2nd Symposium on Security and Privacy in Speech Communication
23-24 September 2022, Incheon, Korea

Speaker Anonymization by Pitch Shifting Based on Time-Scale Modification

Candy Olivia Mawalim, Shogo Okada, Masashi Unoki

Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa 923–1292 Japan
{candylim, okada-s, unoki}@jaist.ac.jp

Abstract

The increasing usage of speech in digital technology raises a privacy issue because speech contains biometric information. Several methods of dealing with this issue have been proposed, including speaker anonymization or de-identification. Speaker anonymization aims to suppress personally identifiable information (PII) while keeping the other speech properties, including linguistic information. In this study, we utilize time-scale modification (TSM) speech signal processing for speaker anonymization. Speech signal processing approaches are significantly less complex than the state-of-the-art x-vector-based speaker anonymization method because it does not require a training process. We propose anonymization methods using two major categories of TSM, synchronous overlap-add (SOLA)-based algorithm and phase vocoder-based TSM (PV-TSM). For evaluating our proposed methods, we utilize the standard objective evaluation introduced in the VoicePrivacy challenge. The results show that our method based on the PV-TSM balances privacy and utility metrics better than baseline systems, especially when evaluating with an automatic speaker verification (ASV) system in anonymized enrollment and anonymized trials (a-a). Further, our method outperformed the x-vector-based speaker method, which has limitations in its complex training process, low privacy in an a-a scenario, and low voice distinctiveness.

Index Terms: speaker anonymization, voice privacy, pitch shifting, phase vocoder, time-scale modification

1. Introduction

Speech is the most common form of human communication. Consequently, digital communication technology utilizes speech to text as its input. However, distribution through public channels such as social media can lead to privacy issues because speech encapsulates linguistic-related and biometric-related content [1]. For instance, the advanced voice conversion system produces fake or cloned human voices of exceptional quality [2, 3].

Further, automatic speaker verification (ASV) systems developed in speech biometric studies have speech features that are related to personally identifiable information (PII) [4]. Without any protection, publicly available speech samples could be used for theft or fraud [5, 6]. Therefore, solutions for protecting the emerging threat are essentially important. One solution is using speaker anonymization (i.e., de-identification) to conceal PII in speech signals [7].

Several methods have been developed to de-identify PII in speech, such as the voice transformation method for anonymization purposes in [8, 9]. Subsequently, the Gaussian mixture model (GMM) mapping and harmonic-stochastic models were used to de-identify speech individuality in [10]. Further, the initiative to formally define speaker anonymization was realized in 2020 through the VoicePrivacy challenge (VPC) [11, 12]. In the VPC 2020, two baseline models were introduced for speaker anonymization [11]. The first baseline system is based on a neural source-filter (NSF) model and state-of-the-art x-vector speaker embedding [7]. The x-vector of a given speaker that represents speaker individuality information is anonymized to the x-vector of a generated pseudo-speaker. Meanwhile, the second baseline system uses the McAdams coefficient [13, 14]. Although the second baseline system does not perform as well as the first, the implementation is considerably simpler because it basically uses a signal processing technique and does not require a training process.

Instead, we investigate the time-scale modification (TSM) signal processing approach to speaker anonymization. We investigate pitch shifting using two major categories of TSM algorithms for speaker anonymization (i.e., synchronous overlap-add (SOLA) and phase vocoder-based TSM (PV-TSM)). Unlike a vocoder, this approach synthesizes speech via frame relocation and adaptation [15]. However, while TSM can synthesize higher-quality voices than a conventional vocoder, it cannot be used to analyze pitch and timbre independently [16]. For instance, the STRAIGHT vocoder [17] can be utilized to analyze the fundamental frequency (F_0), spectral envelope, and aperiodic signal. Prior studies related to speaker anonymization often used the SOLA-based TSM method for F_0 modification, e.g., [18, 19, 20]. F_0 modification using SOLA-based TSM is a relatively simple approach yet effective for manipulating the speaker's pitch. However, Patino et al. [14] reported that solely using SOLA-based TSM is inadequate for voice privacy protection in the state-of-the-art ASV system. Besides, SOLA-based TSM was reported less appropriate for signals with harmonic content [15]. Human voice contains harmonic structures; thus, applying PV-TSM that is more suited to a harmonic component could benefit speaker anonymization. Subsequently, the phase adaptation may manipulate not only F_0 but also the PII-related acoustics features. We follow the VPC protocols, which utilize the ASV system that was trained using the x-vector for comparing the performance of SOLA-based TSM and PV-TSM in the evaluation.

The rest of this paper is as follows. Section 2 discusses the speaker anonymization system based on the VPC. Section 3 introduces the TSM approach and our methods for using TSM algorithms for speaker anonymization. Section 4 describes our experiments, including the datasets, experimental setup, and results. Section 5 discusses the findings in this work. Finally, Section 6 concludes the paper and discusses our future work.

2. Speaker Anonymization

2.1. Definition

Speaker anonymization is defined in the Voice Privacy challenge (VPC) for voice privacy protection [11, 21]. It suppresses

10.21437/SPSC.2022-7

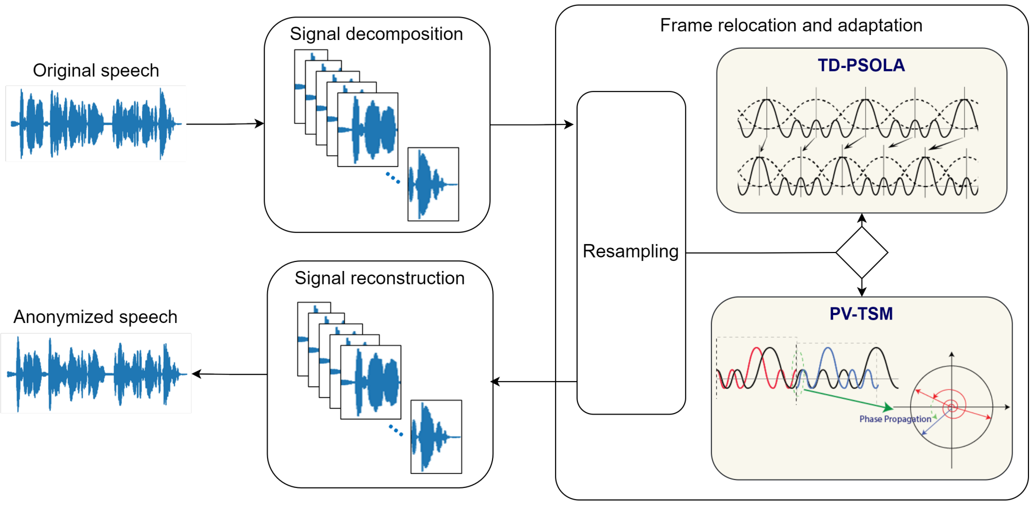


Figure 1: *Block diagram of proposed methods.*

the PII in speech signals while retaining information, such as linguistic content. There are four requirements of speaker anonymization based on the VPC 2020:

1. output should be a speech waveform,
2. speaker identity should be hidden,
3. output speech should be natural and intelligible, and
4. anonymized utterances of a given speaker should be different from those of other speakers.

2.2. Evaluation Metrics

Privacy and utility were the two metrics considered for objective evaluation in the VPC 2020 [11]. The privacy metrics were comprised of an equal error rate (EER) and log-likelihood-ratio cost function ($\mathrm{C_{llr}}$ and $\mathrm{C_{llr}^{min}}$, proposed in [22]). These metrics were evaluated by an ASV system (ASVeval) while the utility metric was evaluated by an automatic speaker recognition (ASR) system (ASReval). The ASReval computed the word error rate (WER) to identify the intelligibility of anonymized speech relative to the original speech. Both systems were trained on the LibriSpeech-train-clean-360 (a subset from LibriSpeech [23]) dataset using a Kaldi toolkit [24].

Furthermore, two secondary utility metrics were introduced in the VPC 2022 [21], including pitch correlation ρ^{F_0} and gain of voice distinctiveness $\mathrm{G_{VD}}$.A semi-informed attack model was also introduced, where the ASVeval and ASReval were trained using anonymized speech data.

2.3. Baseline Systems

Two baseline systems were introduced in the VPC 2020 [11]. The primary baseline system (B1a) was chiefly developed based on an NSF model and x-vector speaker embedding [7]. The main idea was to manipulate speaker individuality information by isolating the input speech into a fundamental frequency (F_0), speaker individuality feature (x-vector), and bottleneck feature (linguistic-related information). Subsequently, the secondary baseline system (B2a) was a speaker anonymization system using the McAdams coefficient (α) [14]. The speech analysis and synthesis process was based on linear prediction analysis.

Additionally, modifications of two baseline systems in the VPC 2020 were reported as other systems in the VPC 2022 (namely, B1b and B2b). While similar to the B1a, the B1b utilized a unified HiFi-GAN NSF model as the speech synthesizer. Unlike the B2a, the B2b utilized a randomized value of the McAdams coefficient (α). The α was randomly selected from a uniform distribution: $\alpha \sim U(0.5, 0.9)$.

3. Proposed Method

This study aims to exploit pitch-shifting methods based on TSM for speaker anonymization. We utilize time-domain pitch synchronous overlap-add (TD-PSOLA) [25] and phase vocoder-based TSM (PV-TSM) approaches [26, 15]. Figure 1 shows the block diagram of our proposed method using these two TSM approaches. The remaining parts of this section overview the TSM and detail our methods.

3.1. Time-Scale Modification (TSM)

TSM algorithms are used in signal processing to compress or stretch audio signals [15]. TSM is often applied for various purposes, especially in music processing. For instance, it is used when adjusting a video clip to the audio stream for faster or slower playback. Apart from music processing, TSM algorithms are also often used as speech synthesizers [16, 27].

The general TSM procedures are comprised of three main components: signal decomposition, frame relocation & adaptation, and signal reconstruction [15]. The performance of TSM algorithms depends upon the procedure in each component. In this study, we manipulate PII using the major conventional TSM procedure based on the synchronous overlap-add (SOLA) and that based on the phase vocoder (PV-TSM).

PII such as speaker individuality is strongly associated with pitch trajectory [28], so we utilize a TD-PSOLA algorithm [25] to represent SOLA-based TSM for anonymization. Meanwhile, we also utilize the PV-TSM algorithm [29] to modify pitch trajectory. The PV-TSM improves the phase of the synthesized speech by phase propagation [15].

3.2. Speaker anonymization by F_0 modification

Several studies have been conducted on speaker anonymization by modifying F_0 [12, 32, 33, 34, 20]. However, most studies utilized the B1a framework for speech synthesis, suggesting that F_0 modification could improve the privacy (in EER)

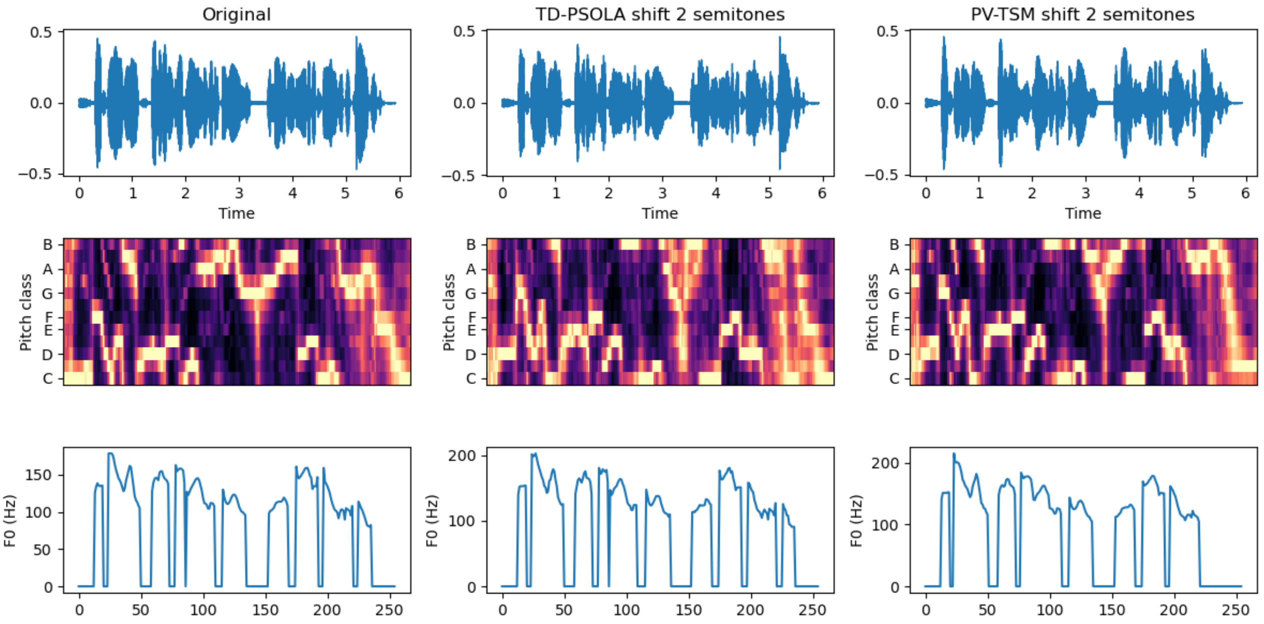


Figure 2: *Pitch shifting at two semitones using TD-PSOLA and PV-TSM algorithms. First row shows waveforms. Second row shows chromagram of speech signals extracted by Librosa [30]. Third row shows F_0 trajectory in Hz extracted by pYIN [31] algorithm in Librosa [30]. Example of pitch-shifting by two semitones: original pitch in C tone becomes D tone after shifting.*

but introducing more WER (reducing the utility). Recently, another study on F_0 manipulation without using the B1a framework is conducted by Tavi et al. [20]. Their study utilized functional data analysis (FDA) on F_0 trajectories to improve speaker anonymization. Although the output anonymization was not comparable to that of the x-vector-based synthesizer, they determined that pitch manipulation by FDA improves privacy (in EER).

3.2.1. Anonymization using TD-PSOLA

The TD-PSOLA algorithm is typically used to manipulate the pitch of a given speaker. Although this method is reportedly insufficient for protecting privacy against x-vector-based ASV systems [20], this algorithm's ability to preserve privacy based on VPC recipes has yet to be thoroughly investigated. We anonymize the given speech signals using the pitch shifting by the TD-PSOLA algorithm.

The main procedures of pitch shifting by the TD-PSOLA algorithm are as follows:

- First, we input speech signal x with an analysis frame x_m. By this iterative process, we obtained the output signal y.
- Second, the original speech frame x_m is resampled to increase or decrease the pitch.
- Next, a Hann window function w is applied to the analysis frame x_m to obtain the synthesis frame y_m.
- Lastly, OLA is conducted to adjust the output signal y duration modified by resampling to the original signal x duration.

3.2.2. Anonymization using PV-TSM

The short-time Fourier transform (STFT) is a fundamental technique for frequency analysis. However, the resulting frequency may be inaccurate depending on the discretization parameters [15]. Therefore, a phase vocoder (different from a general vocoder) is used to improve the frequency estimation by using the derivation of the sinusoidal components' instantaneous frequency (i.e., phase propagation process).

Similar to the speaker anonymization using TD-PSOLA, we can manipulate the PII in a given speech signal with pitch shifting. The pitch shifting is conducted as follows:

- First, the original speech frame x_m is resampled to increase or decrease the pitch.
- Next, STFT is performed to obtain the frequency spectra of the input signal X.
- Then, the phase jumps in each overlapping frame is fixed via phase propagation.
- Finally, the signal is reconstructed from the frequency spectra after phases update X^{Mod}.

4. Experiments

Our experiments follow the protocols and datasets provided in the VPC 2020[1] [11] and VPC 2022[2] [21].

4.1. Datasets

We evaluated our methods using the datasets described in the VPC [11, 21]. The evaluation datasets were comprised of LibriSpeech (Libri) [23] and the voice cloning toolkit (VCTK) [35]. Each dataset was divided into two subsets, i.e., development and test. Further, the VCTK dataset is also split into "common" (comm) and "different" (diff) parts to evaluate the speaker verifiability regarding the linguistic contents (common or different utterances).

[1] https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2020

[2] https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2022

Table 1: *ASVeval of VPC 2020 results for o-a scenario.*

Dataset		**Gender**	**Weight**	**EER (%)**						
				Orig	**B1a**	**B1b**	**B2b**	**TD-PSOLA (2,3)**	**PV-TSM (3,5)**	**PV-TSM (2,3)**
Dev	Libri	female	0.25	8.81	50	52.98	37.93	8.38	42.19	34.67
		male	0.25	1.24	52.78	55.43	38.35	0.93	44.1	22.72
	VCTK (diff)	female	0.20	2.92	49.18	52.78	35.77	4.55	41.89	32.56
		male	0.20	1.44	53.85	54.44	42.33	1.79	49.08	26
	VCTK (comm)	female	0.05	2.62	50	45.93	36.34	3.49	46.22	41.33
		male	0.05	1.43	55.37	55.27	45.01	1.71	42.45	32.2
Weighted average dev				**3.59**	**51.57**	**53.61**	**38.76**	**3.86**	**44.20**	**29.74**
Test	Libri	female	0.25	7.66	50	50.91	31.39	7.3	37.77	34.67
		male	0.25	1.11	52.78	51.89	27.39	1.34	38.53	22.72
	VCTK (diff)	female	0.20	4.94	49.18	49.59	36.32	4.58	40.74	32.56
		male	0.20	2.07	53.85	54.99	38.12	2.41	45.35	26
	VCTK (comm)	female	0.05	2.89	50	50.58	44.51	3.76	51.16	41.33
		male	0.05	1.13	55.37	53.95	40.68	1.41	44.07	32.2
Weighted average test				**3.80**	**51.57**	**51.84**	**33.84**	**3.82**	**41.05**	**29.74**

Table 2: *ASVeval of VPC 2020 results for a-a scenario.*

Dataset		**Gender**	**Weight**	**EER (%)**						
				Orig	**B1a**	**B1b**	**B2b**	**TD-PSOLA (2,3)**	**PV-TSM (3,5)**	**PV-TSM (2,3)**
Dev	Libri	female	0.25	8.81	35.51	32.10	40.62	11.51	50.99	47.44
		male	0.25	1.24	32.45	32.76	43.63	1.40	39.44	38.20
	VCTK (diff)	female	0.20	2.92	27.79	21.56	35.93	3.59	53.06	51.94
		male	0.20	1.44	29.63	24.62	43.37	1.69	31.02	30.27
	VCTK (comm)	female	0.05	2.62	27.62	16.28	54.36	2.62	51.16	49.42
		male	0.05	1.43	30.20	25.07	46.44	1.71	43.30	42.45
Weighted average dev				**3.59**	**31.37**	**27.52**	**41.96**	**4.50**	**44.15**	**42.45**
Test	Libri	female	0.25	7.66	33.39	27.74	42.70	7.48	45.80	44.16
		male	0.25	1.11	33.63	35.86	47.66	1.11	41.43	39.64
	VCTK (diff)	female	0.20	4.94	33.80	23.15	31.02	4.99	47.48	46.97
		male	0.20	2.07	28.07	25.20	38.92	2.58	56.20	56.49
	VCTK (comm)	female	0.05	2.89	32.08	21.97	38.15	3.47	34.97	34.10
		male	0.05	1.13	26.84	25.71	46.61	1.13	49.44	51.98
Weighted average test				**3.80**	**32.08**	**27.95**	**40.82**	**3.89**	**46.76**	**45.95**

4.2. Experimental Setup

For the implementation of our proposed methods, we utilized the TSM algorithms based on the TSM toolbox [29]. We used the default analysis frame size, setting the analysis hop size to 80. The pitch-shifting parameter Δ was uniformly randomized, as follows:

$$\Delta \sim U\{(-\Delta_{\max}, -\Delta_{\min}), (\Delta_{\min}, \Delta_{\max})\}. \qquad (1)$$

where $(\Delta_{\min}, \Delta_{\max})$ are (2, 3) or (3, 5) on the semitone scale (one semitone is equivalent to 100 cents (¢)). A semitone (half step) is a commonly used musical interval, representing different pitches between neighboring notes on the piano. Accordingly, one octave consists of 12 semitones (C, C#, D, D#, E, F, F#, G, G#, A, A#, and B). Pitch shifting by n semitones is expressed as follows:

$$F_{0_y}(t) = 2^{n/12} \times F_{0_x}(t) \qquad (2)$$

where $F_{0_x}(t)$ is the fundamental frequency trajectory of the original signal x in Hz, and $F_{0_y}(t)$ is the fundamental frequency trajectory after shifting.

Figure 2 shows an example of output anonymized speech with the corresponding original speech after pitch shifting at two semitones. We evaluated our methods using the ASVeval and ASReval from the VPC 2020 [11]. Additionally, we considered the VPC 2022 objective evaluation metrics [21] of the best algorithm in this study based on the VPC 2020 evaluation results. We include the evaluation of the privacy performance of the semi-informed attack model (ASVeval in the VPC 2022), pitch correlation (ρ^{F_0}), and the gain of voice distinctiveness (G_{VD}) [36]. Furthermore, we compare the results using the baseline systems (B1a, B1b, and B2b) in the VPC 2022.

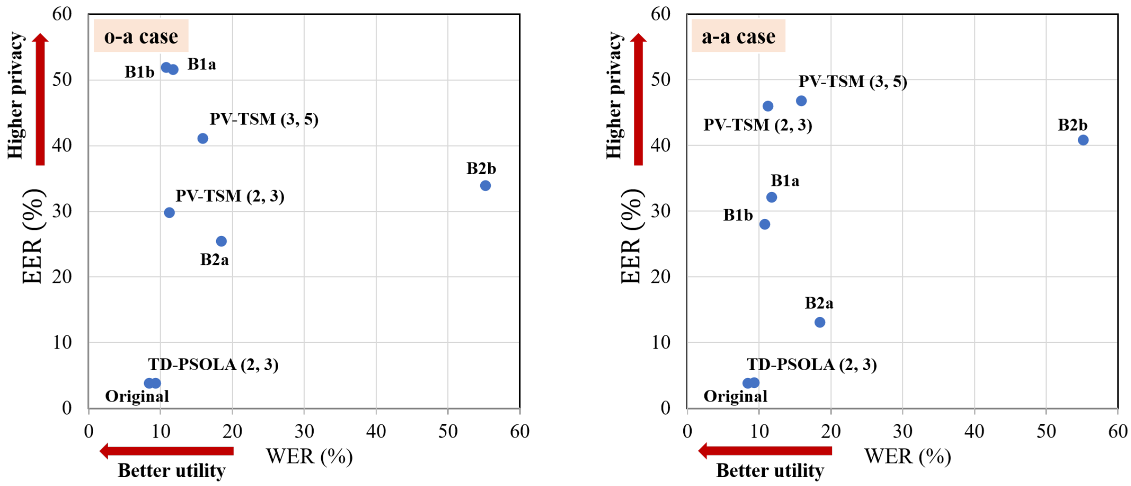
4.3. Results

Carrying out the whole evaluation based on the VPC 2022 recipes requires more resources than the VPC 2020 recipes (huge time and space complexity). Thus, first, we evaluated our method using the VPC 2020 recipes without considering the semi-informed attack model. After verifying the best one, we conducted the VPC 2022 using the best algorithm.

As mentioned in Subsection 2.2, two objective evaluations in the VPC 2020 are used to measure the privacy and utility metrics. We utilized the ASVeval 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). The Orig. column in both Tables 1 and 2 shows the results of the ASVeval in the o-o scenario. The other columns in Table 1 show the results of the ASVeval in the o-a scenario with corresponding methods. In the same manner, Table 2 shows the ASVeval results in the a-a scenario. Table 3 shows the utility evaluation results using the ASReval. Figure 3 compares the privacy metrics versus the

Table 3: *ASReval of VPC 2020 results.*

Dataset		WER (%)						
		Orig	**B1a**	**B1b**	**B2b**	**TD-PSOLA (2,3)**	**PV-TSM (3,5)**	**PV-TSM (2,3)**
Dev	Libri	3.83	6.96	5.91	36.42	4.27	9.26	5.41
	VCTK	10.79	15.96	15.48	52.09	12.14	22.94	15.16
Average-dev		**7.31**	**11.46**	**10.70**	**44.26**	**8.21**	**16.10**	**10.29**
Test	Libri	4.14	7.78	6.08	48.12	4.49	8.18	5.61
	VCTK	12.81	15.74	15.60	62.35	14.24	23.64	16.98
Average-test		**8.48**	**11.76**	**10.84**	**55.24**	**9.37**	**15.91**	**11.30**



(a) *original enrollment – anonymized trials (o-a)*

(b) *anonymized enrollment – anonymized trials (a-a)*

Figure 3: *Privacy vs utility plot based on objective evaluation in VPC 2020 (using test set).*

utility metrics.

We also conducted an objective evaluation based on the metrics in the VPC 2022 [21]. Table 4 shows the ASVeval results of the semi-informed attack model for the test set. Meanwhile, Figure 4 shows the utility metrics in terms of pitch correlation and the gain of voice distinctiveness.

5. Discussion

By comparing the results of all the methods in Table 1, we can conclude that the primary baseline systems (B1a and B1b) have the best privacy in the o-a scenario. If we compare the methods based on signal processing (which does not require a complex training process), the PV-TSM algorithm with wider and higher shift parameters gave better protection than the one from B2b. Interestingly, although TD-PSOLA is also one of the TSM algorithms, it provides totally different results from the one from the PV-TSM. These results confirmed our hypothesis that the PV-TSM method is more beneficial for manipulating voice that contains harmonic structures for voice privacy protection. Additionally, the results aligned with prior work that the TD-PSOLA algorithm is insufficient (in either scenario) for protecting voice privacy against x-vector-based ASV systems [20]. Recently, a speaker anonymization method based on waveform similarity overlap-add (WSOLA) was also proposed in [18, 19]. Although the WSOLA-based method was comparable to the B1a in o–a scenario, its performance in a–a scenario and ASReval were worse than those of baselines.

The VPC 2020 results indicate that all existing methods that achieved high privacy in the o-a scenario attained much lower privacy in the a-a scenario. For instance, the EER of the B1a was reduced from approximately 50% (o-a scenario) to 30% (a-a scenario). This result indicates that privacy is not well-preserved when an attacker has access to several anonymized utterances (anonymization algorithm as a black box) [32]. However, the B2b and our methods that are based on the PV-TSM could provide higher privacy protection in the a-a scenario than in the o-a scenario. Table 2 shows that our PV-TSM (3, 5) method achieved the best speaker verifiability.

Subsequently, the utility metric in the VPC 2020 was evaluated by an ASReval system. Table 3 shows that the speech intelligibility significantly degraded (WER significantly increased) even though the B2b could provide better privacy than the B2a (results in [11]). Despite the simple anonymization method, our methods based on the PV-TSM can provide a more reliable balance between the trade-off in privacy and utility metrics than the anonymization method using the McAdams coefficient (B2a and B2b).

Figure 3 shows the average results overall (with the B2a as an additional reference). These results indicate that the B1a and the B1b perform almost similarly and can be considered the best in the o-a scenario (see left figure). Our methods based on the PV-TSM algorithm could achieve the next highest balance between privacy and utility. The privacy could be improved by increasing the pitch-shifting parameter with a wider range. However, this process slightly degraded speech intelligibility (WER increased from 11.30% to 15.91%). Even so, the PV-TSM-based methods could significantly surpass the methods using the McAdams coefficient (B2a, B2b) that significantly degraded speech intelligibility (WER increased from 18.53%

Table 4: *ASVeval of VPC 2022 results using test set.*

Dataset	Gender	EER (%)				
		Orig	**B1a**	**B1b**	**B2b**	**PV-TSM (3,5)**
Libri	female	7.66	12.04	9.49	7.12	19.34
	male	1.11	8.91	7.80	1.11	6.46
VCTK (diff)	female	4.94	16.00	10.91	16.92	9.77
	male	2.07	10.05	7.52	7.69	4.99
VCTK (comm)	female	2.89	17.34	15.32	10.98	6.65
	male	1.13	9.89	8.19	4.80	1.41
Weighted average test		**3.80**	**11.81**	**9.18**	**7.77**	**9.81**

to 55.24%).

The right side of Figure 3 shows the average results overall using the test set in the a-a scenario. The figure indicates that the methods based on the PV-TSM algorithm could perform the best. Although the range and the value of the pitch-shifting parameters differ, the privacy of the PV-TSM (2, 3) and the PV-TSM (3, 5) in the a-a scenario are comparable. The primary baseline (B1a and B1b) could not achieve as high a privacy performance in the o-a scenario. Similarly, the privacy of the second baseline could be increased by changing the McAdams coefficient from a fixed value to a randomly selected value from a particular range. However, it greatly degraded speech intelligibility.

We also conducted some evaluations based on the VPC 2022. First, we conducted the ASVeval while considering the semi-informed attack model (a-a scenario). In this attack model, the attacker has access to the anonymization method for the utterance level (without knowing the detailed parameters that represent the speaker identity). Table 4 shows the results of this evaluation. Although not as good as the B1a, the PV-TSM (3, 5) could perform slightly better than the B1b. Note that, in the PV-TSM (3, 5), the ASVeval was trained with the anonymization algorithm with prior information on the shifting parameter range (3, 5). If the attacker has no information about the shifting parameter, the performance will be better.

We acknowledge that no significant interpretation can be obtained from the results of the new ASReval in the VPC 2022 (by retraining it with anonymized speech). For instance, the average-dev results using the VPC 2020 ASReval for the B2b indicate that the speech intelligibility is highly distorted in comparison with the B1a (the WER of the B1a and the B2b are approximately 11.46% and 44.26%, respectively). However, the VPC 2022 ASReval outputs nearly similar results for the B1a and the B2b (the WER of the B1a and the B2b are approximately 7.94% and 8.04%, respectively [21]). Therefore, we limited discussion on the new ASReval results in this study.

Alternatively, we calculated the pitch correlation (ρ^{F_0}) and the gain of voice distinctiveness ($\mathrm{G_{VD}}$) that described in [21] as secondary utility metrics (see Fig. 4). The weighted average ρ^{F_0} for test set using B1a, B1b, B2b, and PV-TSM (3,5) are 0.77, 0.80, 0.62, and 0.81, respectively. These results indicate that the PV-TSM (3, 5) successfully preserves the intonation to some extent (as described in [21]). Moreover, the results of $\mathrm{G_{VD}}$ show that our method preserves the voice distinctiveness better than the baseline systems (B1a, B1b, and B2b).

We publicly demonstrate anonymized speech output from our methods compared with baseline systems (B1a, B2a, and B2b)[3]. The stimuli in this demonstration is randomly selected from the LibriSpeech test dataset (one female utterance and one male utterance).

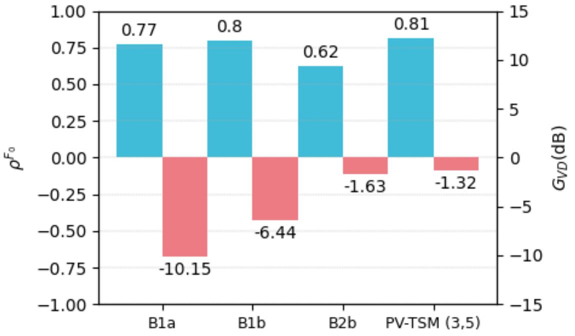


Figure 4: *Objective evaluation in VPC 2022 (weighted average test) using the secondary utility metrics: pitch correlation (ρ^{F_0}) (blue bars) and gain of voice distinctiveness ($\mathrm{G_{VD}}$) (pink bars).*

6. Conclusions

We investigated conventional TD-PSOLA and PV-TSM approaches to speaker anonymization. We determined the possible voice privacy protection against x-vector-based ASV systems via extensive objective evaluations per the VPC 2020 and the VPC 2022. We found that the TD-PSOLA algorithm can be used for pitch shifting but is insufficient for privacy protection in the ASV system. However, pitch shifting by the PV-TSM algorithm for speaker anonymization performs significantly better than one using the McAdams coefficient, providing the highest balance of both privacy and utility metrics in the a-a scenario in comparison with the baseline systems. Additionally, our method using PV-TSM can preserve secondary utility metrics (i.e., pitch correlation can represent intonation and the gain of voice distinctiveness to some extent). In future work, we will more thoroughly investigate appropriate shift parameters. We will also explore non-linear pitch shifting using the PV-TSM algorithm.

7. Acknowledgments

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), and the SCAT Research Foundation. This work was also partially supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (Grant Numbers 22H04860 and 22H00536) and JST AIP Trilateral AI Research, Japan (Grant Number JPMJCR20G6).

[3] `http://www.jaist.ac.jp/~candylim/VP2022/demo_web/demo.html`

8. 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] A. Irum and A. Salman, "Speaker verification using deep neural networks: A review," *International Journal of Machine Learning and Computing*, vol. 9, no. 1, 2019.

[5] R. K. Das and S. M. Prasanna, "Speaker verification from short utterance perspective: a review," *IETE Technical Review*, vol. 35, no. 6, pp. 599–617, 2018.

[6] V. Vestman, T. Kinnunen, R. G. Hautamäki, and M. Sahidullah, "Voice mimicry attacks assisted by automatic speaker verification," *Computer Speech & Language*, vol. 59, pp. 36–54, 2020.

[7] 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," 09 2019, pp. 155–160.

[8] 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.

[9] 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.

[10] 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.

[11] 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

[12] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. 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," 09 2021.

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

[16] 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.

[17] H. Kawahara, "Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical Science and Technology*, vol. 27, no. 6, pp. 349–353, 2006.

[18] H. Kai, S. Takamichi, S. Shiota, and H. Kiya, "Lightweight voice anonymization based on data-driven optimization of cascaded voice modification modules," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 560–566.

[19] ——, "Robustness of signal processing-based pseudonymization method against decryption attack," in *The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 287–293.

[20] L. Tavi, T. Kinnunen, and R. González Hautamäki, "Improving speaker de-identification with functional data analysis of f0 trajectories," *Speech Communication*, vol. 140, pp. 1–10, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639322000498

[21] 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

[22] 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.

[23] 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.

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

[25] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5, pp. 453–467, 1990, neuropeech '89. [Online]. Available: https://www.sciencedirect.com/science/article/pii/016763939090021Z

[26] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323–332, 1999.

[27] 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.

[28] 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.

[29] J. Driedger and M. Müller, "TSM toolbox: MATLAB implementations of time-scale modification algorithms," in *Proceedings of the 17th International Conference on Digital Audio Effects, DAFx-14, Erlangen, Germany, September 1-5, 2014*, S. Disch, J. Herre, R. Rabenstein, B. Edler, M. Müller, and S. Turowski, Eds., 2014, pp. 249–256.

[30] B. McFee, C. Raffel, D. Liang, D. Ellis, M. Mcvicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," 01 2015, pp. 18–24.

[31] M. Mauch and S. Dixon, "Pyin: A fundamental frequency estimator using probabilistic threshold distributions," 05 2014, pp. 659–663.

[32] 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

[33] C. O. Mawalim, K. Galajit, J. Karnjana, S. Kidani, and M. Unoki, "Speaker anonymization by modifying fundamental frequency and x-vector singular value," *Comput. Speech Lang.*, vol. 73, no. C, may 2022. [Online]. Available: https://doi.org/10.1016/j.csl.2021.101326

[34] P. Champion, D. Jouvet, and A. Larcher, “A study of f0 modification for x-vector based speech pseudonymization across gender,” *ArXiv*, vol. abs/2101.08478, 2021.

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

[36] P.-G. Noé, J.-F. Bonastre, D. Matrouf, N. Tomashenko, A. Nautsch, and N. Evans, “Speech pseudonymisation assessment using voice similarity matrices,” 2020. [Online]. Available: https://arxiv.org/abs/2008.13144