INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Voice Privacy Through x-vector and CycleGAN-based Anonymization

Gauri P. Prajapati, Dipesh K. Singh, Preet P. Amin, and Hemant A. Patil

Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT),
Gandhinagar, Gujarat, India.
{gauri_prajapati, dipesh_singh, preet_amin, hemant_patil}@daiict.ac.in

Abstract

With the rise in usage of voice assistants and spoken language interfaces, important concerns regarding voice data privacy have been prompted. In an attempt to reduce the threat of attacks on voice data, in this paper, we propose a speaker anonymization system based on CycleGAN. This method modifies the speaker's gender and accent information from the original speech signal. The proposed method gives a more natural-sounding anonymized voice in addition to a de-identified speaker. We have chosen baseline-1 of The Voice Privacy Challenge-2020 as our baseline system. Training of CycleGAN, ASR, and ASV experiments are performed on the subset of Librispeech corpus. In this paper, the double anonymization technique is also explored in which the CycleGAN-based anonymization technique is adopted on top of the baseline system. Experimental results show that combining the proposed method with the x-vector and neural source-filter (NSF) model-based method (baseline system) gives up to 5.61% relative improvement in EER of original-anonymized, enroll-trial pairs. However, it gives up to 19.30% relative improvement in EER for anonymized-anonymized enroll-trial pairs. We observed that along with the good speaker de-identification, the anonymized utterances have adequate speech intelligibility and naturalness.

Index Terms: Voice privacy, voice anonymization, CycleGAN.

1. Introduction

Speech is widely used as the powerful form of communication between humans and several automated systems. Due to the advancement and ease of voice biometrics and voice assistants, people use them for online banking, security, transcribing meetings, online purchases, etc. [1]. However, voice data contains sensitive and speaker's personal information, such as password, age, gender, health status, geographical background, etc., which may result in privacy risk for the speaker [2]. Hence, the General Data Protection Regularization (GDPR) highlights personal data protection, including speech data [3, 4]. Several approaches exist to protect the speaker-specific data in the speech signal: cryptographic methods, de-identification, pseudonymization, and anonymization of speech. The system that incorporates this method is called a voice privacy system [5].

Speaker anonymization is a process of hiding or modifying a speaker's identity such that the resultant speech sounds as if it was uttered by a different speaker (i.e., *pseudo speaker*) without affecting the linguistic content. The unpublished voice of user is given to the system and the resulting speech signal is called *trial utterance* of the pseudo speaker. Considerable methods for anonymization have been proposed in [6, 7, 8, 9, 10, 11, 12, 13] and much more. Some of these studies include methods based on voice transformation (VT) [14], VoiceMask, vocal tract length normalization (VTLN)-based voice conversion (VC), and noise (i.e., pink noise) addition. In [6], authors have proposed a de-identification approach - voice mask via two frequency warping functions. In addition, it is observed that the age and gender modification of the speaker can also be altered to anonymize the speech signal. To that effect, authors in [15] have changed fundamental frequency (i.e., F_0), the first four formant frequencies (i.e., $F_1 - F_4$), and corresponding $-3dB$ bandwidths (i.e., $B_1 - B_4$) which carry speaker-specific information, in particular, the higher formants.

Fang et. al. [9] proposed to modify speaker embedding (x-vector [16]) after separating it from the linguistic content (i.e., F_0). The modified x-vector is then used with the original linguistic characteristics to generate the anonymized voice with the help of neural source-filter (NSF) model. This approach is proposed as the primary baseline of the Voice Privacy Challenge organized during INTERSPEECH 2020 [17]. Recently, a study represented by Mawalim et al. [18], this x-vector and NSF-based approach is further improved using singular value modification of x-vector. These both approaches need an x-vector pool to have the anonymized x-vector.

Generative Adversarial Networks (GANs) are mainly proposed to estimate probability density function of underlying data [19, 20, 21]. It has many variants proving their effectiveness in several major fields. One such network is CycleGAN, popular in voice conversion [22]. The capability of CycleGAN to translate features from one-domain to another makes it a suitable candidate for the anonymization approach in our proposed study. Hence, we have used CycleGAN to adapt features of male and female speakers. We have also explored the effect of combining proposed approach to the baseline system, i.e., *double anonymization*.

2. CycleGAN for Voice Privacy

GANs are an impressive field due to their ability to generate realistic outputs for various generative problem domains, such as image-to-image translation, generating realistic photos, and VC, etc. There are many variants of GANs, namely, conditional GANS (cGAN), stackGAN, CycleGAN, Wasserstein GANs (WGAN), etc. The CycleGAN is a well known architecture, which has concept of cycle-consistency in the GAN architecture [23]. In this study, we have translated features between two domains, male and female using CycleGAN, which is discussed in this Section.

2.1. CycleGAN architecture

We have used CycleGAN-VC2 architecture similar to described in [22], with some minor improvements. For discriminator network, instead of using FullGAN which fails to capture high frequency structures, a more efficient PatchGAN with three down-

http://dx.doi.org/10.21437/Interspeech.2021-1573

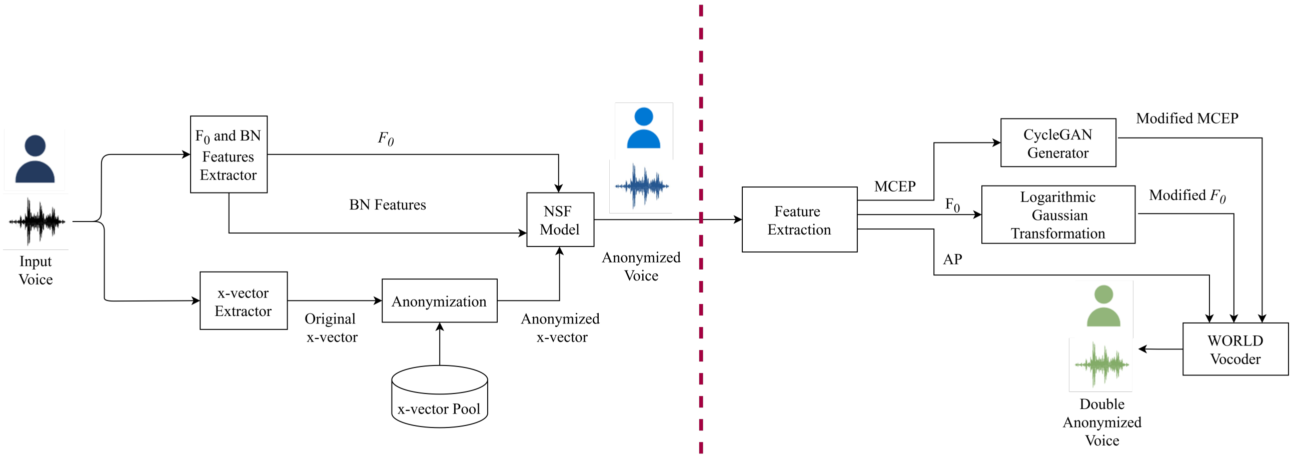

Figure 1: *Functional block diagram of proposed double anonymization approach.*

sampling layers is used [24]. 2-1-2D generator described in [22] is used with some modification to reconstruct features as close to real as possible. For the two downsampling layers, strided 2D convolution layers with instance normalization and gated linear units (GLU) as activation function were used [25]. To perform feature transformation six residual layers of 1D CNN with instance normalization and GLU were used. To make network learn its own upsampling, instead of using interpolation with pixel-shuffle, strided transpose-convolution layers are used as two upsampling layers. To adjust feature dimensions during reshaping, done before and after residual layers, 1×1 1-D convolution is applied. For CycleGAN-based voice transformation, instead of separating dataset speaker-wise, we have formed two classes based on the gender of speakers, namely, female class and male class. Through CycleGAN training, we attempt to transform the male speech into female speech and vice-versa.

Let $M \subset \mathbb{R}^{D\times T}$ be feature space representing male speaker utterances, and $F \subset \mathbb{R}^{D\times T}$ be representing feature space of female utterances. D is the dimension of feature vector being used, while T corresponds to the number of speech frames. Our goal is to convert $\mathbf{m} \in M$ to $\mathbf{f} \in F$, and vice-versa. For which mapping functions $G_{M\rightarrow F} : M \rightarrow F$, and $G_{F\rightarrow M} : F \rightarrow M$ are learned with the use of CycleGAN network. To improve the performance of these mapping functions, they are trained in adversary with discriminating functions $D_M : \mathbb{R}^{D\times T} \rightarrow [0,1]$, and $D_F : \mathbb{R}^{D\times T} \rightarrow [0,1]$.

CycleGAN is trained with four different loss functions to make these mappings more accurate [22].

Adversarial loss: To perform the task of voice transformation, and learn the mapping functions, generators are trained in adversary with discriminators, whose job is to discriminate between real features and transformed features. For a pair of generator and discriminator, $G_{M\rightarrow F}$ and D_F, the goal of generator is to make $G_{M\rightarrow F}(m) \sim F$, while the goal of D_B is to discriminate between them. Adversarial loss for the given generator-discriminator-pair is defined as:

$$\mathcal{L}_{adv}(G_{M\rightarrow F}, D_F) = E_{f\sim P_F(f)}[logD_F(f)] \quad + \\ E_{m\sim P_M(m)}[log(1 - D_F(G_{M\rightarrow F}(m)))]. \tag{1}$$

Cycle-consistency loss: Since the task of generators is to convert feature from one class to another, we should be able to reconstruct back original feature by using another generator. In the other words, two mapping functions $G_{M\rightarrow F}$, and $G_{F\rightarrow M}$ should be inverse of each other. This ensures that two mappings are one-to-one mappings (and also onto), as well as $G_{F\rightarrow M}(G_{M\rightarrow F}(m)) \approx m$.

$$\mathcal{L}_{cyc}(G_{M\rightarrow F}, G_{F\rightarrow M}) = \\ E_{m\sim P_M(m)}[\|G_{F\rightarrow M}(G_{M\rightarrow F}(m)) - m\|_1] + \\ E_{f\sim P_F(f)}[\|G_{M\rightarrow F}(G_{F\rightarrow M}(f)) - f\|_1], \tag{2}$$

where $\|.\|_1$ represents L_1-norm, and $E[.]$ represents expectation operator.

Identity-mapping loss: Identity-mapping loss ensures that a feature belonging to target feature space in mapping function is mapped to itself, i.e., $G_M M\rightarrow F(f) = f$. Other than that it is also used to preserve the intelligibility of the utterance, since adversarial loss and cycle-consistency loss are used to perform speaker identity transformation between class of speakers, and they alone cannot ensure that linguistic content does not become corrupted.

$$\mathcal{L}_{id}(G_{M\rightarrow F}, G_{F\rightarrow M}) = E_{m\sim P_M(m)}[\|G_{F\rightarrow M}(m) - m\|_1] + \\ E_{f\sim P_F(f)}[\|G_{M\rightarrow F}(f) - f\|_1]. \tag{3}$$

Two-step adversarial loss: To further stabilize the training of CycleGAN and to negate the effect of smoothing caused by L_1-norm from cycle-consistency loss, we also apply a second adversarial loss for reconstructed features.

$$\mathcal{L}_{2-step}(G_{M\rightarrow F}, G_{F\rightarrow M}, D_F) = \\ E_{f\sim P_F(f)}[logD_F(f)] \quad + \\ E_{m\sim P_M(m)}[log(1 - D_F(G_{M\rightarrow F}(G_{F\rightarrow M}(f))))]. \tag{4}$$

All the four functions are used to optimize the generators, while discriminators are only optimized using adversarial loss. To calculate the total objective of the generator, all the loss functions added with their associated weights. These weights are taken as hyperparameters λ_{adv}, λ_{cyc}, and λ_{id}.

$$\begin{aligned}\mathcal{L}_{gen} = {} & \lambda_{adv}\mathcal{L}_{adv}(G_{F\to M}, D_M) \quad + \\ & \lambda_{adv}\mathcal{L}_{adv}(G_{M\to F}, D_F) \quad + \\ & \lambda_{adv}\mathcal{L}_{2-step}(G_{F\to M}, G_{M\to F}, D_M) \quad + \\ & \lambda_{adv}\mathcal{L}_{2-step}(G_{M\to F}, G_{F\to M}, D_F) \quad + \\ & \lambda_{cyc}\mathcal{L}_{cyc}(G_{F\to M}, G_{M\to F}) \quad + \\ & \lambda_{id}\mathcal{L}_{id}(G_{F\to M}, G_{M\to F}). \end{aligned} \tag{5}$$

2.2. Double Anonymization

We have also explored the double anonymization approach for anonymization in which CycleGAN-based technique is applied on top of the baseline system. The overall double anonymization system block diagram is shown in Fig. 1, where the left side of the dashed line is the baseline system, whereas the right side part is the proposed CycleGAN-based anonymization system. The baseline system is explained in next Section. For CycleGAN-based anonymization, Mel Cepstral Coefficients (MCEP), fundamental frequency (F_0) contour, and aperiodicity (AP) were extracted. Then only the MCEP and F_0 is modified keeping AP the same as that of original. These modified MCEP, F_0, and original AP is used to have the anonymized utterances from the WORLD vocoder.

3. Experimental Setup

3.1. Dataset Used

Three subsets of Librispeech corpus [26] are used to train, develop, and evaluate the proposed voice privacy system. The Libri-train-clean-360 dataset is used for the training of the ASV and ASR models. The remaining two subsets, Libri-dev-clean and Libri-test-clean are used as development and evaluation datasets, respectively. These both the datasets are subdivided into trials and enrolls for the ASV performance assessment. More details related to datasets are given in [17].

3.2. Baseline System

We have used the anonymization method proposed in [9] as our baseline system. Three features, namely, fundamental frequency (F_0), bottleneck (BN) features, and speaker-specific x-vector are extracted from the original input speech signal. For anonymization, x-vectors are modified, to have a different x-vector (i.e., pseudo x-vector) corresponding to a pseudo speaker. For getting the pseudo x-vector that is farthest from N (i.e., 200), x-vectors (taken from an external pool) are averaged. This pseudo x-vector and the original F_0 and BN features are used to get the anonymized utterance using the NSF model. This algorithm is shown in the left part of the dashed line in the functional block diagram represented in Fig. 1. This approach is chosen as the baseline system since our proposed system is also a deep learning-based privacy preservation method.

3.3. CycleGAN

3.3.1. Input Features

To achieve the task of Voice Privacy through VC, 36-dimensional (0^{th} + 35) MCEP, and F_0 were used. MCEP, F_0 contour, and AP were extracted using WORLD vocoder [27]. For effective voice conversion, MCEPs are transformed using trained CycleGAN network, while F_0 contour is transformed using logarithmic Gaussian normalized transformation [28]. AP of the utterance is left untouched.

3.3.2. CycleGAN training

Normalized MCEPs with zero mean and unit variance were given as an input to train the CycleGAN network. $7,000$ utterances from each male and female class were taken to train our network. The network was trained with randomly selected 256 frames, which added more randomness on top of randomly selected utterances. To converge our GAN network with proper stability, and to help generator in learning, Least Squares GAN (LSGAN) were used [29]. For generators initial learning rate was taken as 4×10^{-4} and for discriminators, it was taken as 2×10^{-4}. Network was trained for a total of 5×10^4 iterations with a batch size of 8. A multi-step learning rate exponential decay function was implemented which decreased the learning rate by a factor of $\gamma = 2$ at iterations 4×10^{-4} and 2×10^{-4}. To optimize the network model weights, Adam optimizer was used with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. Weights of the loss functions, λ_{adv}, λ_{id}, and λ_{cyc} were initialized as 1, 5, and 10 respectively, of which λ_{id} was reduced to 0 once 2×10^4 had passed.

3.4. Objective evaluation

The voice privacy system needs to be evaluated for speaker verification and speech intelligibility. An ASV system verifies the speaker identity modification by producing a high equal error rate (EER) (i.e., less accuracy of the ASV system for a particular target speaker). ASV system performance is evaluated for the three cases: (i) original-original (O-O) utterances, (ii) original-anonymized (O-A) utterances, and (iii) anonymized-anonymized (A-A) utterances. For speech intelligibility, an ASR system is used which checks the loss that occurred in the information provided by the anonymized speech signal in terms of word error rate (WER). Hence, lesser WER reflects lesser linguistic information loss during anonymization. We have further used the recently introduced similarity measure for pseudonymization, namely, de-identification (DeID) measure [30]. It resembles the uncertainty for linkability between given utterance and the speaker identity.

4. Experimental Results

This Section presents results on the CycleGAN-based anonymization method. The performance is measured with the help of ASV and ASR systems. Table 1 and Table 2 shows the experimental results for CycleGAN and double anonymization approach.

4.1. Results of ASV and ASR Systems :

As explained earlier in sub-Section 3.4, the ASV evaluation is carried out for three combinations of enroll and trial pairs. The O-A case gives the ASV score stating the effectiveness of the anonymization method, whereas the A-A case reflects the capability of a method to anonymize a speaker in a different way for every utterance. We can observe from Table 1 that the CycleGAN-based approach lacks performance in terms of EER, however, the corresponding ASR results reflects that with approximately 40% of EER, it still gave 4.71% WER which is nearer to WER of the original speech. These results show that the CycleGAN-based approach gives better speaker de-identification with lesser damaged linguistic content via less WER.

From the above discussion, we can say that the baseline system has the capability to do de-identification, however, it lacks performance w.r.t. speech intelligibility and naturalness. Whereas the CycleGAN-based approach gives comparatively good EER, with effectively protecting speech intelligibility. Hence, we propose to use the CycleGAN network on top of the baseline system in order to get both better speaker de-identification and intelligibility. It was observed that using the double anonymization approach gives approximately 50% EER

Table 1: *ASV results of CycleGAN-based and double anonymization methods for development and test data (O-original, A-anonymized speech; Gen denotes gender: f-female, m-male)*

#	Dataset	Enroll	Trial	Gen	EER (%)		
					Baseline	**CycleGAN**	**Double Anon.** (relative % improvement)
1		O	O	f	8.66	8.66	8.66 (0)
2	libri_dev	O	A	f	50.14	43.32	**53.12 (+5.94)**
3		A	A	f	36.79	26.14	**38.78 (+5.41)**
4		O	O	m	1.24	1.24	1.24 (0)
5	libri_dev	O	A	m	57.76	33.23	52.80 (-8.58)
6		A	A	m	34.16	24.19	**38.98 (+14.11)**
7		O	O	f	7.66	7.66	7.66 (0)
8	libri_test	O	A	f	47.26	37.77	**49.45 (+4.63)**
9		A	A	f	32.12	23.18	**38.32 (+19.30)**
10		O	O	m	1.11	1.11	1.11 (0)
11	libri_test	O	A	m	52.12	32.07	47.88 (-8.13)
12		A	A	m	36.75	24.01	**40.31 (+9.68)**

Table 2: *ASR results of Baseline-1, CycleGAN, and double anonymization for development and test data (O-original, A-anonymized speech)*

Dataset	Data	WER (%)		
		Baseline	**CycleGAN**	**Double Anon.**
libri_dev	O	3.82	3.82	3.82
	A	6.39	**4.71**	8.62
libri_test	O	4.15	4.15	4.15
	A	6.73	**4.75**	9.11

of O-A case and with approximately 9% WER for anonymized voices. We can also observe that the EER for the A-A case is approximately 39%, which shows that the anonymization approach gives better privacy because it gives different pseudo speakers for a single original speaker when we change the utterance.

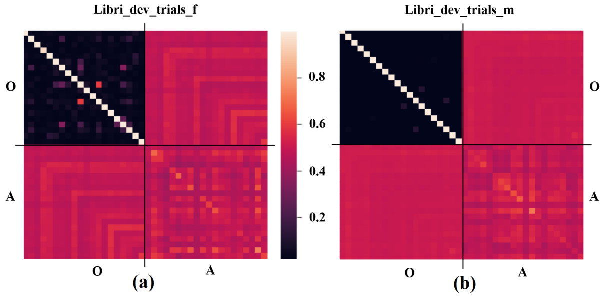

Figure 2: *Voice similarity matrices for the proposed method on (a) libri-dev-trials-f, and (b) libri-dev-trials-m datasets.*

4.2. Analysis of Voice Similarity:

In addition to the ASV and ASR performance, we have also analyzed the recently proposed similarity metric, DeID which is derived from similarity matrices, as shown in Fig. 2 and Fig. 3. In Fig. 2(a), upper-left quadrant (M_{OO}) reflects the voice similarity within the original speech signals [30] for development set. And the same is for anonymized speech signals in the bottom-right quadrant (M_{AA}) as shown in Fig. 2(b). The upper-right (M_{OA}) and bottom-left (M_{AO}) quadrants show the similarity between original and anonymized speech signals. We can observe that there is less similarity for the A-A case (i.e., less dominant diagonal in (M_{AA})) with diagonal values near to 0.5. However, the O-A case has almost no similarity (i.e., no dominant diagonal is observed), which reflects that the proposed double anonymization method is giving good speaker de-identification. In particular, the DeID was found to be 99.54% for libri-dev-trials-f, and 99.89% for libri-dev-trials-m set. Similar observations were also found in the test set as shown in Fig. 3.

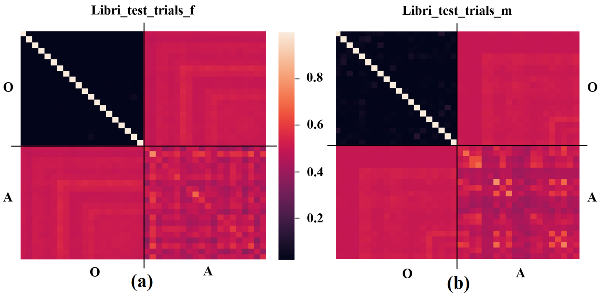

Figure 3: *Voice similarity matrices for the proposed method on (a) libri-test-trials-f, and (b) libri-test-trials-m datasets.*

5. Summary and Conclusions

This study presents a CycleGAN-based anonymization approach for voice privacy system. We have used CycleGAN framework for converting gender-specific information in the utterance (i.e., female class to male class), instead of one-to-one mapping. We have used MCEP, F_0, and AP features for this purpose. From the results, it can be observed that this architecture gives good anonymization with very less degradation speech quality. The WER obtained for the CycleGAN-based approach is found to be near to the WER of original speech signal. Hence, we have also explored the effect of double anonymization approach, where this framework is applied on top of the baseline system approach. We observed that this double anonymization approach gives better speaker anonymization (i.e., more EER), however, the speech quality was degraded from that of the original speech signal. However, the WER from double anonymization approach is still less than the baseline system. Hence, we can say that the proposed anonymization method can be used for preserving speaker-specific information. This system can be improved by focusing on the speech intelligibility improvement techniques. In addition, other GAN architectures also can be explored for anonymization.

6. References

[1] M. B. Hoy, “Alexa, Siri, Cortana, and more: An introduction to voice assistants,” *Medical Reference Services Quarterly*, vol. 37, no. 1, pp. 81–88, 2018.

[2] A. Nautsch, A. Jiménez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa *et al.*, “Preserving privacy in speaker and speech characterization,” *Computer Speech & Language,*, vol. Special Issue, 58, pp. 441–480, 2019.

[3] A. Nautsch, C. Jasserand, E. Kindt, M. Todisco, I. Trancoso, and N. Evans, “The GDPR and speech data: Reflections of legal and technology communities, first steps towards a common understanding,” *arXiv preprint arXiv:1907.03458*, 2019, {Last Accessed: 15-03-2021}.

[4] G. D. P. Regulation, “Regulation eu 2016/679 of the European parliament and of the council of 27 April 2016,” *Official Journal of the European Union. Available at: http://ec. europa. eu/justice/data-protection/reform/files/regulation_oj_en. pdf*, 2016, {Last Accessed: 15-03-2021}.

[5] B. M. L. Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, “Evaluating voice conversion-based privacy protection against informed attackers,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Virtual Barcelona, May 2020, pp. 2802–2806.

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

[7] M. Pobar and I. Ipšić, “Online speaker de-identification using voice transformation,” in *2014* 37^{th} *International convention on information and communication technology, electronics and microelectronics (mipro)*, Opatija, Croatia, May 2014, pp. 1264–1267.

[8] C. Magarinos, P. Lopez-Otero, L. Docio-Fernandez, E. Rodriguez-Banga, D. Erro, and C. Garcia-Mateo, “Reversible speaker de-identification using pre-trained transformation functions,” *Computer Speech & Language*, vol. 46, pp. 36–52, 2017.

[9] 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,” *arXiv preprint arXiv:1905.13561*, 2019, {Last Accessed: 15-03-2021}.

[10] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, “Speaker anonymization using the McAdams coefficient,” *arXiv preprint arXiv:2011.01130*, 2020, {Last Accessed: 15-03-2021}.

[11] P. Gupta, G. P. Prajapati, S. Singh, M. R. Kamble, and . Hemant A. Patil, “Design of voice privacy system using linear prediction,” in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Auckland, New Zealand, 2020, pp. 543–549.

[12] Q. Jin, A. R. Toth, T. Schultz, and A. W. Black, “Speaker de-identification via voice transformation,” in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) Workshop*, Merano, Italy, 13 - 17 December, 2009, pp. 529–533.

[13] S. Ahmed, A. R. Chowdhury, K. Fawaz, and P. Ramanathan, “Preech: A system for privacy-preserving speech transcription,” in 29^{th} *{USENIX} Security Symposium*, 12-14 August 2020, pp. 2703–2720.

[14] Y. Stylianou, “Voice transformation: A survey,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 19-24 April 2009, pp. 3585–3588.

[15] J. Přibil, A. Přibilová, and J. Matoušek, “Evaluation of speaker de-identification based on voice gender and age conversion,” *Journal of Electrical Engineering*, vol. 69, no. 2, pp. 138–147, 2018.

[16] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust DNN embeddings for speaker recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Calgary, Canada: IEEE, 15-20April, 2018, pp. 5329–5333.

[17] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. L. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, “Introducing the Voice Privacy initiative,” in *INTERSPEECH*, Shanghai, China, 24-28 October, 2020.

[18] C. O. Mawalim, K. Galajit, J. Karnjana, and M. Unoki, “X-vector singular value modification and statistical-based decomposition with ensemble regression modeling for speaker anonymization system,” *INTERSPEECH*, pp. 1703–1707, 24-28 October, 2020.

[19] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Proceedings of the* 27^{th} *International Conference on Neural Information Processing Systems (NIPS)*, vol. 2, Montreal, Quebec, Canada, 8-13 December 2014, pp. 2672–2680.

[20] Y. Saito, S. Takamichi, and H. Saruwatari, “Statistical parametric speech synthesis incorporating generative adversarial networks,” *arXiv preprint arXiv:1709.08041*, 2017, {Last Accessed: 15-03-2021}.

[21] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, “StarGAN-VC2: Rethinking conditional methods for stargan-based voice conversion,” *arXiv preprint arXiv:1907.12279*, 2019, {Last Accessed: 15-03-2021}.

[22] T. Kaneko, H. Kameoka, K. Tanaka, and Hojo, “Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion,” in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019.* IEEE, 2019, pp. 6820–6824.

[23] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 22-29 October 2017, pp. 2223–2232.

[24] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, 21-26 July, 2017, pp. 1125–1134.

[25] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language modeling with gated convolutional networks,” in *International Conference on Machine Learning.* Sydney, Australia: PMLR, 6-11 August, 2017, pp. 933–941.

[26] 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)*, Queensland, Australia, 19-24 April, 2015, pp. 5206–5210.

[27] M. Morise, F. Yokomori, and K. Ozawa, “World: A vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[28] K. Liu, J. Zhang, and Y. Yan, “High quality voice conversion through phoneme-based linear mapping functions with straight for Mandarin,” in *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, vol. 4,. Haikou, China: IEEE, 24-27 August 2007, pp. 410–414.

[29] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, “Least squares generative adversarial networks,” in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 22-29 October, 2017, pp. 2794–2802.

[30] P.-G. Noé, J.-F. Bonastre, D. Matrouf, N. Tomashenko, A. Nautsch, and N. Evans, “Speech pseudonymization assessment using voice similarity matrices,” *arXiv preprint arXiv:2008.13144*, 2020, {Last Accessed: 15-03-2021}.