18.76	15.65	43.93	10.53
Ignorant	50.55	26.08	49.15	49.15	49.95	47.74	49.85

Table 3. WER (%) achieved using end-to-end ASR.

	VoiceMask	**VTLN-based VC**			**Disentangl.-based VC**		
Subset ↓ / Strategies →	*random*	*const*	*perm*	*random*	*const*	*perm*	*random*
test-clean	18.1	19.8	18.4	15.9	41.5	23.7	115.1

obtain an EER of 4.61% and 4.31% for i-vector and x-vector, respectively, and a WER of 9.4% for ASR.

Tables 1 and 2 present the EER for x-vector and i-vector based speaker verification for the three attackers and the various VC methods and target selection strategies. Interestingly, the *Informed* attacker achieves similar or even slightly lower EER than the baseline. This indicates that, when the attacker has complete knowledge of the VC scheme and target speaker mapping, none of the VC methods is able to protect the speaker identity. While an attacker with such complete knowledge is not very realistic in most practical cases, our results show that speaker information has not been totally removed and is somehow still present in the converted speech.

For the more realistic *Semi-Informed* attacker, we observe that strategy *perm* is quite effective in protecting privacy and shows the highest gains in EER. This is due to the fact that the target speaker in the enrolled data may not be same as the one in trial, hence greater confusion is induced during inference. We also notice that strategy *random* is not much affected by the change of speaker mapping, which is intuitive because in this case the utterances are already being mapped randomly to different speakers. Such mapping would be ineffective due to averaging of randomness. Strategy *const* is also slightly affected by the change of mapping because the training and enrollment speaker is not same as that of test speaker, but the effect is not as significant as strategy *perm*.

Consistently with past results in the literature, the *Ignorant* attacker performs worst in terms of EER. This confirms that, when the attacker is oblivious to the privacy-preserving mechanism, we can protect speaker identity completely. Figure 1 shows the distribution of i-vector PLDA scores for genuine and impostor trials, i.e., the log-likelihood ratios between *same-speaker* and *different-speaker* hypotheses. For full unlinkability, the distributions of genuine and impostor scores must be identical. We observe that the overlap between the two distributions decreases as we move from the *Ignorant* to the *Informed* attacker, hence increasing linkability.

Table 3 gives the WER obtained for each VC method, which we use as a proxy for the usefulness of the converted speech. Note that there is no difference between converted data in different attack scenarios, hence the WER does not depend on the attacker. VoiceMask and VTLN-based VC achieve reasonable WER compared to the untransformed data, while the disentangled representation based VC produces unreasonably high WER. Note that these WERs are achieved when ASR is trained solely using converted data. In practice, many techniques can be used optimize the WER, such as using converted data to augment clean data.

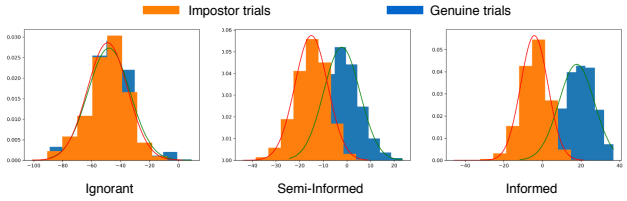

Fig. 1. I-vector score distribution for trials conducted on VTLN (strategy *random*) converted data by *Ignorant*, *Semi-Informed*, or *Informed* attackers. The orange distribution indicates impostor scores, while the blue distribution indicates genuine scores. The crossing between the two curves indicates the threshold for EER. More overlap means greater confusion, hence greater privacy protection.

5. CONCLUSION AND FUTURE WORK

We investigated the use of VC methods to protect the privacy of speakers by concealing their identity. We formally defined target speaker selection strategies and linkage attack scenarios based on the knowledge of attacker. Our experimental results indicate that both aspects play an important role in the strength of the protection. Simple methods such as VTLN-based VC with appropriate target selection strategy can provide reasonable protection against linkage attacks with partial knowledge.

Our characterization of strategies and attack scenarios opens up several avenues for future research. To increase the naturalness of converted speech, we can explore intra-gender VC as well as the use of a supervised phonetic classifier in VTLN. We also plan to conduct experiments with a broader range of attackers and use standard local and global unlinkability metrics [11] to precisely evaluate the privacy protection in various scenarios. More generally, designing a privacy-preserving transformation which induces a large overlap between genuine and impostor distributions even in the *Informed* attack scenario remains an open question. In the case of disentangled representations, this calls for avoiding any leakage of private attributes into the content embeddings.

6. REFERENCES

[1] Anil Jain, Lin Hong, and Sharath Pankanti, "Biometric identification," *Communications of the ACM*, vol. 43, no. 2, pp. 90–98, 2000.

[2] Éva Székely, Gustav Eje Henter, Jonas Beskow, and Joakim Gustafson, "Spontaneous conversational speech synthesis from found data," in *Proc. INTERSPEECH*, 2019, pp. 4435–4439.

[3] Ville Vestman, Tomi Kinnunen, Rosa González Hautamäki, and Md Sahidullah, "Voice mimicry attacks assisted by automatic speaker verification," *Computer Speech & Language*, vol. 59, pp. 36–54, 2020.

[4] Jaime Lorenzo-Trueba, Fuming Fang, Xin Wang, Isao Echizen, Junichi Yamagishi, and Tomi Kinnunen, "Can we steal your vocal identity from the internet?: Initial investigation of cloning Obama's voice using GAN, WaveNet and low-quality found data," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2018, pp. 240–247.

[5] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.

[6] Tomi Kinnunen, Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. INTERSPEECH*, 2017, pp. 2–6.

[7] Document ISO/IEC 24745:2011, "Information Technology—Security techniques—Biometric Information Protection," *ISO/IEC JTCI SC27 Security Techniques*, 2011.

[8] Andreas Nautsch et al., "Preserving privacy in speaker and speech characterisation," *Computer Speech & Language*, vol. 58, pp. 441–480, 2019.

[9] Fahimeh Bahmaninezhad, Chunlei Zhang, and John HL Hansen, "Convolutional neural network based speaker de-identification.," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2018, pp. 255–260.

[10] Brij Mohan Lal Srivastava, Aurélien Bellet, Marc Tommasi, and Emmanuel Vincent, "Privacy-preserving adversarial representation learning in ASR: Reality or illusion?," in *Proc. INTERPSPEECH*, 2019, pp. 3700–3704.

[11] Marta Gomez-Barrero, Javier Galbally, Christian Rathgeb, and Christoph Busch, "General framework to evaluate unlinkability in biometric template protection systems," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 6, pp. 1406–1420, 2017.

[12] Fuming Fang, Xin Wang, Junichi Yamagishi, Isao Echizen, Massimiliano Todisco, Nicholas Evans, and Jean-Francois Bonastre, "Speaker anonymization using x-vector and neural waveform models," in *Proc. 10th ISCA Speech Synthesis Workshop*, 2019, pp. 155–160.

[13] Slobodan Ribaric, Aladdin Ariyaeeinia, and Nikola Pavesic, "De-identification for privacy protection in multimedia content: A survey," *Signal Processing: Image Communication*, vol. 47, pp. 131–151, 2016.

[14] Miran Pobar and Ivo Ipšić, "Online speaker de-identification using voice transformation," in *Proc. 37th International convention on information and communication technology, electronics and microelectronics (mipro)*, 2014, pp. 1264–1267.

[15] Jianwei Qian, Haohua Du, Jiahui Hou, Linlin Chen, Taeho Jung, and Xiang-Yang Li, "Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity," in *Proc. the 16th ACM Conference on Embedded Networked Sensor Systems*. ACM, 2018, pp. 82–94.

[16] Qin Jin, Arthur R Toth, Tanja Schultz, and Alan W Black, "Speaker de-identification via voice transformation," in *Proc. ASRU*, 2009, pp. 529–533.

[17] Rebecca T Mercuri and Peter G Neumann, "Security by obscurity," *Communications of the ACM*, vol. 46, no. 11, pp. 160, 2003.

[18] ISO/IEC FDIS 30136, "Information Technology—Performance Testing of Biometric Protection Schemes," *ISO/IEC JTCI SC37 Biometrics*, 2017.

[19] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[20] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proc. INTERPSPEECH*, 2018, pp. 2207–2211.

[21] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *Proc. Spoken Language Technology Workshop (SLT)*, 2018, pp. 266–273.

[22] David Sundermann and Hermann Ney, "VTLN-based voice conversion," in *Proc. 3rd IEEE International Symposium on Signal Processing and Information Technology*, 2003, pp. 556–559.

[23] Ju chieh Chou and Hung-Yi Lee, "One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization," in *Proc. INTERSPEECH*, 2019, pp. 664–668.

[24] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[25] Kun Liu, Jianping Zhang, and Yonghong Yan, "High quality voice conversion through phoneme-based linear mapping functions with STRAIGHT for Mandarin," in *Proc. Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, 2007, vol. 4, pp. 410–414.

[26] Masanori Morise, "CheapTrick, a spectral envelope estimator for high-quality speech synthesis," *Speech Communication*, vol. 67, pp. 1–7, 2015.

[27] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *Proc. CVPR*, 2017, pp. 6924–6932.