X-Vector Singular Value Modification and Statistical-Based Decomposition with Ensemble Regression Modeling for Speaker Anonymization System

Candy Olivia Mawalim[1], *Kasorn Galajit*[1,2], *Jessada Karnjana*[2], *Masashi Unoki*[1]

[1] Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa 923–1292 Japan
[2] NECTEC, National Science and Technology Development Agency, Pathum Thani, Thailand

candyolivia@jaist.ac.jp, kasorn@jaist.ac.jp, unoki@jaist.ac.jp,
jessada.karnjana@nectec.or.th

Abstract

Anonymizing speaker individuality is crucial for ensuring voice privacy protection. In this paper, we propose a speaker individuality anonymization system that uses singular value modification and statistical-based decomposition on an x-vector with ensemble regression modeling. An anonymization system requires speaker-to-speaker correspondence (each speaker corresponds to a pseudo-speaker), which may be possible by modifying significant x-vector elements. The significant elements were determined by singular value decomposition and variant analysis. Subsequently, the anonymization process was performed by an ensemble regression model trained using x-vector pools with clustering-based pseudo-targets. The results demonstrated that our proposed anonymization system effectively improves objective verifiability, especially in anonymized trials and anonymized enrollments setting, by preserving similar intelligibility scores with the baseline system introduced in the VoicePrivacy 2020 Challenge.

Index Terms: speaker anonymization, x-vector, singular value modification, statistical-based decomposition, ensemble regression modeling

1. Introduction

As speech is generally preferred over text communication, voice-input features have become widely implemented in recent technology. However, voice recordings can contain personal, sensitive information, which may lead to security and privacy risks when exposed [1]. Such risks are due to advancements in speech synthesis and conversion technology that have enabled increasingly accurate voice cloning even with limited speech samples [2, 3]. Consequently, there have been growing efforts to preserve voice security and privacy, one of the main proposed approaches being speaker anonymization.

Speaker anonymization or de-identification is a method for suppressing or concealing speaker identity in their speech data [4]. According to the VoicePrivacy 2020 Challenge [5], the following four requirements are important for a speaker anonymization system: (i) the speaker identity must be hidden, (ii) the output speech should be natural and intelligible, (iii) the language information should be preserved, and (iv) a speaker-to-speaker correspondence must be followed.

Several methods have been proposed for anonymization systems [1, 4, 6, 7, 8]. Previously, an anonymization system was developed by suppressing speaker identity using a voice transformation system [6, 7]. For instance, a diphone-based syntactic source speech (kaldiphone) is transformed to fit a set of speakers to attack the speaker identification system. It was suggested that this voice transformation could fluster the Gaussian mixture model (GMM) based speaker identification system [6]. Subsequently, a voice transformation method to de-identify speech using GMM mapping and harmonic-stochastic models was proposed [8]. De-identification of online speakers was feasible with this method. Next, a technique for concealing speaker identity through voice transformation was developed using the natural speech of a target person instead of a synthetic voice [9]. Another approach was implemented using cepstral frequency warping plus amplitude scaling to transform speech and hide the identity [10].

Fang et al. [4] proposed a method based on a neuro source-filter (NSF) model to separate the speaker identity and the linguistic content from the input speech before resynthesizing the speech data with modification of speaker identity information (x-vector). This method is referred to as the first baseline system in the VoicePrivacy 2020 Challenge [5]. The x-vector was chosen since it could effectively encode speaker identity as a feature in speaker verification system [11]. In the first baseline system, the original x-vector was replaced with the mean x-vector from the farthest x-vector group in the anonymization x-vector pool. On the other hand, our proposed method offers two different approaches for anonymizing speaker identity information (x-vector): (1) modifying its singular value, and (2) decomposition based on the x-vectors' statistical properties and transforming it with ensemble regression models. We predicted that by modifying the significant elements of x-vectors, the speaker-to-speaker correspondence requirement of anonymization system could be satisfied. Furthermore, we investigated the performance of the synthesis system to improve the quality of anonymized speech.

The rest of this paper is organized as follows. Section 2 describes the proposed anonymization system in detail. Section 3 presents the experimental setup and results of the proposed method. Finally, Section 4 presents the conclusion and future work.

2. Proposed Model

Figure 1 shows our proposed model for a speaker anonymization system. The analysis and synthesis framework from input speech to anonymized system using an x-vector and a neural-source filter (NSF) model were based on the first baseline system in Voice Privacy 2020 Challenge [4, 5, 12]. Four pre-trained models were employed in the baseline system, including an ASR acoustic model [4, 13] for extracting linguistic-related features or bottleneck (BN) features, an x-vector extractor [11] trained by VoxCeleb datasets [14, 15], a speech synthesis acoustic model [4], and an NSF [16] for generating a speech signal with F0, Mel-filterbank, and an anonymized x-vector as input. We modified the baseline model by replacing the F0 extractor with the one provided by another speech analysis toolkit. The experiment by Morise et al. demonstrated that WORLD

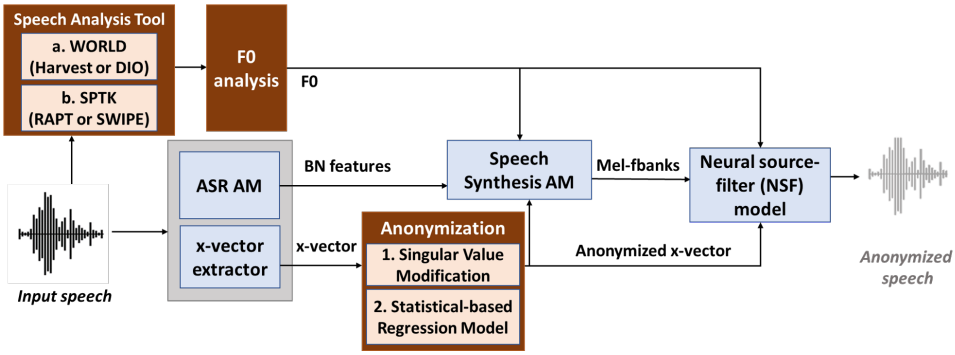

Figure 1: *Schematic diagram of proposed speaker anonymization system*

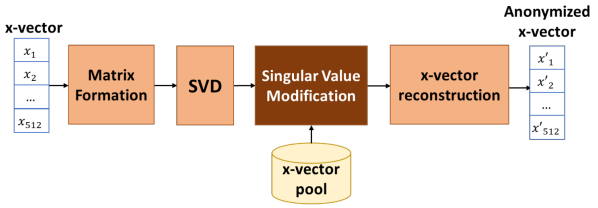

Figure 2: *Schematic diagram of x-vector modification by singular-value decomposition*

provides computationally intensive and robust F0 estimation [17]. Furthermore, SPTK gives the relatively best precision with other F0 extractors, including Yaapt (Kaldi) [18]. Therefore, we investigate the F0 extractors from WORLD [17] and SPTK [19] in this study.

We propose two approaches for anonymizing the x-vector: (1) modifying the singular value of the input x-vector, and (2) decomposing the input x-vector based on its statistical properties and transforming it with regression models.

2.1. X-vector Anonymization using Singular Value Modification

The first approach is based on the concept of matrix factorization using singular value decomposition (SVD), which has a variety of applications such as recommender systems and data reduction [20]. SVD provides the constituent elements of the matrix; the modification of the x-vector matrix singular values from a speaker is expected to provide the anonymized x-vector with similar constituent elements that represent intra-speaker information. Figure 2 shows our proposed x-vector anonymization process. Each step is explained in detail below.

X-vector Pool Construction. First, we constructed the x-vector pool to obtain the input x-vector and pseudo target x-vector. The pseudo target x-vector was determined from the least similar centroid using a clustering method.

Matrix Formation. An x-vector matrix ($\mathbf{X}$) was constructed using the x-vectors of all available utterances of a speaker. The output is the x-vector matrix for the pseudo target x-vectors with dimension $M \times N$, where M is the number of utterances and N is the dimension of x-vector (512).

Singular Value Decomposition (SVD) and Modification. The pseudo target x-vector matrix obtained from the previous step was decomposed into two singular matrices and a diagonal singular values matrix. The decomposition is expressed as:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}, \tag{1}$$

where $\mathbf{U}$ and $\mathbf{V}$ are the orthonormal eigenvectors of $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ and of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, respectively, and $\mathbf{\Sigma}$ consists of the square roots of the eigenvalues of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$.

In our approach, we interpreted $\mathbf{U}$ as the utterance-to-concept similarity matrix, and $\mathbf{V}$ as the x-vector-to-concept similarity matrix. $\mathbf{\Sigma}$ represents the strength of each concept involved. By reducing the dimension of $\mathbf{\Sigma}$, we expect to obtain more general constituent elements of the x-vector. Thus, x-vector anonymization is conducted by controlling $\mathbf{\Sigma}$ with a threshold parameter (singular value threshold). Figure 3 shows the anonymization of an x-vector singular value.

X-vector Reconstruction. Lastly, the anonymized x-vector of a speaker's utterance was obtained from the anonymized x-vector matrix reconstructed from $\mathbf{U}$, $\mathbf{V}$, and the modified $\mathbf{\Sigma}$.

2.2. Statistical-Based Regression Modelling

Figure 4 shows the second approach of our anonymization system, which comprises the following four steps:

X-vector Variant-based Decomposition. First, the variant of intraspeaker x-vectors in x-vector pool 1 was analyzed to observe the distribution of the x-vector of a speaker in different utterances. The standard deviation of the intraspeaker x-vectors were calculated with a given threshold to decompose the x-vector into two parts, i.e., high-variant x-vector (y_i) and low-variant x-vector (z_i). This decomposition is based on our hypothesis that the low-variant x-vector is a stable part of the x-vector that contains the uniqueness of the speaker identity; therefore, it is an important cue for one-to-one mapping from the original to anonymized speech.

Anonymization Pool Construction. After the x-vector was decomposed into high and low-variant parts, we built clustering models to create pseudo-target x-vectors. The clustering model was trained using x-vector pool 2. The clustering model produced several centroids which are assigned as the candidates of the pseudo-target x-vectors. The pseudo-target x-vector was determined by the centroid least similar to the pseudo-input x-vector. The pseudo-target x-vectors were fit into a regression model in two consecutive processes, and then pairs of pseudo input-target x-vectors were fit into the regression model. In other words, we defined the x-vectors pairs as our anonymization pool.

Ensemble Learning for Regression Modeling. Two ensemble regression models were constructed by fitting the anonymization pool x-vectors. A non-linear regression model was trained for the high-variant x-vector, and a linear regression model was trained for the low-variant x-vector. We predicted that by transforming the low-variant x-vector linearly, the uniqueness of each speaker's x-vector could be preserved. In other words, we fit a linear function ($z_i' = Az_i + B$) for transforming the original low-variant x-vector (z) to the anonymized

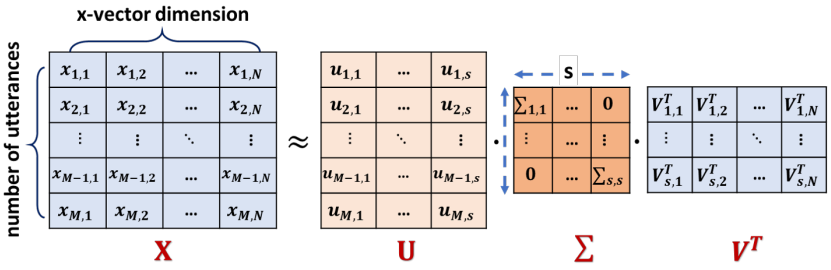

Figure 3: *Modification of x-vector singular values*

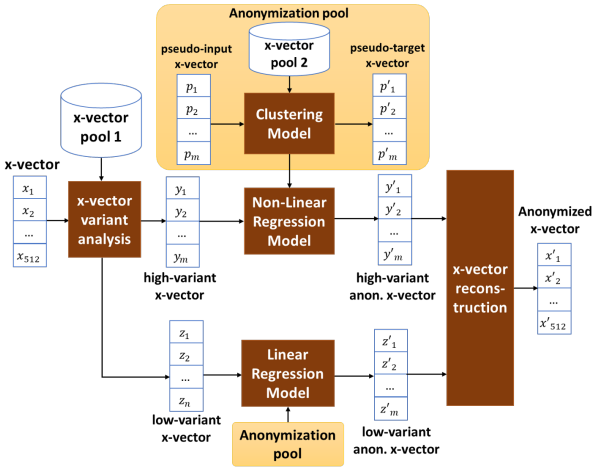

Figure 4: *Schematic diagram of x-vector modification by statistical-based ensemble regression modeling. The dimensions of high-variant x-vector and low-variant x-vector are m and n, respectively (with $m + n = 512$).*

low-variant x-vector (z'). The subscript i in z_i represents the index of z, while A and B are constants obtained from the training process with the low-variant anonymization pool. The original high-variant x-vector (y) was transformed to the anonymized high-variant x-vector (y') with a non-linear regression model trained by the high-variant anonymization pool. From this step, two pre-trained regression models were obtained.

Anonymized X-vector Reconstruction. Lastly, we concatenated the high-variant and low-variant anonymized x-vectors (y' and z') to form the anonymized x-vector (x').

3. Experiments

3.1. Datasets

All datasets utilized in the experiments were based on the VoicePrivacy 2020 Challenge [12]. Table 1 shows the training data used as our x-vector pools. In the variant analysis, we randomly selected subsets of libriTTS train-other-500 and train-clean-100 [21] with 30 utterances from each speaker (with 30 total speakers per dataset). The full set of x-vectors extracted from train-other-500 was then utilized in the singular value (SV) pool and pool-2. In x-vector pool-2, we also utilized the full set of train-clean-100. We fit our regression models with 95% of the total data and the remaining 5% was used to evaluate our models by R^2 score and root-mean squared error (RMSE). The development and test sets of LibriSpeech (Libri) [22] and VCTK [23] were utilized to evaluate speaker verifiability (ASVeval) and intelligibility (ASReval).

Table 1: *Training data for x-vector pools*

Subset	LibriTTS (x-vector)	Female	Male	Total	#Utter
SV pool	train-clean-100	123	124	247	33,236
	train-other-500	560	600	1160	205,044
pool-1	train-clean (rand)	15	15	30	600
	train-other (rand)	15	15	30	600
pool-2	train-clean-100	123	124	247	33,236
	train-other-500	560	600	1160	205,044

Table 2: *Comparison of F0 extractors in terms of intelligibility assessment*

Subset	Data	WER (%)				
		Yaapt	DIO	Harvest	RAPT	SWIPE
Libri (dev)	ori	3.83	3.82	3.82	3.82	3.82
	resyn	6.5	6.77	6.52	6.59	6.41

3.2. Experimental Setup

The main part of our experiment was conducted using the Kaldi toolkit [24]. WORLD and Speech Processing Toolkit (SPTK) were used to extract F0. We investigated four different F0 estimation algorithms, i.e., DIO and Harvest from WORLD, and RAPT and SWIPE from SPTK. In addition, we used Scikit-learn [25] to construct the machine learning model for our anonymization system.

To generate the anonymization x-vector pool, we employed a Gaussian mixture model as our clustering model. We investigated the variation in the number of clusters and the measured x-vector similarity (cosine distance or probabilistic linear discriminant analysis (PLDA)) to build the anonymization pool as the input for the regression model. Consequently, we built gender-dependent regression models that mapped the high-variant and low-variant parts (as in second approach) of the x-vector. The RandomForest algorithm was used as the regression models for the high-variant x-vector. RandomForest uses ensemble learning and can produce a highly accurate model and control overfitting even when the amount of data is large [25]. In our experiment, we tuned the RandomForest regressor parameters, i.e., the number of estimators ($\mathrm{n_{est}}$) and maximum depths ($\mathrm{max_{depth}}$). Our final model used $\mathrm{n_{est}} = 10$ and the default $\mathrm{max_{depth}}$ from scikit-learn since it performed oprimally in our evaluation (in predicting 5% of the training data). For the low-variant x-vector linear regressor, we investigated several pairs of constants A and B for each gender obtained from the parameter variation used while constructing the anonymization pool.

3.3. Results

We evaluated each component of our proposed model by conducting an ablation test. Table 2 shows the evaluation of the resynthesis process using NSF with several F0 estimators (Kaldi (Yaapt), WORLD (DIO and Harvest), and SPTK (RAPT and SWIPE)). The intelligibility assessment (ASReval) was con-

Table 3: *Ablation test on proposed model in terms of objective speaker verifiability. The anonymization system was built using a RandomForest regressor which trained by anonymization pool with GMM clustering with 200 centroids and PLDA scoring. For anonymization model 1, the singular value threshold was between 10% to 20%. The F0 was extracted by the SWIPE algorithm in SPTK. Since the results of model 1 were more optimal than model 2, only the combination of F0 modification with model 1 is reported in this table. Gen stands for gender (F: female and M: male).*

Dataset	Gen	Anonymization		F0 (Resynthesis)			Anon. Model 1			Anon. Model 2			F0 + Anon. Model 1		
		Enroll	Trial	EER (%)	Cllr	Cllr	EER (%)	Cllr	Cllr	EER (%)	Cllr	Cllr	EER (%)	Cllr	Cllr
Libri (dev)	F	ori	ori	8.67	0.31	42.93	8.67	0.31	42.93	8.67	0.31	42.93	8.67	0.31	42.93
			anon	27.56	0.77	116.28	51.99	1.00	147.21	47.16	0.99	167.31	50.99	1.00	145.09
		anon		22.02	0.66	14.25	32.95	0.86	14.25	33.10	0.87	16.78	33.81	0.87	13.55
	M	ori	ori	1.24	0.04	14.28	1.24	0.04	14.28	1.24	0.04	14.28	1.24	0.04	14.28
			anon	24.84	0.72	115.02	58.70	1.00	170.42	56.37	1.00	167.00	55.90	1.00	167.71
		anon		20.34	0.60	8.54	28.88	0.78	18.43	33.85	0.87	23.34	29.19	0.79	17.92
Libri (test)	F	ori	ori	7.66	0.18	26.80	7.66	0.18	26.80	7.66	0.18	26.80	7.66	0.18	26.80
			anon	27.55	0.73	117.11	48.72	1.00	151.98	50.00	1.00	165.24	47.99	1.00	152.84
		anon		18.80	0.58	10.77	28.65	0.78	12.73	31.02	0.81	16.97	28.10	0.76	10.91
	M	ori	ori	1.11	0.04	15.34	1.11	0.04	15.34	1.11	0.04	15.34	1.11	0.04	15.34
			anon	21.38	0.66	121.58	54.34	1.00	168.93	50.78	1.00	165.51	51.45	1.00	166.52
		anon		19.82	0.61	9.29	30.73	0.81	24.20	34.74	0.89	33.09	31.18	0.81	21.89
VCTK common (dev)	F	ori	ori	2.62	0.09	0.87	2.62	0.09	0.87	2.62	0.09	0.87	2.62	0.09	0.87
			anon	30.81	0.80	98.98	50.87	1.00	167.48	49.71	0.99	184.60	50.00	1.00	163.63
		anon		18.02	0.53	5.72	24.42	0.70	7.12	23.84	0.72	9.44	25.58	0.73	8.07
	M	ori	ori	1.43	0.05	1.57	1.43	0.05	1.57	1.43	0.05	1.57	1.43	0.05	1.57
			anon	23.36	0.63	110.95	57.26	1.00	191.60	52.99	1.00	189.56	55.27	1.00	188.91
		anon		15.38	0.48	6.33	25.93	0.71	18.20	31.05	0.82	22.44	28.49	0.75	16.38
VCTK diff (dev)	F	ori	ori	2.92	0.10	1.15	2.92	0.10	1.15	2.92	0.10	1.15	2.92	0.10	1.15
			anon	30.60	0.79	113.40	50.70	0.99	164.37	50.53	0.97	175.98	50.76	0.99	162.20
		anon		16.56	0.53	4.66	26.78	0.77	8.72	28.52	0.81	12.14	26.50	0.77	8.99
	M	ori	ori	1.44	0.05	1.16	1.44	0.05	1.16	1.44	0.05	1.16	1.44	0.05	1.16
			anon	22.43	0.69	104.46	55.98	1.00	166.42	52.56	1.00	164.59	54.74	1.00	163.99
		anon		15.73	0.52	10.28	25.31	0.74	18.28	30.22	0.83	21.75	27.20	0.78	18.17
VCTK common (test)	F	ori	ori	2.89	0.09	0.86	2.89	0.09	0.86	2.89	0.09	0.86	2.89	0.09	0.86
			anon	27.17	0.74	89.41	48.84	0.99	157.68	46.82	0.99	155.28	48.55	0.99	157.68
		anon		21.10	0.60	5.88	28.61	0.80	8.81	32.37	0.86	10.86	28.61	0.80	8.82
	M	ori	ori	1.13	0.04	1.03	1.13	0.04	1.03	1.13	0.04	1.03	1.13	0.04	1.03
			anon	21.75	0.64	118.50	55.65	1.00	186.48	53.39	1.00	187.84	55.37	1.00	186.50
		anon		11.86	0.40	4.28	20.34	0.62	9.79	28.53	0.78	19.91	20.34	0.62	9.78
VCTK diff (test)	F	ori	ori	4.99	0.17	1.50	4.99	0.17	1.50	4.99	0.17	1.50	4.99	0.17	1.50
			anon	27.78	0.77	97.37	49.64	1.00	142.88	48.10	1.00	140.78	49.54	1.00	142.87
		anon		18.21	0.58	6.95	32.66	0.87	11.36	34.67	0.90	12.20	32.77	0.87	11.36
	M	ori	ori	2.07	0.07	1.82	2.07	0.07	1.82	2.07	0.07	1.82	2.07	0.07	1.82
			anon	26.98	0.75	112.89	54.31	1.00	164.68	52.76	1.00	166.21	54.31	1.00	164.69
		anon		16.65	0.54	9.94	21.81	0.67	13.26	32.32	0.86	21.54	21.81	0.67	13.25

Table 4: *Speaker intelligibility attained by the pretrained ASReval model*

Subset	Data	WER (%)			
		F0	Anon 1	Anon 2	F0+Anon 1
Libri (dev)	ori	3.82	3.82	3.82	3.82
	anon	6.41	6.67	6.4	6.41
Libri (test)	ori	4.15	4.15	4.15	4.15
	anon	6.78	6.76	6.63	6.78
VCTK (dev)	ori	10.79	10.79	10.79	10.79
	anon	15.35	15.46	15.55	15.16
VCTK (test)	ori	12.81	12.81	12.81	12.81
	anon	15.21	15.31	15.65	15.32

ducted using the LibriSpeech development set. Table 3 shows the evaluation results for speaker verifiability, which include equal error rate (EER) and log-likelihood-ratio cost function (Cllr) metrics of the system from F0 modification (only resynthesis), for both anonymization models and a combination of F0 modification and the anonymization model. Additionally, Table 4 shows the objective intelligibility evaluation in terms of word error rate (WER) in the ASR evaluation system (ASReval).

The results highlight two main findings. First, speech distortion occurs in the analysis-synthesis process using NSF with an x-vector. When only resynthesis was conducted, the intelligibility metric of the output speech decreased (WER increased), as shown in Table 2. The performance of several F0 extractors were not significantly affected in terms of objective intelligibility metric. Since the resynthesis process tends to alter the speech, the resynthesis process itself contributes to the anonymization process, as shown in Table 3 (F0 (resynthesis) ASVeval). The ASV objective metrics of anonymization in the pair enrollment-trials, ori-ori and ori-anon, differed significantly (e.g., EER rate increased by more than 15% in all cases). Second, our proposed anonymization model improved the objective anonymization metrics compared with when only resynthesis was used. Compared with the first baseline system, our proposed method (first approach) was more effective but not significantly so in terms of the objective verifiability metrics. We predict that this limitation is caused by the limited amount of training data and the similarity in the main frameworks of the analysis-synthesis process of the baseline system.

4. Conclusion and Future Work

We proposed two x-vector anonymization approaches: singular value modification and statistical-based decomposition with regression models. The main concept was that one-to-one mapping from input speech to anonymized speech could be obtained by modifying the significant elements of the x-vector. The evaluation results demonstrated that our proposed anonymization system was effective in increasing the anonymization rate (ASVeval) compared with resynthesis only. We intend to increase the amount of training data and study state-of-the-art regression models for anonymizing x-vector to improve our system. We will also investigate how to construct an analysis-synthesis system that better suits the anonymization process.

5. Acknowledgements

This work was supported by a Grant-in-Aid for Scientific Research (B) (No. 17H01761) and JSPS KAKENHI Grant (No. 20J20580).

6. References

[1] 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 16th ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 82–94. [Online]. Available: https://doi.org/10.1145/3274783.3274855

[2] S. O. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," 2018.

[3] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, and L.-R. Dai, "Wavenet vocoder with limited training data for voice conversion," in *INTERSPEECH*, 2018, pp. 1983–1987.

[4] F. Fang, X. Z. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker anonymization using x-vector and neural waveform models," *ArXiv*, vol. abs/1905.13561, 2019.

[5] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J.-F. Bonastre, P.-G. Noé, M. Todisco *et al.*, *The VoicePrivacy 2020 Challenge Evaluation Plan*, 2020. [Online]. Available: https://www.voiceprivacychallenge.org/docs/VoicePrivacy_2020_Eval_Plan_v1_2.pdf

[6] Q. Jin, A. R. Toth, A. W. Black, and T. Schultz, "Is voice transformation a threat to speaker identification?" in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4845–4848.

[7] Q. Jin, A. R. Toth, T. Schultz, and A. W. Black, "Speaker de-identification via voice transformation," in *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2009, pp. 529–533.

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

[9] M. Abou-Zleikha, Z.-H. Tan, M. G. Christensen, and S. H. Jensen, "A discriminative approach for speaker selection in speaker de-identification systems," in *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, 2015, pp. 2102–2106.

[10] C. Magariños, 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.

[11] 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)*, 2018, pp. 5329–5333.

[12] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "Introducing the voiceprivacy initiative," 2020.

[13] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH*, 2015.

[14] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," 2017.

[15] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," 2018.

[16] X. Wang and J. Yamagishi, "Neural harmonic-plus-noise waveform model with trainable maximum voice frequency for text-to-speech synthesis," 2019.

[17] 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. E99.D, pp. 1877–1884, 2016.

[18] D. Jouvet and Y. Laprie, "Performance analysis of several pitch detection algorithms on simulated and real noisy speech data," in *2017 25th European Signal Processing Conference (EUSIPCO)*, 2017, pp. 1614–1618.

[19] K. Tokuda, K. Oura, Y. Takenori, A. Tamamori, S. Sako, H. Zen, T. Nose, T. Takahashi, J. Yamagishi, and Y. Nankaku, *Speech Signal Processing Toolkit (SPTK) Version 3.11*, 2017. [Online]. Available: http://sp-tk.sourceforge.net/

[20] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, 2016.

[21] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," 2019.

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

[23] C. Veaux, J. Yamagishi, and K. Macdonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," in *arXiv*, 2017.

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.