F_0 Modification via PV-TSM Algorithm for Speaker Anonymization Across Gender

Candy Olivia Mawalim, Shogo Okada, and Masashi Unoki
Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa 923–1292 Japan
Email: candyolivia@jaist.ac.jp, okada-s@jaist.ac.jp, unoki@jaist.ac.jp

***Abstract*—Speaker anonymization has been developed to protect personally identifiable information while retaining other encapsulated information in speech. Datasets, metrics, and protocols for evaluating speaker anonymization have been defined in the Voice Privacy Challenge (VPC). However, existing privacy metrics focus on evaluating general speaker individuality anonymization, which is represented by an x-vector. This study aims to investigate the effect of anonymization on the perception of gender. Understanding how anonymization caused gender transformation is essential for various applications of speaker anonymization. We proposed speaker anonymization methods across genders based on phase-vocoder time-scale modification (PV-TSM). Subsequently, in addition to the VPC evaluation, we developed a gender classifier to evaluate a speaker's gender anonymization. The objective evaluation results showed that our proposed method can successfully anonymize gender. In addition, our proposed methods outperformed the signal processing-based baseline methods in anonymizing speaker individuality represented by the x-vector in ASVeval while maintaining speech intelligibility.**

I. Introduction

In recent years, speech communication has been significantly developed and utilized in daily applications. However, many novice users are still unaware of the privacy issues that may be caused by publicly distributed speech. Speech encapsulates linguistic-related and biometric-related content [1]. Therefore, it is vulnerable to being misused by an unauthorized person (attacker), e.g., fake speech created with an advanced speech synthesizer [2], [3]. Consequently, research on protecting against the emerging threat caused by voice privacy violations is essential.

One of the solutions is using the speaker anonymization approach defined in the Voice Privacy Challenge (VPC) [4], [5], [6]. Speaker anonymization aims to conceal personal identifiable information (PII) while retaining other information. The PII in the VPC is referred to as a state-of-the-art feature in speech biometric studies and is used for developing automatic speaker verification (ASV) systems. Subsequently, reliable speaker anonymization could be achieved when the anonymized speech causes high errors in the ASV system. In other words, the anonymized speech from a given speaker should not resemble any existing speech known by the ASV system [7].

In an earlier study, methods for anonymizing PII in speech consisted of combinations of prosodic and spectral modifications using pitch-synchronous overlap-add algorithm [8] and the voice transformation techniques described in [9], [10]. Next, the method proposed by Pobar and Ipsic [11] utilized Gaussian mixture model mapping and harmonic-stochastic models to anonymize speech individuality. Furthermore, VPC 2020 proposed two baseline speaker anonymization systems [4]. A neural source-filter (NSF) model and state-of-the-art x-vector speaker embedding were utilized in the primary baseline system [12]. Anonymization is performed through the x-vector that represents PII. Furthermore, modification of the McAdams coefficient using linear prediction analysis was carried out in the secondary baseline system [13], [14]. The secondary baseline is not a machine learning approach; thus, it does not require a training process. However, the overall speaker anonymization evaluation results showed that the primary baseline system outperformed the secondary baseline.

In our previous work [15], we investigated methods based on the time-scale modification (TSM) signal processing approach for speaker anonymization according to the VPC protocols. Unlike vocoder-based systems, such as those from both baseline systems, the TSM-based approach synthesizes speech via frame relocation and adaptation [16]. Although it cannot be used to analyze pitch and timbre independently, the TSM-based approach was reported to output better quality voices than those from conventional vocoders [17]. The objective evaluation results showed that the TSM-based method using phase propagation (PV-TSM) outperformed the secondary baseline system [15].

Gender is one of the pieces of information included in PII. How speaker anonymization caused a specific PII (e.g., gender) perception change has not been defined in the VPC evaluation. On the other hand, several prior studies indicated that gender recognition or selection is necessary for speaker anonymization systems [7], [18], [19], [20], [21]. In addition, the general goal of speaker anonymization is to change the speaker identity to a pseudotarget speaker that has different characteristics than the source speaker (which simply causes a change in gender) [18].

In this study, we aim to investigate the effect of the F_0 modification by speaker anonymization across genders using the PV-TSM algorithm. The anonymization effect on gender perception has yet to be investigated in [15]. Understanding how anonymization caused gender transformation is essential for speaker anonymization and its applications. Furthermore, unlike other prior works on gender anonymization, we devel-

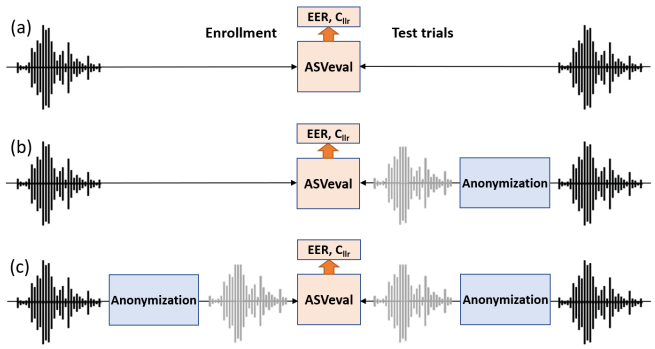

Fig. 1: VPC 2020 ASV evaluation for (a) original trial and original enrollment (o-o), (b) original enrollment and anonymized trial (o-a), and (c) anonymized enrollment and trial (a-a) [4]

oped a voice gender classifier to evaluate gender perception change in addition to an extensive speaker anonymization evaluation in the VPC.

The remaining parts of this paper are organized as follows. Section 2 discusses the speaker anonymization system according to the VPC. Section 3 introduces the PV-TSM algorithm for voice gender anonymization. Section 4 describes our experiments, including the datasets, experimental setup, and discussion of the results. Finally, Section 5 concludes the paper and discusses our future work.

II. Speaker Anonymization

A. Definition

Speaker anonymization or speaker deidentification is a method for concealing the personally identifiable information of a given speaker and aims to protect voice privacy while maintaining other information, such as linguistic information [4], [5]. The VPC initiates the generalization of the speaker anonymization task and metrics [4]. Speaker anonymization should satisfy all of the following requirements:

1) It outputs a speech waveform,
2) the speaker identity should be concealed,
3) the output speech should be natural and intelligible, and
4) anonymized utterances of a given speaker should be perceived as unique from those of other speakers.

Several open-source corpora are utilized in the VPC to develop speaker anonymization systems, including LibriSpeech [22], LibriTTS [23], VCTK [24], and VoxCeleb-1,2 [25], [26]. The detailed description and statistics of these datasets were explained in the VPC 2020 evaluation plan [4]. The recent VPC (VPC 2022) introduced a modified primary baseline system, an attack model scenario, and additional metrics for evaluating speaker anonymization [5]. More details about the update in VPC 2022 are explained in the following subsections.

B. Evaluation Metrics

Assessment of an anonymization system consists of privacy and utility metrics. The speaker verifiability is used for measuring the privacy metric. The utility metric measures how speaker anonymization preserves characteristics other than the PII of the given voice. VPC 2020 [4] introduced an automatic speaker verification (ASV) system for evaluating speaker verifiability and an automatic speech recognition (ASR) system for evaluating utility metrics. Hereafter, we refer to the ASV system as ASVeval 2020 and the ASR system as ASReval. Both of the systems are trained on the Kaldi toolkit using a subset of the LibriSpeech dataset (LibriSpeech-train-clean-360) [27].

In VPC 2022 [5], a semi-informed attack model was introduced to evaluate speaker verifiability when an attacker has prior information about the original speech for enrollment data and the speaker anonymization system. Hereafter, we refer to the ASV system for a semi-informed attack as ASVeval 2022. In addition, two complementary utility metrics are also introduced, namely, pitch correlation (ρ^{F_0}) and gain of voice distinctiveness $\mathrm{G_{VD}}$. Pitch correlation is used as a measurement to check if the anonymization method preserves the original utterance. Furthermore, the gain of voice distinctiveness measures the voice similarity preservation of utterances from the same speaker after anonymization.

ASVeval was developed by utilizing probabilistic linear discriminant analysis (PLDA) on the x-vector (state-of-the-art speaker embedding) [28]. In ASVeval, the equal error rate (EER) and log-likelihood-ratio cost function (C_{llr} and C_{llr}^{min}, as proposed in [29]) are computed as the objective verifiability metrics. There are three scenarios for ASVeval-2020: o–o, o–a, and a–a, as shown in Fig. 1. Moreover, for ASVeval 2022, the attack model is evaluated only in the a–a scenario with anonymized data for training the ASV system. ASReval was developed based on a factorized time delay neural network (TDNN-F) acoustic model [12], [30] with a trigram language model using a Kaldi recipe for the LibriSpeech dataset. The word error rate (WER) is used to measure the speech intelligibility of the output anonymized speech.

C. Baseline Systems

In VPC 2020, two speaker anonymization systems were introduced as the baseline systems [4]. The primary baseline (B1a) system is based on a deep learning approach using x-vectors and an NSF model [12]. The second baseline (B2a) system is based on a signal processing approach using the McAdams coefficient [13] in linear prediction analysis.

The idea of B1a is to separate linguistic content and speaker individuality features from the input speech and then anonymize the extracted speaker individuality features. The B1a system consists of an F_0 extractor, an ASR acoustic model, an x-vector extractor, an x-vector anonymization model, a pool of x-vectors, a speech synthesis AM, and an NSF model. The x-vector anonymization process is as follows:

1) Feature extraction: extraction of F_0, a bottleneck feature (as linguistic feature representation using an ASR acoustic model [12], [30]) and a speaker individuality feature (x-vector based on [28]);

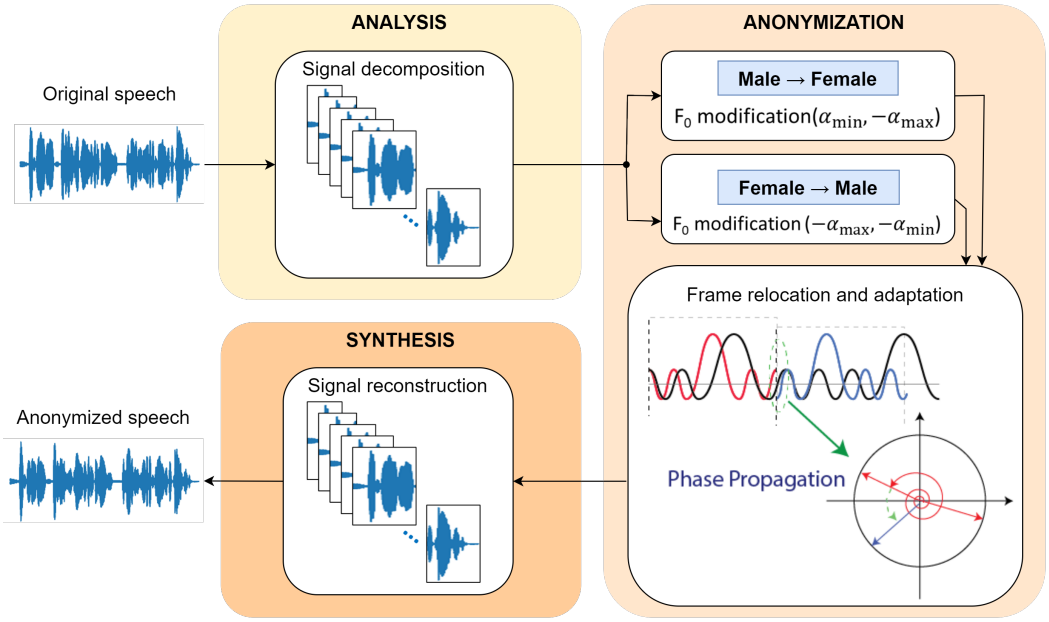

Fig. 2: Block diagram of the proposed method.

2) X-vector anonymization: modification of the extracted x-vector by averaging a set of candidate x-vectors from the pool of x-vectors; and
3) Speech synthesis: speech synthesis using F_0, the bottleneck features, and the modified/anonymized x-vector based on the speech synthesis AM [12] and an NSF [3] model.

Additionally, extended primary and secondary baseline systems were introduced in VPC 2022 (B1b and B2b). B1b utilized a unified HiFi-GAN NSF model as the speech synthesizer. On the other hand, B2b utilized a uniformly randomized value of the McAdams coefficient ($\alpha \sim U(0.5, 0.9)$).

III. Proposed Method

TSM algorithms are signal processing algorithms that compress or stretch audio signals [16]. They are often used in music processing as well as for speech synthesizers [16], [17], [31]. This study utilized a TSM algorithm based on a phase vocoder (PV-TSM) and pitch shifting for speaker anonymization [15]. The implementation of the PV-TSM algorithm is based on [16]. The PV-TSM improves speech quality by reducing the artifacts that occur in synthesized speech via phase propagation [16]. Figure 2 shows the block diagram of our proposed method.

The modification of the F_0 trajectory affects the perception of speaker individuality [32]. For instance, the F_0 range of female speakers is generally higher than the F_0 range of male speakers. We evaluate the effect of F_0 modification on speaker anonymization using PV-TSM. The anonymization process follows the following steps:

- First, the original signal ($x(r)$) is decomposed into a number of frames ($x_m(r)$). The original frame is then resampled to increase or decrease F_0 with a factor of α as follows:

$$\alpha = F_{0y_m(r)}/F_{0x_m(r)}. \quad (1)$$

where $m \in \mathbb{Z}$ is the frame index, the sample $r \in [0, L-1]$, and L is the signal length.

- Next, the short-time Fourier transform (STFT) is performed to obtain the frequency spectra of the input signal X after resampling. This process is expressed as follows:

$$X(m,k) = \sum_{r=-N/2}^{N/2-1} x'_m(r)w(r)\exp(-2\pi ikr/N), \quad (2)$$

where k is the frequency index ($k \in [0, N-1]$), N is the frame length, $x'_m(r)$ is the input signal frame after resampling, $w(r)$ is a Hann window function, and i is imaginary unit. Complex $X(m,k)$ can also be expressed as the combination of a magnitude $|X(m,k)| \in \mathbb{R}^+$ and a phase $\varphi(m,k) \in [0,1)$:

$$X(m,k) = |X(m,k)|\exp(2\pi i\varphi(m,k)). \quad (3)$$

- To reconstruct the output signal $y(r)$, we need to concatenate the time-domain frames x_m^{Mod} by using the inverse Fourier transform. However, this process caused phase errors or phase jumps in each overlapping frame (as shown in the frame relocation and adaptation process in Fig. 2). Consequently, the phase jumps in each overlapping frame are fixed via phase propagation $\varphi^{\text{Mod}}(m,k)$. More detail of phase propagation is available in [16].

$$X^{\text{Mod}}(m,k) = |X(m,k)|\exp(2\pi i\varphi^{\text{Mod}}(m,k)). \quad (4)$$

- Subsequently, we derive the time-domain frames x_m^{Mod} by using the inverse Fourier transform as follows:

$$x_m^{\text{Mod}}(r) = \frac{1}{N}\sum_{k=0}^{N-1} X^{\text{Mod}}(m,k)\exp(2\pi ikr/N). \quad (5)$$

- Finally, the signal $(y'_m(r))$ is reconstructed from the frequency spectra after phase updating X^{Mod}, and it is calculated as follows:

$$y'_m(r) = \frac{w(r)x_m^{\text{Mod}}(r)}{\sum_{n\in\mathbb{Z}} w(r-nH_s)^2}, \tag{6}$$

where H_s is the synthesis hop size. The output anonymized speech $y(r)$ is obtained by concatenating the time-domain frames $y'_m(r)$ as follows:

$$y(r) = \sum_{m\in\mathbb{Z}} y'_m(r - mH_s). \tag{7}$$

Several works on F_0 modification for speaker anonymization across gender have been proposed [7], [21]. The results showed that speaker verifiability could be improved by cross-gender transformation. Unlike our prior work [15], we carefully investigate the effect of gender perception caused by F_0 modification. We carry out a nonlinear transformation for F_0 modification with a factor α in the unit of semitone. We decrease F_0 by α for anonymizing a female voice, while we increase F_0 by α for anonymizing a male voice. Mathematically, α is expressed by:

$$\alpha(n) = 2^{n/12}, \tag{8}$$

where n is the number of semitones for modification. The ranges of semitones investigated in this study are $n \in [1, 2]$, $n \in [2, 3]$, and $n \in [3, 4]$. We later define the proposed methods with the respective ranges as "PV-TSM-12", "PV-TSM-23", and "PV-TSM-34".

IV. Evaluation

To evaluate our proposed methods, we carried out an objective evaluation based on the protocols and datasets provided in the VPC [4], [5]. Additionally, we also developed a gender classifier based on state-of-the-art x-vector embedding to objectively evaluate the effect on gender perception after anonymization. The description of the dataset, gender classifier, and evaluation based on the VPC are subsequently explained.

A. Dataset

We utilized three publicly available open-source corpora for evaluation. Two corpora were included in the VPC [4], [5]: LibriSpeech [22] and VCTK [24]. LibriSpeech is an English corpus that was designed for ASR development. On the other hand, VCTK is also an English corpus spoken by 109 speakers with various accents and was designed for text-to-speech (TTS) development. Each corpus was split into nonoverlapping training, development, and testing data. The number of female and male speakers in both corpora are relatively balanced (approximately 50–55% female and 45–50% male). For the VCTK dataset, a "common part" and "different part" are defined in the VPC to evaluate speaker verifiability regardless of text dependency. The common part consisted of identical utterances spoken by multiple speakers, while the different part consisted of distinct utterances spoken by multiple speakers. We utilized these two corpora for the privacy and utility evaluation based on the VPC. Additionally, we also utilized the TIMIT dataset [33] in addition to the LibriSpeech dataset for constructing the objective gender classifier. TIMIT is an English corpus with 630 speakers speaking in various dialects and was designed for the evaluation of an ASR system. Unlike the LibriSpeech and VCTK corpora, the comparison between female and male speakers in the TIMIT corpus is quite imbalanced (70% male and 30% female). The TIMIT corpus was incorporated to develop a more general classifier that covers more dialects.

B. Voice Gender Perception Evaluation

The evaluation of voice gender perception assumes that if the anonymized speech affects the perception of gender from the original gender existing in the dataset (female or male), the gender classification accuracy will be reduced. For this evaluation, we construct a binary classifier based on the state-of-the-art x-vector speaker embedding model [28]. We fine-tuned the time-delay neural network (TDNN) model that was coupled with statistical pooling, and the model was trained on the VoxCeleb dataset with categorical cross-entropy loss. The SpeechBrain[1] toolkit was utilized to develop the gender classifier [34]. The training data used for fine-tuning comprises a subset of the LibriSpeech dataset (train-clean-5) and the TIMIT dataset. The total number of speakers for training is 490, while for testing, it is 174. The data for training and testing do not overlap. The overall classification accuracy of the testing phase in terms of the F1 score is 95.20%. Furthermore, we evaluate the anonymized speech using this classifier.

To compare the effect of anonymization with regard to gender, we calculate the accuracy of anonymized speech spoken by female and male speakers separately and the overall F1 score for various shift factors α. Figure 3 shows the results of the gender classification task using the LibriSpeech dataset. The results show that by increasing or decreasing the F_0 trajectory, the gender perception of the source speakers changed (the accuracy decreased). This means that gender anonymization is successful, and the results improved to some extent when we increased the F_0 gap to the original signal based on our assumption in this evaluation.

C. The VPC Evaluation

We also carried out the VPC evaluation, as described in Subsection II-B. Since our proposed method is based on a signal processing approach, we did not train any machine learning models using the development data. In other words, the development data used for evaluation can be regarded as another type of testing data. Additionally, we compared the performance of the proposed methods with that of the secondary baselines introduced in VPC 2020 (B2a) and in VPC 2022 (B2b), which were also based on a signal processing approach.

[1] https://speechbrain.github.io/

TABLE I: ASVeval of the VPC 2020 results for the o-a scenario. The direction of the arrow indicates the criteria of a better anonymization performance.

Dataset		Gender	Weight	EER (%) ↑					
				Orig	B2a	B2b	PV-TSM-12	PV-TSM-23	PV-TSM-34
Source speaker: Female									
Dev	Libri	female	0.50	8.81	35.37	37.93	26.28	35.51	41.34
	VCTK (diff)	female	0.40	2.92	35.43	35.77	31.5	40.31	45.37
	VCTK (comm)	female	0.10	2.62	34.01	36.34	32.27	40.7	47.09
Weighted average dev				**5.84**	**35.26**	**36.91**	**28.97**	**37.95**	**43.53**
Test	Libri	female	0.50	7.66	26.09	31.39	20.07	34.67	35.4
	VCTK (diff)	female	0.40	4.94	29.99	36.32	19.5	39.35	42.9
	VCTK (comm)	female	0.10	2.89	30.92	44.51	18.5	36.42	45.09
Weighted average test				**6.10**	**28.13**	**34.67**	**19.69**	**36.72**	**39.37**
Source speaker: Male									
Dev	Libri	male	0.50	1.24	17.86	38.35	8.39	20.81	32.61
	VCTK (diff)	male	0.40	1.44	28.14	42.33	17.22	33.5	43.87
	VCTK (comm)	male	0.10	1.43	23.93	45.01	13.11	30.48	39.6
Weighted average dev				**1.34**	**22.58**	**40.61**	**12.39**	**26.85**	**37.81**
Test	Libri	male	0.50	1.11	17.82	27.39	12.25	22.94	30.96
	VCTK (diff)	male	0.40	2.07	28.3	38.12	9.36	25.49	39.27
	VCTK (comm)	male	0.10	1.13	24.29	40.68	7.63	25.42	37.57
Weighted average test				**1.50**	**22.66**	**33.01**	**10.63**	**24.21**	**34.95**

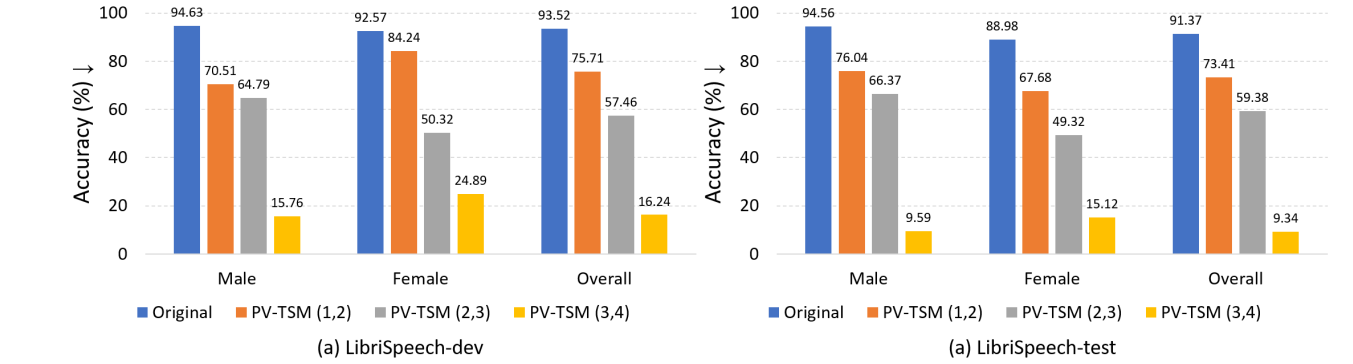

Fig. 3: Results of gender classification on anonymized speech using the LibriSpeech dataset

We utilized the ASVeval in VPC 2020 in three scenarios: (1) original enrollment–original trials (o-o), (2) original enrollment–anonymized trials (o-a), and (3) anonymized enrollment–anonymized trials (a-a). A higher EER in the ASVeval indicates better privacy preservation (as shown in Fig. 1). The Orig. column in Tables I and II shows the results of the ASVeval in the o-o scenario. The other columns in Table I show the results of the ASVeval in the o-a scenario of the corresponding methods with regard to the gender of the source speakers. Furthermore, Table II shows the ASVeval results in the a-a scenario. The a-a scenario considers the "lazy-informed" attack model (without retraining the ASV with anonymized speech). We also conducted the evaluation using ASVeval 2022 (considering the "semi-informed" attack model). However, since the results are similar to those reported in [15] and our aim in this study is to investigate the effect of anonymization on gender perception, we focus on the discussion of speaker verifiability using the results from ASVeval 2020.

The ASVeval 2020 results indicate that in the o-a scenario, the proposed methods could give better privacy than the original secondary baseline (B2a) when the shift factor α is higher than (1,2). However, the EER of B2a was reduced significantly in the a-a scenario, which means that privacy is not well preserved when an attacker has access to the black box anonymization algorithm [35]. B2b could significantly improve this shortcoming, but it causes a significant degradation in speech intelligibility (the WER increased from approximately 19% to 55%, as shown in Table III). Increasing the gap between α of the anonymized signal and the original signal results in better privacy for both the o-a and a-a scenarios. For instance, when α is set to (3,4), the weighted average EER in the o-a scenario of anonymized test data is improved to 37%, while the weighted average EER when α is equal to (1,2) and (2,3) are 15% and 30%, respectively.

To evaluate utility, we used the ASReval provided in VPC

TABLE II: ASVeval of the VPC 2020 results for the a-a scenario.

Dataset		Gender	Weight	EER (%) ↑ Orig	B2a	B2b	PV-TSM-12	PV-TSM-23	PV-TSM-34
Source speaker: Female									
Dev	Libri	female	0.50	8.81	23.58	40.62	36.51	49.01	50.85
	VCTK (diff)	female	0.40	2.92	15.78	35.93	40.88	44.97	47.16
	VCTK (comm)	female	0.10	2.62	11.63	54.36	41.28	48.26	49.71
Weighted average dev				**5.84**	**19.27**	**40.12**	**38.74**	**47.32**	**49.26**
Test	Libri	female	0.50	7.66	15.15	42.70	27.37	35.95	40.15
	VCTK (diff)	female	0.40	4.94	16.98	31.02	49.33	49.79	51.44
	VCTK (comm)	female	0.10	2.89	14.45	38.15	48.55	50.87	52.89
Weighted average test				**6.10**	**15.81**	**37.57**	**38.27**	**42.98**	**45.94**
Source speaker: Male									
Dev	Libri	male	0.50	1.24	10.56	43.63	29.04	42.39	47.67
	VCTK (diff)	male	0.40	1.44	11.12	43.37	31.71	31.86	31.22
	VCTK (comm)	male	0.10	1.43	10.54	46.44	36.75	42.45	41.03
Weighted average dev				**1.34**	**10.78**	**43.81**	**30.88**	**38.18**	**40.43**
Test	Libri	male	0.50	1.11	8.46	47.66	29.18	42.76	44.77
	VCTK (diff)	male	0.40	2.07	12.23	38.92	31.80	45.24	51.95
	VCTK (comm)	male	0.10	1.13	11.86	46.61	33.05	46.61	47.18
Weighted average test				**1.50**	**10.31**	**44.06**	**30.62**	**44.14**	**47.88**

TABLE III: ASReval of the VPC 2020 results.

Dataset		WER (%) ↓ Orig	B2a	B2b	PV-TSM-12	PV-TSM-23	PV-TSM-34
Dev	Libri	3.83	8.74	36.42	4.59	5.23	6.16
	VCTK	10.79	25.56	52.09	12.59	14.32	16.51
Average-dev		**7.31**	**17.15**	**44.26**	**8.59**	**9.78**	**11.34**
Test	Libri	4.14	8.90	48.12	4.85	5.53	6.26
	VCTK	12.81	28.15	62.35	14.97	16.49	18.66
Average-test		**8.48**	**18.53**	**55.24**	**9.91**	**11.01**	**12.46**

2020. Table III shows the utility evaluation results in terms of speech intelligibility (the WER). High utility is achieved when speech intelligibility can be preserved (the WER is as close to the original utterances as possible). Based on this definition, our results indicate that the speech intelligibility of anonymized signals obtained by PV-TSM methods are better than those obtained by the speaker anonymization methods that use the McAdams coefficient (B2a and B2b). Increasing the gap between α of the anonymized signal and the original signal caused slightly more distortion (the WER increased by $< 5\%$).

Furthermore, we illustrate the performance in terms of the privacy metric and utility metrics in Fig. 4 on four different categories based on gender and enrollment–trials scenarios. The first row shows the privacy versus utility results in the o-a scenario. Moreover, the second row shows the privacy versus utility results in the a-a scenario. All subfigures indicate that the proposed methods can better balance the privacy and utility metrics than the secondary baseline methods (B2a and B2b). The privacy preservation of B2b is comparable to that of PV-TSM-23; however, its utility preservation is greatly reduced. Considering the gender of the source speakers, the anonymization of female speakers has a better balance of privacy and utility than from the anonymization of male speakers in both the o-a and a-a scenarios.

In addition to ASReval, we also carried out an evaluation using the secondary utility metrics described in VPC 2022 [5]. The secondary utility metrics include pitch correlation and the gain of voice distinctiveness. The average pitch correlation values of PV-TSM-12, PV-TSM-23, and PV-TSM-34 are 0.87, 0.85, and 0.82, respectively. The average gain of voice distinctiveness of the proposed methods with any α is between -2.00 and -1.20. The results obtained by these two utility metrics indicate that our proposed method can preserve other speech attributes to some extent, as described in the VPC 2022 evaluation plan [5].

D. Limitations

An extensive objective evaluation was carried out to evaluate the proposed methods. The results of the proposed methods showed superior performance in comparison to the secondary baseline systems. However, the aim of speaker anonymization is to conceal the PII, and using only the objective evaluation results is not conclusive. In addition, there is no ground truth (label) for speaker anonymization. For instance, the objective evaluation based on a gender classifier could only indicate that most of the anonymized female voice was no longer recognized as a female voice (an increased classification error). For further

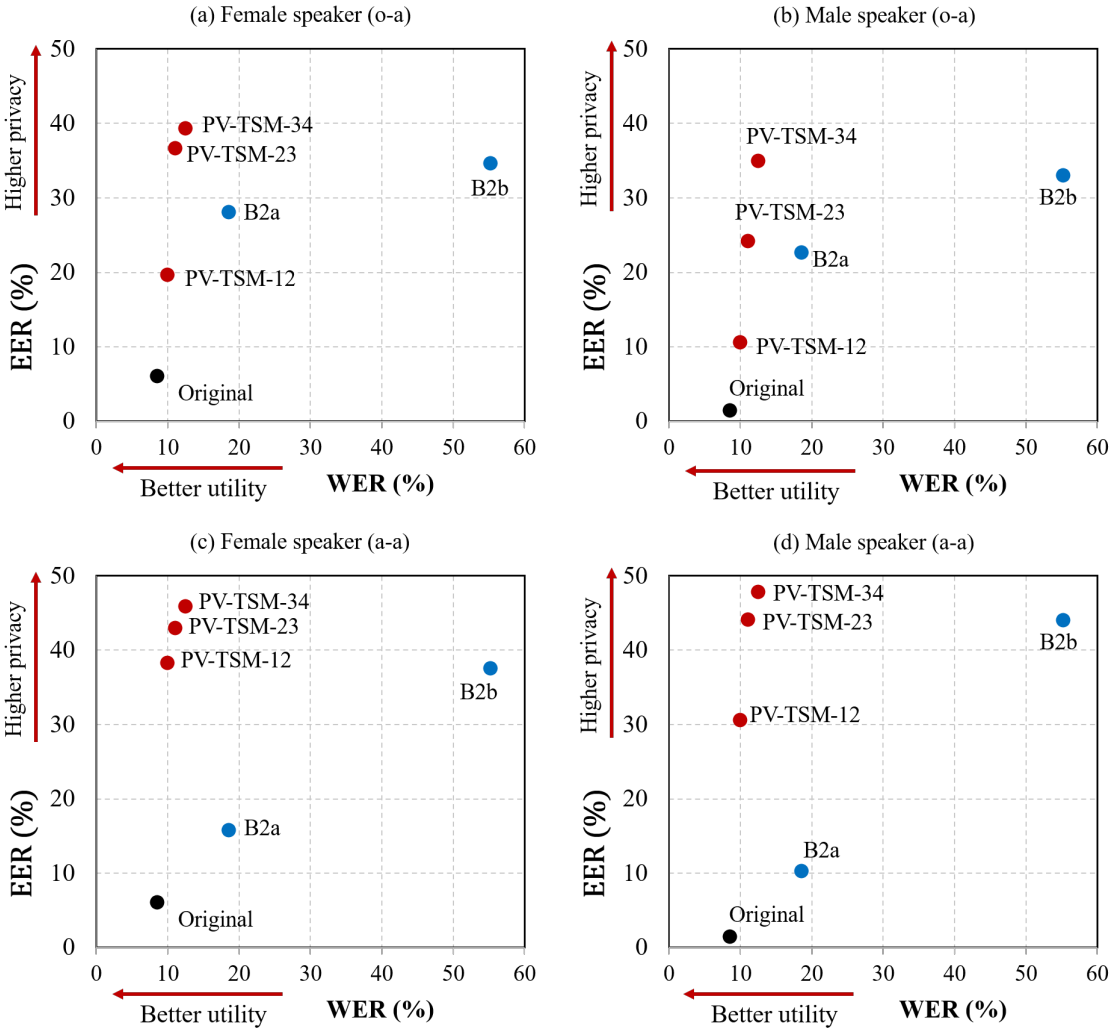

Fig. 4: Privacy vs utility plot based on ASVeval and ASReval of VPC 2020 using the test dataset.

analysis, a subjective evaluation is required to evaluate more detailed anonymization performance in the future.

Another limitation is that the prosodic feature that we considered in this study is limited to F_0. In the future, we will consider other speaker individuality features to improve anonymization performance. Additionally, the PII investigated in this study is limited to the speaker's gender perception. More attributes are included in the PII that are worth further investigation.

V. Conclusions

This study demonstrated the effect of F_0 trajectory modification on speaker anonymization based on PV-TSM in the perception of a speaker's gender. To evaluate the effectiveness of anonymization on gender, we developed a binary classifier based on x-vector speaker embedding. In addition, we also carried out an objective evaluation for speaker anonymization based on the VPC. The results showed that the proposed methods could successfully anonymize a speaker's gender (in terms of gender classification accuracy) and general PII (in terms of privacy and utility metrics in the VPC). In the future, we will carry out a subjective evaluation to carefully investigate speaker anonymization performance. In addition, further attributes in PII will also be investigated as the objective of anonymization.

Acknowledgment

This work was supported by the Fund for the Promotion of Joint International Research (Fostering Joint International Research (B)) (Grant Number 20KK0233), the KDDI Foundation (Research Grant Program), a Japan Society for the Promotion of Science (JSPS) Kakenhi Activity Start-up (Grant Number 22K21304), and the SCAT Research Foundation. This work was also partially supported by the JSPS KAKENHI (Grant Numbers 22H04860 and 22H00536) and JST AIP Trilateral AI Research, Japan (Grant Number JPMJCR20G6).

References

[1] A. Nautsch, C. Jasserand, E. Kindt, M. Todisco, I. Trancoso, and N. Evans, "The GDPR & speech data: Reflections of legal and technology communities, first steps towards a common understanding," in *Interspeech 2019*. ISCA, sep 2019.

[2] F. Fang, J. Yamagishi, I. Echizen, and J. Lorenzo-Trueba, "High-quality nonparallel voice conversion based on cycle-consistent adversarial network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada*. IEEE, 2018, pp. 5279–5283.

[3] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom*. IEEE, 2019, pp. 5916–5920.

[4] 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, visited on 2022-06-02. [Online]. Available: https://www.voiceprivacychallenge.org/vp2020/docs/VoicePrivacy_2020_Eval_Plan_v1_4.pdf

[5] N. Tomashenko, X. Wang, X. Miao, H. Nourtel, P. Champion, M. Todisco, E. Vincent, N. Evans, J. Yamagishi, and J.-F. Bonastre, *The VoicePrivacy 2022 Challenge Evaluation Plan*, 2022, visited on 2022-06-02. [Online]. Available: https://www.voiceprivacychallenge.org/vp2020/docs/VoicePrivacy_2022_Eval_Plan_v1.0.pdf

[6] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O'Brien, A. Chanclu, J.-F. Bonastre, M. Todisco, and M. Maouche, "The voiceprivacy 2020 challenge: Results and findings," *Comput. Speech Lang.*, vol. 74, no. C, jul 2022. [Online]. Available: https://doi.org/10.1016/j.csl.2022.101362

[7] 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, pp. 138–147, 03 2018.

[8] H. Valbret, E. Moulines, and J. Tubach, "Voice transformation using psola technique," *Speech Communication*, vol. 11, no. 2, pp. 175–187, 1992, eurospeech '91. [Online]. Available: https://www.sciencedirect.com/science/article/pii/016763939290012V

[9] Q. Jin, A. R. Toth, A. W. Black, and T. Schultz, "Is voice transformation a threat to speaker identification?" in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2008, Caesars Palace, Las Vegas, Nevada, USA*. IEEE, 2008, pp. 4845–4848.

[10] 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, ASRU 2009, Merano/Meran, Italy, December 13-17, 2009*. IEEE, 2009, pp. 529–533.

[11] M. Pobar and I. Ipsic, "Online speaker de-identification using voice transformation," in *37th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2014, Opatija, Croatia*. IEEE, 2014, pp. 1264–1267.

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

[13] S. McAdams, "Spectral fusion, spectral parsing and the formation of auditory images," *Ph. D. Thesis, Stanford*, 1984.

[14] J. Patino, N. A. Tomashenko, M. Todisco, A. Nautsch, and N. W. D. Evans, "Speaker anonymisation using the McAdams coefficient," *CoRR*, vol. abs/2011.01130, 2020.

[15] C. O. Mawalim, S. Okada, and M. Unoki, "Speaker Anonymization by Pitch Shifting Based on Time-Scale Modification," to appear in the 2nd Symposium on Security and Privacy in Speech Communication joined with 2nd VoicePrivacy Challenge (SPSC 2022), 2022.

[16] J. Driedger and M. Müller, "A review of time-scale modification of music signals," *Applied Sciences*, vol. 6, p. 57, 2016.

[17] M. Morise, "Platinum: A method to extract excitation signals for voice synthesis system," *Acoustical Science and Technology*, vol. 33, no. 2, pp. 123–125, 2012.

[18] 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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230816302959

[19] M. Faundez-Zanuy, E. Sesa-Nogueras, and S. Marinozzi, "Speaker identification experiments under gender de-identification," in *2015 International Carnahan Conference on Security Technology (ICCST)*, 2015, pp. 1–6.

[20] 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)*, 2015, pp. 2102–2106.

[21] P. Champion, D. Jouvet, and A. Larcher, "A study of f0 modification for x-vector based speech pseudonymization across gender," 2021.

[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] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria*. ISCA, 2019, pp. 1526–1530.

[24] C. Veaux, J. Yamagishi, and K. Macdonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit," in *arXiv*, 2017.

[25] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*. ISCA, 2018, pp. 1086–1090.

[26] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*. ISCA, 2017, pp. 2616–2620.

[27] 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, Hawaii, US*. IEEE Signal Processing Society, Dec. 2011.

[28] 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, Calgary, AB, Canada*. IEEE, 2018, pp. 5329–5333.

[29] N. Brümmer and J. A. du Preez, "Application-independent evaluation of speaker detection," *Comput. Speech Lang.*, vol. 20, no. 2-3, pp. 230–275, 2006.

[30] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany*. ISCA, 2015, pp. 3214–3218.

[31] S. Yong and J. Nam, "Singing expression transfer from one voice to another for a given song," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 151–155.

[32] M. Akagi and T. Ienaga, "Speaker individuality in fundamental frequency contours and its control," *Journal of the Acoustical Society of Japan (E)*, vol. 18, no. 2, pp. 73–80, 1997.

[33] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 11 1992.

[34] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[35] U. E. Gaznepoglu and N. Peters, "Exploring the importance of f0 trajectories for speaker anonymization using x-vectors and neural waveform models," 2021. [Online]. Available: https://arxiv.org/abs/2110.06887