# Hiding Speaker's Sex in Speech Using Zero-Evidence Speaker Representation in an Analysis/Synthesis Pipeline

Paul-Gauthier Noé, Xiaoxiao Miao, Xin Wang, Junichi Yamagishi,
Jean-François Bonastre, Driss Matrouf

# HIDING SPEAKER'S SEX IN SPEECH USING ZERO-EVIDENCE SPEAKER REPRESENTATION IN AN ANALYSIS/SYNTHESIS PIPELINE

*Paul-Gauthier Noé*[1], *Xiaoxiao Miao*[2], *Xin Wang*[2]
*Junichi Yamagishi*[2], *Jean-François Bonastre*[1], *Driss Matrouf*[1]

[1]Laboratoire Informatique d'Avignon, Avignon Université, France
[2]National Institute of Informatics, Japan

## ABSTRACT

The use of modern vocoders in an analysis/synthesis pipeline allows us to investigate high-quality voice conversion that can be used for privacy purposes. Here, we propose to transform the speaker embedding and the pitch in order to hide the sex of the speaker. ECAPA-TDNN-based speaker representation fed into a HiFiGAN vocoder is protected using a neural-discriminant analysis approach, which is consistent with the zero-evidence concept of privacy. This approach significantly reduces the information in speech related to the speaker's sex while preserving speech content and some consistency in the resulting protected voices.

***Index Terms***— attribute privacy, voice conversion, privacy preservation, sex-neutral voice

## 1. INTRODUCTION

Privacy considerations are rising in speech technology research [1]. The VoicePrivacy challenge [2] focuses on hiding the full identity of the speaker. However, there might be situations where the user requires the protection of only one or a few of their personal attributes, for instance, their sex, native language, emotional or health state. This approach to privacy is known as *user configurable* [3] and *attribute driven* privacy [4]. In this work, we focus on sex as an attribute to hide in speech signals.

Recently, methods have been proposed for this purpose. In [5], in order to avoid sex-related bias in speech model training, the authors proposed making speech *sex-neutral* beforehand by automatically searching for pitch and formants shifting that would lead to the maximum uncertainty sex classifier score, i.e., 50%. However, there is no guarantee that the classifier they used is well calibrated [6], and a search for shifting parameters must be done for each utterance. Here, we prefer to have a single transformation that can be applied regardless of the input utterance, which appears to us more suitable for a real-life application of privacy systems. In [7], the authors proposed removing the speaker's sex using an adversarial approach. Their approach also aims to protect the speaker's identity instead of leaving the speaker's other information unchanged.

Here, we want to alter only the speaker's sex while preserving the other speaker-related variabilities. We therefore consider the explicit disentanglement of sex information as a desirable step. In [4], we proposed a neural-discriminant-analysis-based approach for disentanglement. The sex variable is represented as a log-likelihood-ratio (LLR) that can be set to zero for privacy, which is consistent with the zero-evidence recognition framework [8] and Shannon's perfect secrecy [9]. However, this approach has been designed for vector inputs and applied to speaker embeddings. Extending it to waveforms is challenging, and we therefore want, as a first step, to include it in an analysis/synthesis framework for voice conversion for sex protection. We use the HiFiGAN vocoder [10] fed by the $f_0$ trajectory, a HuBERT soft content representation [11], and an ECAPA-TDNN [12] speaker embedding. Once the analysis/synthesis pipeline is trained, we apply the protection proposed in [4] to the speaker vector and an affine transformation to the $f_0$ to remove sex-related information. In our experiments, we test the protection ability of our approach in terms of sex recognition performance with both *ignorant* and *semi-informed* attacks [13], respectively considered *weak* and *strong* attacks. We evaluate the performance of automatic speech recognition (ASR) and speaker verification (ASV) as downstream tasks. Listening tests are also done in order to assess human ear perception of protected speech.

Randomly assigning a target sex to each speaker could lead to better protection results using our evaluation protocol. However, we want to inform the reader that our concept of privacy is not to fool the attacker but rather to not provide any evidence about speaker's sex, resulting in some kind of sex-neutral voice[1].

## 2. ATTRIBUTE PRIVACY AND ZERO-EVIDENCE

Most of the approaches in voice conversion for privacy aim to hide the full speaker identity [1, 2]. Attribute privacy aims instead to hide only one or a few attributes of the speaker [4], making it possible to look for a better compromise between utility and *user configurable* privacy [3]. The attributes can be personal information such as the speaker's sex, emotional and health state and so on. The attacker's knowledge on an attribute he or she wants to infer is represented by a discrete probability distribution over the possible outcomes (male and female in our case). The Bayes' rule provides a natural way to update the attacker's belief in light of observed data. For perfect secrecy/privacy, posterior and prior knowledge has to be the same [9], which corresponds to a LLR equal to zero; this is *zero-evidence* [8].

In attribute privacy, the attributes to conceal have a relatively low number of possible outcomes. For instance, if the speaker appears to be a French native and an attacker wants to infer from which region the speaker comes from, the number of possible outcomes is 18, i.e.,

This work was done when Paul-Gauthier Noé was visiting Yamagishi Laboratory, National Institute of Informatics, Japan. It was supported by the VoicePersonae project ANR-18-JSTS-0001, JST CREST Grants (JPMJCR18A6 and JPMJCR20D3), MEXT KAKENHI Grants (21K17775, 21H04906, 22K21319), and Google AI for Japan program.

[1]Audio samples and model are available at `https://github.com/nii-yamagishilab/speaker_sex_attribute_privacy`

the number of administrative regions in France. In this case, where the number of classes is significantly lower than the dimensionality of the data, the latter can be transformed such that a group of components embeds the attribute-related variability, while the other components contain the residual variability. Once this separation has been done, the attribute variability can be annihilated for privacy. This approach differs from common privacy tasks where the number of classes to make indistinguishable can be arbitrarily large.

In [4], we proposed a nonlinear discriminant analysis that allows for manipulating the LLR related to the sex attribute in speaker representation. For privacy, the LLR can be set to zero, which is consistent with *zero-evidence*. However, in its current design, this approach cannot be applied to raw speech data because speech is time dynamical. In the next section, we propose a way to get around this problem by protecting intermediate features in an analysis/synthesis pipeline instead of trying to sanitize raw speech directly.

## 3. PROPOSED PROTECTION SYSTEM

Analysis/synthesis is the process of extracting speech features from which the original speech signal can be recovered using a vocoder. In speech technology, this approach has been widely used. The intermediate characterisation of speech can be used for speech transmission, voice conversion, speech anonymisation... In speech anonymisation, we want a part of the intermediate features to represent speaker-related information that can be manipulated for privacy. In [14], the authors proposed updating the first VoicePrivacy's baseline [2] by replacing the neural source-filter vocoder [15] with a HiFiGAN vocoder [10]. They also replaced the Kaldi TDNN speaker embedding [16] (xvector) with an ECAPA-TDNN speaker embedding [12] considered to be the state-of-the-art representation for ASV. They finally got rid of the acoustic model by using instead a HuBERT-based soft content representation [11]. In [14], they used this system for the VoicePrivacy task and studied its application to unseen languages. In this paper, we use this system, but we replace the speaker embedding averaging used for voice anonymisation with the discriminant-analysis-based protection in [4] and add an affine transformation of the $f_0$ trajectory for sex protection.

**Speaker representation protection:** In [14, 2], for anonymisation, the original xvector of an utterance is replaced with an average of xvectors randomly selected from a pool of speakers. Here, we want to conceal the sex of the speaker only. We propose using the discriminant-analysis-based approach presented in [4] and discussed above. We recall here in more detail how this can be used for the concealment of sex-related information in speaker embeddings. The idea is to use normalizing-flow neural-transformation [17] to learn an invertible mapping from the speaker embedding space to a base space where the class-conditional densities are carefully chosen such that only the first component embeds the sex information in the form of a LLR $\log \frac{P(x|C=0)}{P(x|C=1)}$ (where $C$ is for class, 0 for male, 1 for female). When the LLR is zero, the observation $x$ is equally likely to come from both classes, resulting in no change in the belief of the observer/attacker. Therefore, for protection, the observed embedding is mapped into the base space where the first dimension (LLR) is set to zero before mapping back to the observation space.

**$f_0$ protection:** The fundamental frequency ($f_0$) is known to contain information about the sex. We therefore apply an affine transformation to the $f_0$ to force a fixed target $f_0$ trajectory mean and standard deviation that we expect to be sex-neutral. They are computed from a training set where the means and standard deviations from $f_0$ trajectories are first averaged at the speaker level and are then averaged over all males and all females resulting in two means and two standard deviations $f_0$ (one for each sex). Then, the target mean $f_0$ is obtained by taking the average between the male and the female mean $f_0$, and the target standard deviation is obtained by taking the average between the male and the female standard deviation. This careful averaging is done in order to avoid bias due to an unbalanced number of utterances per speaker and speakers per sex.

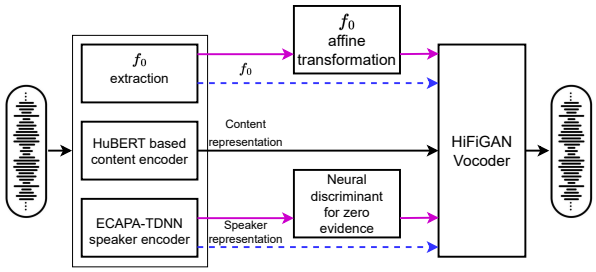



**Fig. 1**: Architecture of our system. Blue-dashed path is used during training and purple one during protection.

Figure 1 shows an outline of our system. The blue-dashed arrows show the training path of the HiFiGAN. The feature extractors are pretrained and fixed. Once the vocoder has been trained, the purple path is used. Both the $f_0$ and the speaker representation are transformed to reduce the sex-related information they contain. The content representation is assumed to not contain sex-related information. However, in real applications, the speaker might explicitly reveal their sex but we do not consider this scenario and instead focus on hiding the sex information in the acoustic features.

## 4. EXPERIMENTS

This section presents the sets used for training and testing the system, the baselines with which we compare it and experimental results.

### 4.1. Training and testing sets

**Vocoder's training:** LibriTTS train-clean-100 [18] was used to train the HiFiGAN as in [14]. The feature extraction modules were pretrained and fixed. ECAPA-TDNN [12] with 80-coefficient FBank features [14] was used for the 192-dimensional speaker representation extraction and was trained on VoxCeleb2 development set [19]. The 200-dimensional content representation was extracted by a HuBERT soft content encoder [11] fine-tuned from a pretrained HuBERT base model[2] on LibriTTS train-clean-100. Its training procedure is detailed in [14]. We used YAAPT [20] for the $f_0$ extraction, which does not require any training.

**Training of protection modules:** Once the analysis/synthesis system is trained, it can be used for privacy by manipulating the speaker representation and the $f_0$. In our experiments, the former was protected using the discriminant analysis for zero-evidence sex recognition presented in [4] and summarised in Section 3. It was trained on ECAPA-TDNN speaker embeddings [12] from LibriTTS train-other-500. The target $f_0$ mean and standard deviation were computed as described in 3 from LibriTTS train-other-500 also.

**Testing sets:** The VoicePrivacy challenge provides a complete evaluation protocol. For conciseness, we merged its libri_dev and libri_test sets to assess our system, resulting in 35 females with a total of 1185 utterances and 34 males with a total of 1136 utterances.

[2] https://github.com/pytorch/fairseq/tree/main/examples/hubert

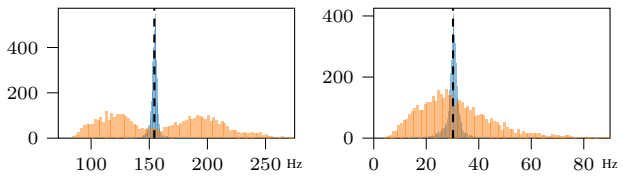



**Fig. 2**: Mean $f_0$ (left) and standard deviation $f_0$ (right) histograms for original speech (orange) and for generated protected speech (blue) with target $f_0$ mean and standard deviation (dashed lines).

### 4.2. Baselines

We compared the proposed approach with two baselines. The first, called *global*, is the same as our proposed approach but instead of using the neural-discriminant-analysis-based protection of the speaker representation, we simply fed the HiFiGAN with the same global averaged xvector for all utterances. The averaging was done in such a way as to avoid bias as it was done for the computation of the target $f_0$ trajectory moments. In this case, we expect that the sex of the original speaker will be hidden but that all the speaker information will be altered such that the resulting voices all look the same. The second baseline transforms only the $f_0$ using time domain pitch synchronous overlap add (TDPSOLA) [21] where, because we know that sex information is also contained in the spectral envelope, we expect that the sex of the speaker will not be satisfactorily hidden.

### 4.3. Results

We report the results for *original* speech, *synthesised*, i.e., fed into our system but without xvector and $f_0$ transformations, protected with the *proposed* approach, i.e., with xvector and $f_0$ transformations as presented in Section 3, together with the *global* approach and the *TDPSOLA* approach. To assess the protection, we report to which extent an automatic sex classifier is able to detect the sex of the speaker. We then present results for ASR and ASV as downstream tasks and voice similarity matrices. Finally, we present listening test results. First, Figure 2 shows histograms of generated $f_0$ trajectory mean and standard deviation when the proposed approach was used. We can see that the mean and standard deviation of the generated $f_0$ follow the target ones. Indeed, their histograms (in blue) are narrow around the target values shown by the dashed lines.

#### *4.3.1. Protection assessment: Automatic sex classification*

To objectively assess the protection performance of the systems, we report the results of automatic sex recognition. This section is concerned with automatic attacks only. An attacker may try to infer the sex of the speaker by listening manually to the data. This point will be discussed in Section 4.3.3. We propose two kinds of attack: one where the classifier is trained on original speech data and one on protected speech data. The former corresponds to an *ignorant* attacker and the latter is analogous to a *semi-informed* attacker [13]. The *ignorant* attacker does not have access to the protection system or may not be aware that the data has been protected. In this case, it uses a sex classifier trained on natural non-protected speech. The *semi-informed* one is the strongest attack we consider. In this case, the attacker has access to the protection system. He or she can apply it to data he or she will be using for training the automatic sex classifier. The resulting classifier therefore benefits from the sex-related information that could remain in the protected data. The classifier

**Table 1**: Sex classification results for protection assessment, and automatic speech recognition WER.

| | ignorant | | semi-informed | | ASR |
|---|---|---|---|---|---|
| system | EER [%] | $D_{\text{ECE}}$ [bit] | EER [%] | $D_{\text{ECE}}$ [bit] | WER [%] |
| original | 3.67 | 0.578 | | | 4.02 |
| synthesised | 4.32 | 0.542 | 4.01 | 0.593 | 4.79 |
| global | 24.95 | 0.198 | 20.60 | 0.233 | 4.92 |
| proposed | 28.99 | 0.128 | 24.13 | 0.200 | 4.81 |
| TDPSOLA | 6.30 | 0.504 | 4.36 | 0.542 | 4.43 |

we used in our experiments is based on fine-tuned HuBERT base features extraction (with frozen convolution), statistical pooling and multilayer perceptron. Table 1 reports the results in terms of two metrics: the equal error rate (EER) and the $D_{\text{ECE}}$ [8]. The latter is a positive measure of the expected amount of information disclosed to the attacker when observing the output of the classifier. For privacy, we want a low $D_{\text{ECE}}$. The first line shows the initial ability to distinguish the sex of the speakers. We can see from the second line that this is slightly altered when processed even without protection applied. The next two lines show how the classification performance drops when protection is applied. For the ignorant attack, we have a drop in $D_{\text{ECE}}$ of 78% for the *proposed* approach and 66% for *global*. The methods are also robust to the semi-informed attack with a drop in $D_{\text{ECE}}$ of 65% and 60%. The *TDPSOLA* baseline is not competitive, which is not a surprise because it alters only the $f_0$ while it is known that differences in vocal tract shape between males and females are significantly related to the spectral envelope. However, we do not have a clear explanation as to why the *proposed* approach protects better than *global* baseline does. This could be due to uncontrolled bias in the data but this requires further study.

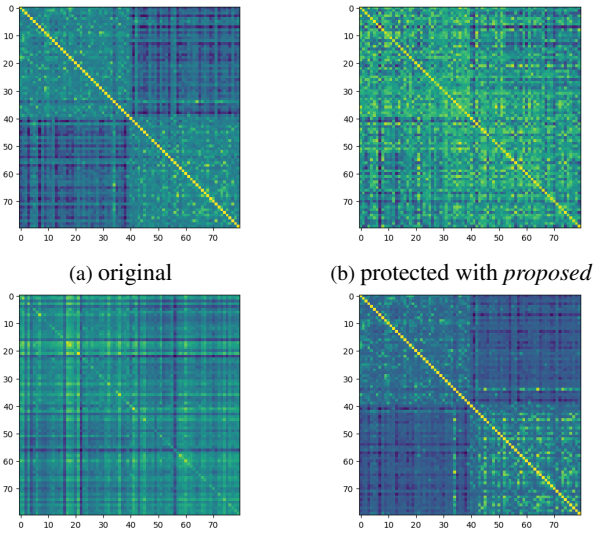
#### *4.3.2. Automatic speech recognition and speaker verification*

In this section, we want to ensure that automatic speech recognition and automatic speaker verification can still be performed as downstream tasks. We used the same ASR evaluation as in the VoicePrivacy challenge [2]. The word error rate (WER) is reported in Table 1. Processing the speech increases the WER slightly, but among the two systems that provide good protection, the proposed approach seems to alter the ASR performance less. At worst, 0.9% is added to the WER which is a relatively low price to pay for privacy.

We also report ASV results and voice similarities matrices to check if, after protection, ASV can still perform. We used the same ASV system used for evaluation in the VoicePrivacy challenge. It consists of a Kaldi TDNN speaker embedding extractor [16] with a PLDA backend. Both enrolment and test utterances were processed by the system. The EER and $C_{\text{llr}}^{\text{min}}$ [22] are reported in Table 2. We can see that processing the data without protection already slightly reduces the ASV performance. This suggests that the HiFiGAN vocoder results in a small distortion or domain shift. However, applying protection further reduces the ASV performance. For *global*, all the speaker variability in the xvector is annihilated by the global averaging, therefore increasing the confusion between voices. With the *proposed* approach, the xvector is disentangled in order to alter only the speaker's sex. Other speaker variabilities are preserved and, as expected, the protected voices remain consistent to some extent. Indeed, the proposed approach does far better than the global one although, compared with original data, significant ASV ability is lost with an increase in $C_{\text{llr}}^{\text{min}}$ from 0.278 to 0.445 and from 0.040 to 0.345 for female and male respectively. In addition to the domain shift induced by the HiFiGAN synthesis, this drop in performance could be

**Table 2**: Automatic speaker verification results. F and M refer respectively to in-between female and in-between male trials, while FM refers to cross-sex trials.

| system | EER [%] | | | $C_{\text{llr}}^{\text{min}}$ [bit] | | |
|---|---|---|---|---|---|---|
| | F | M | FM | F | M | FM |
| original | 8.15 | 1.13 | 5.77 | 0.278 | 0.040 | 0.204 |
| synthesised | 9.42 | 7.18 | 6.86 | 0.325 | 0.245 | 0.240 |
| global | 39.88 | 39.81 | 35.86 | 0.931 | 0.943 | 0.903 |
| proposed | 13.22 | 9.85 | 11.55 | 0.445 | 0.345 | 0.407 |
| TDPSOLA | 9.36 | 1.26 | 6.38 | 0.332 | 0.046 | 0.237 |



(a) original (b) protected with *proposed*

(c) protected with *global* (d) protected with *TDPSOLA*

**Fig. 3**: Voice log-similarity matrices [23].

explained by both the reduction in sex information as a component of the speaker variability that helps in distinguishing speakers, not only a male speaker from a female one but also between speakers with the same sex, and also by imperfect disentanglement of the sex component from other speaker-related information. In [23, 24], the authors proposed, in the context of the VoicePrivacy initiative, to visualise speaker voice similarity matrices to investigate the behavior of a protection system at both a speaker and global level. Here, our task is different, but we can still visualise voice similarity matrices to assess the consistency of the protected voices and to pay attention to any sex-related patterns that could appear in the matrices. We report four of these matrices in Figure 3. Speakers were grouped by sex such that squares appear in the matrix (a). Indeed, male speakers generally look more like other males than females and vice versa. When we have good sex protection in (b) and (c), we can see that these squares tend to disappear. The near disappearance of the diagonal in (c) confirms that *global* is not suitable enough to preserve other speaker variabilities compared with *proposed*.

### *4.3.3. Listening tests*

The results presented so far show the machine's perception. In this section, we discuss how the human ear perceives protected speech by reporting listening test results. 19 listeners, all native English speakers, were asked to assess the naturalness of speech on a discrete scale from 1 (unnatural) to 10 (natural) and whether the speech sounded like a male (1), a female (5), or neutral (3), also allowing for some nuance with scores of 2 and 4. *Neutral* refers here to the zero-evidence formulation of privacy where we want the data to provide no evidence about the speaker's sex such that the listener posterior belief remains equal to the prior one. Figure 4 shows the naturalness and sex perception scores. The speech processed by the analysis/synthesis even without protection does not sound as natural as the original speech. However, applying the protection does not further decrease the naturalness. As expected, *TDPSOLA* does not sufficiently change the perception of the sex. While *global* seems to change the perception of the speech from males, it does not have the expected behavior for females[3]. The *proposed* approach works for both male and female with a good average score close to 3 which tends to make attacks by listening inefficient. However, for better zero-evidence protection of each utterance, it would have been better to have narrower distributions around the neutral score.

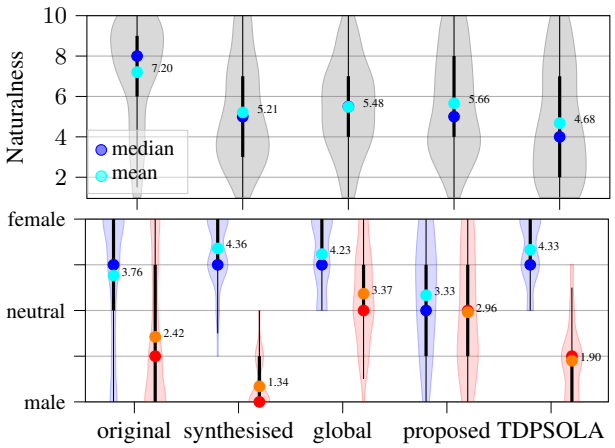



**Fig. 4**: Listening test results. Violin plots of perceived speech naturalness (top). Violin plots of perceived speaker's sex (bottom), blue for female and red for male; blue and red dots show medians, cyan and orange dots show means.

## 5. CONCLUSION

For privacy reasons, this paper proposed removing the sex of the speaker in speech using an analysis/synthesis-based voice conversion pipeline. An affine transformation is applied to the pitch, and the speaker representation is disentangled using a neural-discriminant analysis in order to conceal the speaker's sex-related information. The latter is consistent with the zero-evidence framework. The protection ability of the system was checked by means of an automatic sex classifier considering both an *ignorant* and a *semi-informed* attacker. Automatic speech recognition can still be applied on protected speech. Although the automatic speaker verification is deteriorated, the protected voices remain consistent to a certain extent. A listening test showed that the naturalness of the protected speech is satisfactory and the perception of the speaker's sex is altered, making attacks by listening more difficult.

In the future, we are interested in extending this work to other attributes like, for instance, accents. However, as accents involve more classes and are rarely labeled in large datasets, we expect that handling them will be even more challenging.

[3]Again, we do not have a clear explanation as to why *global* does not protect the data as well as *proposed*.

## 6. REFERENCES


[1] Andreas Nautsch, Abelino Jiménez, Amos Treiber, Jascha Kolberg, Catherine Jasserand, Els Kindt, Héctor Delgado, Massimiliano Todisco, Mohamed Amine Hmani, Aymen Mtibaa, Mohammed Ahmed Abdelraheem, Alberto Abad, Francisco Teixeira, Driss Matrouf, Marta Gomez-Barrero, Dijana Petrovska-Delacrétaz, Gérard Chollet, Nicholas Evans, Thomas Schneider, Jean-François Bonastre, Bhiksha Raj, Isabel Trancoso, and Christoph Busch, "Preserving privacy in speaker and speech characterisation," *Computer Speech & Language*, vol. 58, pp. 441 – 480, 2019.

[2] N. Tomashenko, Brij Mohan Lal Srivastava, Xin Wang, Emmanuel Vincent, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Jose Patino, Jean-François Bonastre, Paul-Gauthier Noé, and Massimiliano Todisco, "Introducing the VoicePrivacy Initiative," in *Proc. Interspeech*. ISCA, 2020, pp. 1693–1697.

[3] Ranya Aloufi, Hamed Haddadi, and David Boyle, "Privacy-preserving voice analysis via disentangled representations," in *Proc. SIGSAC Conference on Cloud Computing Security Workshop*. ACM, 2020, pp. 1–14.

[4] Paul-Gauthier Noé, Andreas Nautsch, Driss Matrouf, Pierre-Michel Bousquet, and Jean-François Bonastre, "A bridge between features and evidence for binary attribute-driven perfect privacy," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 3094–3098.

[5] Davit Rizhinashvili, Abdallah Hussein Sham, and Gholamreza Anbarjafari, "Gender neutralisation for unbiased speech synthesising," *Electronics*, vol. 11, no. 10, 2022.

[6] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, Doina Precup and Yee Whye Teh, Eds. 06–11 Aug 2017, vol. 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, PMLR.

[7] Dimitrios Stoidis and Andrea Cavallaro, "Generating gender-ambiguous voices for privacy-preserving speech recognition," in *Proc. Interspeech 2022*, 2022, pp. 4237–4241.

[8] Andreas Nautsch, Jose Patino, N. Tomashenko, Junichi Yamagishi, Paul-Gauthier Noé, Jean-François Bonastre, Massimiliano Todisco, and Nicholas Evans, "The Privacy ZEBRA: Zero Evidence Biometric Recognition Assessment," in *Proc. Interspeech*. ISCA, 2020, pp. 1698–1702.

[9] C. E. Shannon, "Communication theory of secrecy systems," *The Bell System Technical Journal*, vol. 28, no. 4, pp. 656–715, 1949.

[10] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 17022–17033, Curran Associates, Inc.

[11] Benjamin van Niekerk, Marc-André Carbonneau, Julian Zaïdi, Matthew Baas, Hugo Seuté, and Herman Kamper, "A comparison of discrete and soft speech units for improved voice conversion," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6562–6566.

[12] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *Proc. Interspeech 2020*, pp. 3830–3834, 2020.

[13] Brij Mohan Lal Srivastava, Nathalie Vauquier, Md Sahidullah, Aurélien Bellet, Marc Tommasi, and Emmanuel Vincent, "Evaluating voice conversion-based privacy protection against informed attackers," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

[14] Xiaoxiao Miao, Xin Wang, Erica Cooper, Junichi Yamagishi, and Natalia Tomashenko, "Language-Independent Speaker Anonymization Approach Using Self-Supervised Pre-Trained Models," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 279–286.

[15] Xin Wang and Junichi Yamagishi, "Neural Harmonic-plus-Noise Waveform Model with Trainable Maximum Voice Frequency for Text-to-Speech Synthesis," in *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)*, 2019, pp. 1–6.

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

[17] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio, "Density estimation using real NVP," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. 2017, OpenReview.net.

[18] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, "Libritts: A corpus derived from librispeech for text-to-speech," 2019.

[19] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*. ISCA, 2018, pp. 1086–1090.

[20] Kavita Kasi and Stephen A. Zahorian, "Yet another algorithm for pitch tracking," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, vol. 1, pp. I–361–I–364.

[21] Eric Moulines and Francis Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5, pp. 453–467, 1990, Neuropeech '89.

[22] Niko Brümmer and Johan du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2, pp. 230–275, 2006, Odyssey 2004: The speaker and Language Recognition Workshop.

[23] Paul-Gauthier Noé, Andreas Nautsch, Nicholas Evans, Jose Patino, Jean-François Bonastre, Natalia Tomashenko, and Driss Matrouf, "Towards a unified assessment framework of speech pseudonymisation," *Computer Speech & Language*, vol. 72, pp. 101299, 2022.

[24] Paul-Gauthier Noé, Jean-François Bonastre, Driss Matrouf, N. Tomashenko, Andreas Nautsch, and Nicholas Evans, "Speech Pseudonymisation Assessment Using Voice Similarity Matrices," in *Proc. Interspeech 2020*, 2020, pp. 1718–1722.