IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 10, 2020, accepted October 26, 2020, date of publication November 3, 2020, date of current version November 12, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3035416

Speaker Anonymization for Personal Information Protection Using Voice Conversion Techniques

IN-CHUL YOO, (Member, IEEE), KEONNYEONG LEE, SEONGGYUN LEEM, HYUNWOO OH, BONGGU KO, AND DONGSUK YOOK, (Member, IEEE)
Artificial Intelligence Laboratory, Department of Computer Science and Engineering, Korea University, Seoul 02841, South Korea

Corresponding author: Dongsuk Yook (yook@korea.ac.kr)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning under Grant NRF-2017R1E1A1A01078157, in part by the Ministry of Science and ICT (MSIT) through the Information Technology Research Center (ITRC) Support Program supervised by the Institute for Information & Communications Technology Promotion (IITP) under Grant IITP-2018-0-01405, in part by the IITP grant funded by the Korean Government (MSIP) (A research on safe and convenient big data processing methods) under Grant 2018-0-00269, and in part by the Korea University Grant.

ABSTRACT As speech-based user interfaces integrated in the devices such as AI speakers become ubiquitous, a large amount of user voice data is being collected to enhance the accuracy of speech recognition systems. Since such voice data contain personal information that can endanger the privacy of users, the issue of privacy protection in the speech data has garnered increasing attention after the introduction of the General Data Protection Regulation in the EU, which implies that restrictions and safety measures for the use of speech data become essential. This study aims to filter the speaker-related voice biometrics present in speech data such as voice fingerprint without altering the linguistic content to preserve the usefulness of the data while protecting the privacy of users. To achieve this, we propose an algorithm that produces anonymized speeches by adopting many-to-many voice conversion techniques based on variational autoencoders (VAEs) and modifying the speaker identity vectors of the VAE input to anonymize the speech data. We validated the effectiveness of the proposed method by measuring the speaker-related information and the original linguistic information retained in the resultant speech, using an open source speaker recognizer and a deep neural network-based automatic speech recognizer, respectively. Using the proposed method, the speaker identification accuracy of the speech data was reduced to 0.1–9.2%, indicating successful anonymization, while the speech recognition accuracy was maintained as 78.2–81.3%.

INDEX TERMS Data privacy, deep neural networks, speaker anonymization, variational autoencoder, voice conversion.

I. INTRODUCTION

Speech-based user interfaces are commonly utilized in various applications, owing to their feasibility and simplicity [1]. Speech recognition algorithms enable us to control various devices using natural languages. Deep learning-based algorithms, which are prominently used for speech recognition, require large quantities of training data [2], [3]. The collection of voice data from users is an attractive task because they contain various types of real-life speeches that can significantly enhance the accuracy of speech recognition systems. However, since such data contain personal information, it is not advisable to use them directly. Especially, since an increasing number of applications use speech data in their user authentication mechanisms, openly available speech data may be vulnerable to security threats. In other words, speaker identification algorithms can be used to determine the identity of a speaker through speech data. Therefore, if such data are stored without a proper anonymization process, they can be vulnerable to various exploits, which must be avoided. Furthermore, there is an increasing demand for the protection of speech data-related personal information, usually referred to as privacy protection. This can be partly attributed to the introduction of the General Data Protection Regulation (GDPR) in the EU [4], which implies that restrictions and preventive measures on the use of speech data become essential [5].

Several studies have attempted to tackle privacy protection in speech processing systems by extracting privacy-preserving features from speeches [6], [7], extracting

The associate editor coordinating the review of this manuscript and approving it for publication was Kuo-Hui Yeh.

features from encrypted signals [8], augmenting models with adversarial representations [9], and applying score normalizations [10]. However, such feature- or model-level privacy protection techniques have a critical drawback, wherein the users cannot verify that their personal information is actually removed from the resultant features or models. This drawback of the privacy protection methods can make users reluctant for storing their data to enhance future models. Moreover, determining the level of privacy protection systematically can lead to problems, since the experience of privacy differs in various situations as demonstrated in [11]. Therefore, in this study, we propose the use of a raw data-level privacy protection technique, where an algorithm outputs the conventional waveform audio files that can be easily accessed by the users. If users are allowed to access the resultant speech data, they can intuitively decide whether the level of privacy protection is ''good enough'' for them to provide their speech data for future use. This process can help in complying with the GDPR, which requires explicit consent to use speech data.

Speaker anonymization aims to retain the linguistic content of speech data while removing the voice biometrics of the speakers, thereby considerably reducing the potential risks concerning the exploitation of voice-related personal information. The key challenge in speaker anonymization is the preservation of the linguistic content of speech data to ensure that various speech recognition systems can correctly recognize them. Such anonymized speeches can be collected to train the speech recognition systems to enhance their accuracy.

In this study, we propose the adoption of voice conversion techniques for speaker anonymization. Voice conversion aims to convert the identity of one speaker to that of another speaker while preserving the linguistic content of speech data. Speaker anonymization can be achieved by utilizing voice conversion techniques to modify the identity of a speaker to that of an anonymous speaker. In Section II, related works on speaker anonymization and voice conversion are reviewed. Section III describes the proposed method to create the identities of anonymous speakers by altering one-hot speaker identity vectors. Section IV summarizes various results of the speech recognition and speaker identification experiments using open source libraries and publicly available speech data. Section V concludes the paper with some future research directions.

II. RELATED WORKS

A. VOICE MODULATION

Voice modulation alters various features of voice data, such as pitch and intonation, to create different speech styles. Its goal is similar to that of speaker anonymization as it attempts to remove the identity of the speaker from a given speech. Examples of voice modulation can be frequently found in TV news to ensure the anonymity of suspects and witnesses.

Voice modulation can be implemented by various methods. A simple method of implementing voice modulation involves the application of acoustic filters to alter the spectral characteristics of a given speech. Such methods, however, are not highly recommended because the original speech can be easily recovered using inverse filters. Vocoders can also be used to implement voice modulations by altering the features of input speech data during synthesis [12], [13]. For example, the WORLD vocoder [14] uses the values of fundamental frequencies (F0s) and aperiodicities (APs) with spectrograms. By using the fundamental frequencies of other speakers, the resultant speech can have different speaker characteristics than those of the input speech.

The main difference between these voice modulation techniques and speaker anonymization is that such techniques are not explicitly designed to retain the comprehensibility of the modulated speech. For example, in many cases of voice modulation in TV news, the modulated speech is typically accompanied with closed captions because the comprehensibility of the modulated speech is reduced. Furthermore, the values of speech data altered by voice modulation are limited for training speech recognition systems as the speech may not sound normal voice.

B. VOICE CONVERSION

Voice conversion aims to convert the identity of the speaker of an input speech to that of the target speaker while retaining the linguistic content of the input speech. The simplest form of voice conversion requires parallel data for training and it is capable of one-to-one speaker conversion. Parallel data include same transcription utterances spoken by the source and target speakers and they are highly expensive to collect. Thus, several studies attempted to use non-parallel data to train voice conversion models. In the case of multiple-speaker voice conversion, one-to-one speaker conversion algorithms may be applied to obtain separately trained models for all possible combinations of speaker pairs. However, this approach becomes impractical as the number of speakers increases.

Variational autoencoders (VAEs) can be used for many-to-many voice conversion using a single model and non-parallel training data [15]. VAEs can be combined with generative adversarial networks (GANs) [16] to enhance the quality of the converted speech, where the decoder of the VAE is shared with the generator of the GAN [17]. The VAE-GAN can be extended to include the cycle-consistency loss [18], [19] to further improve the voice quality, especially for non-parallel training data. This is known as a cycle-consistent variational autoencoding generative adversarial network (CycleVAE-GAN) [19]. Fig. 1 shows each component of the CycleVAE-GAN.

The loss function of the VAE is given by

$$\mathcal{L}_{\mathrm{VAE}}(\phi,\theta;x,X) = \mathbb{D}_{\mathrm{KL}}\left(q_{\phi}(z|x) \parallel p(z)\right) - \mathbb{E}_{z\sim q_{\phi}(z|x)}\left[\log p_{\theta}(x|z,I_X)\right], \quad (1)$$

where x, X, I_X, ϕ, and θ denote the input speech, speaker of the input speech, speaker identity vector, encoder parameters, and decoder parameters, respectively. $\mathbb{D}_{\mathrm{KL}}$ and $\mathbb{E}$ represent the Kullback-Leibler divergence and expectation,

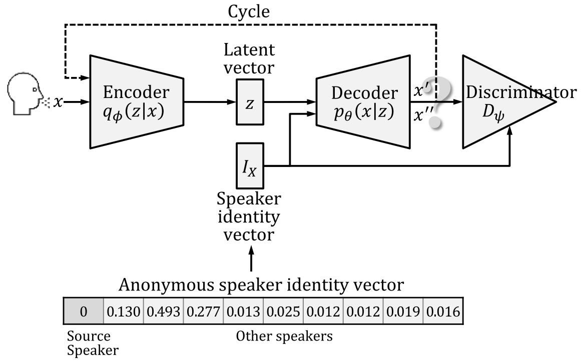


FIGURE 1. Speaker anonymization using the CycleVAE-GAN. *x* is the input speech, *x′* is either the reconstructed speech (when I_X is the source speaker identity vector) or the converted speech (when I_X is replaced with the target speaker identity vector), *x″* is the converted back speech which should recover the original input speech *x*. The dashed line represents the cyclic conversion path that produces *x″*. When the speaker identity vector I_X is replaced with an anonymous speaker identity vector (see Section III), the input speech can be anonymized.

respectively. By minimizing (1), the encoder is trained to extract the latent vector z which corresponds to the linguistic information of the input speech, and the decoder is trained to reconstruct the input speech from the latent vector z and the source speaker identity vector I_X. To convert the speech from a source speaker to a target speaker, the source speaker identity vector I_X is replaced with the target speaker identity vector. That is, the target speaker identity vector, which is a one-hot vector containing 1 for the target speaker and 0s for other speakers (see Fig. 2 (a) for example), is fed into the decoding process of the VAE to convert the source speaker speech to the target speaker speech.

The cycle-consistency loss is computed as follows:

$$\begin{aligned}\mathcal{L}_{\text{Cycle}}(\phi,\theta;x,X,Y) = {} & \mathbb{D}_{\text{KL}}\left(q_\phi\left(z|x'_{X\to Y}\right) \parallel p(z)\right) \\ & - \mathbb{E}_{z\sim q_\phi\left(z|x'_{X\to Y}\right)}\left[\log p_\theta\left(x|z,I_X\right)\right],\end{aligned} \tag{2}$$

where $x'_{X\to Y}$ denotes the speech converted from the source speaker X to the target speaker Y. The input speech x from speaker X goes through the encoder and the decoder with speaker identity vector I_Y to generate the converted speech $x'_{X\to Y}$ which has the same linguistic content as x but in speaker Y's voice. Then, the converted speech goes through the encoder and the decoder with speaker identity vector I_X to generate the converted back speech $x''_{X\to Y\to X}$ which should recover the original input speech x. This cyclic conversion encourages the explicit training of the conversion paths without parallel data.

Now, given the input speeches x and y from speakers X and Y, respectively, the loss function of the CycleVAE is defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{CycleVAE}}(\phi,\theta;x,y,X,Y) = {} & \mathcal{L}_{\text{VAE}}(\phi,\theta;x,X) \\ & + \mathcal{L}_{\text{VAE}}(\phi,\theta;y,Y) \\ & + \lambda_1\mathcal{L}_{\text{Cycle}}(\phi,\theta;x,X,Y) \\ & + \lambda_1\mathcal{L}_{\text{Cycle}}(\phi,\theta;y,Y,X),\end{aligned} \tag{3}$$

where λ_1 decides the weight of the cycle-consistency loss in the CycleVAE.

Finally, after optimizing the VAE module, the GAN module is used to train the CycleVAE-GAN model. The decoder of the VAE is considered as the generator of the GAN. The discriminator of the GAN helps the generator to produce a speech similar to that of the target speaker. Given the input speeches x and y from speakers X and Y, respectively, the loss function of the CycleVAE-GAN is defined as follows:

$$\begin{aligned}&\mathcal{L}_{\text{CycleVAE-GAN}}(\phi,\theta,\psi;x,y,X,Y) \\ &= \mathcal{L}_{\text{CycleVAE}}(\phi,\theta;x,y,X,Y) \\ &\quad + \lambda_2\mathbb{E}_{y|Y}\left[D_\psi(y)\right] - \lambda_2\mathbb{E}_{z\sim q_\phi(z|x)}\left[D_\psi\left(G_\theta\left(z,I_Y\right)\right)\right] \\ &\quad + \lambda_2\mathbb{E}_{x|X}\left[D_\psi(x)\right] - \lambda_2\mathbb{E}_{z\sim q_\phi(z|y)}\left[D_\psi\left(G_\theta\left(z,I_X\right)\right)\right],\end{aligned} \tag{4}$$

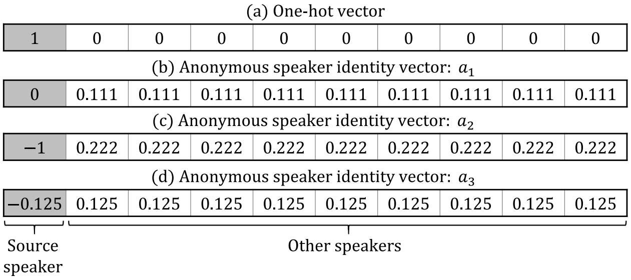
where G_θ and D_ψ denote the generator with parameter θ and the discriminator with parameter ψ, respectively, and λ_2 decides the weight of the GAN loss in the CycleVAE-GAN. In this work, the Wasserstein GAN is used instead of the vanilla GAN [17], [20]. Equation (4) is minimized for the VAE and the generator, and it is maximized for the discriminator.

III. SPEAKER ANONYMIZATION USING VOICE CONVERSION TECHNIQUES

A. ANONYMOUS SPEAKER IDENTITY VECTORS USING UNIFORM VALUES

As explained in Section II-B, the target speaker identity vector guides the decoder of the VAE that the output speech has similar characteristics as that of the target speaker. If appropriate values are selected for the speaker identity vector, we believe that it can be used to force the decoder to output speech that has novel characteristics as well (Fig. 1). In this study, we propose various anonymous speaker identity vectors that can be used for the decoder to generate anonymized speech, and evaluate their performances in terms of speaker identification and speech recognition accuracies. Because we aim to retain the linguistic content of the speech and remove the identity of the original speaker, our goal is to obtain high speech recognition accuracy with low speaker identification accuracy.

A simple method to create anonymous speaker identity vectors, which minimize the voice biometrics of a given speaker, is the use of reversed one-hot vectors that assign 0 to a source speaker and 1s to the non-source speakers. However, preliminary experiments indicated that the resultant speech did not retain the linguistic content. We suspect that the decoder of the VAE is trained to handle a speaker identity vector which is a unit vector. That is, the summation of all its elements should be equal to 1, while the summation of the reversed one-hot vector is equal to $n-1$ for n training speakers. To match the condition of the summation being 1, we assign 0 to the source speaker and assign the value of



(a) One-hot vector

1	0	0	0	0	0	0	0	0	0

(b) Anonymous speaker identity vector: a_1

0	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111

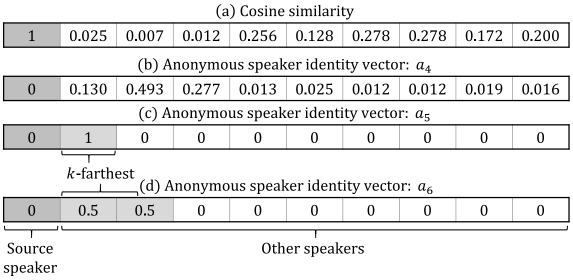
(c) Anonymous speaker identity vector: a_2

−1	0.222	0.222	0.222	0.222	0.222	0.222	0.222	0.222	0.222

(d) Anonymous speaker identity vector: a_3

−0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125

FIGURE 2. Examples of anonymous speaker identity vectors for speaker anonymization using lowered values for a source speaker and uniformly distributed values for other speakers. For convenience, the values for the source speaker are shown in the first column. The number of speakers is 10 in this example.



(a) Cosine similarity

1	0.025	0.007	0.012	0.256	0.128	0.278	0.278	0.172	0.200

(b) Anonymous speaker identity vector: a_4

0	0.130	0.493	0.277	0.013	0.025	0.012	0.012	0.019	0.016

(c) Anonymous speaker identity vector: a_5

0	1	0	0	0	0	0	0	0	0

(d) Anonymous speaker identity vector: a_6

0	0.5	0.5	0	0	0	0	0	0	0

FIGURE 3. Examples of anonymous speaker identity vectors for speaker anonymization: a) cosine similarities; b) normalized version of the inverse of a) with target speaker being set to 0; c) one-farthest speaker is set to 1 and others are set to 0; d) two-farthest speakers are set to 0.5 and others are set to 0.

$1/(n-1)$ to other speakers uniformly in the anonymous speaker identity vector as follows (see a_1 in Fig. 2 (b) for example).

- Value for source speaker: 0
- Values for other speakers: $1/(n-1)$

Another possible method of creating anonymous speaker identity vectors is by assigning negative values to the source speaker to further suppress the voice biometrics of the speaker; for example, -1 is assigned to the source speaker and $2/(n-1)$ to other speakers in the anonymous speaker identity vector as follows (see a_2 in Fig. 2 (c) for example).

- Value for source speaker: -1
- Values for other speakers: $2/(n-1)$

To match the absolute values between speakers, we also tested the possibility of assigning $-1/(n-2)$ to the source speaker and $1/(n-2)$ to other speakers in the anonymous speaker identity vector as follows (see a_3 in Fig. 2 (d) for example).

- Value for source speaker: $-1/(n-2)$
- Values for other speakers: $1/(n-2)$

This yielded a vector of summation 1 with the same absolute values for all speakers.

Fig. 2 illustrates the anonymous speaker identity vector creation methods, where a_1, a_2, and a_3 represent the anonymous speaker identity vectors that have lowered values for the source speaker and uniformly distributed values for other speakers. A conventional one-hot speaker identity vector utilized for voice conversion is also shown in Fig. 2 (a).

B. ANONYMOUS SPEAKER IDENTITY VECTORS USING NON-UNIFORM VALUES

A more sophisticated method of creating anonymous speaker identity vectors can utilize the relative distances between pairs of speakers. If the features of speakers, which are significantly different from those of a source speaker, are boosted, the resultant speech can be expected to have dissimilar characteristics from those of the source speaker. This can be implemented by assigning higher values to the elements in the speaker identity vectors that correspond to the speakers who are separated by a significant distance from the source speaker. The similarities of the speakers can be determined by using i-vectors [21], the most widely used method for speaker recognition. We used the Kaldi toolkit [22] to extract the i-vectors and computed the cosine similarities for each pair of speakers (see Fig. 3 (a) for example).

Because our objective is to assign higher values to dissimilar speakers, the similarities must be converted to distances. Additionally, as evident from Section III-A, the sum of the distance values should be 1. Various methods can be used to convert the cosine similarities to distances and normalize them. One of such methods includes taking the inverse of the cosine similarity values, setting the source speaker value to 0, and normalizing the values to obtain a sum of 1 (see a_4 in Fig. 3 (b) for example).

- Value for source speaker X: 0
- Value for other speaker Y: $\frac{1}{i_X^T i_Y} \Big/ \sum_{Y'} \frac{1}{i_X^T i_{Y'}}$

where i_X is a speaker X's i-vector. Since the application of the softmax function seems to over-compress the ranges of weights, thereby resulting in values similar to that of a_1, we did not used the softmax function for normalization.

Another method involves the determination of k-farthest speakers and uniformly assigning the value of $1/k$ to them and 0s to others (see a_5 and a_6 in Fig. 3 (c) and (d) for example).

- Values for farthest k speakers: $1/k$
- Values for other speakers: 0

Fig. 3 illustrates the examples of anonymous speaker identity vectors using speaker distances. The effects of each method are evaluated in the next section.

IV. EXPERIMENTS

A. SPEECH DATA AND SYSTEM CONFIGURATIONS

We used the voice conversion challenge (VCC) 2016 corpus [23] to evaluate the performance of speaker anonymization techniques in terms of speaker identification and speech recognition accuracies. The VCC 2016 comprises ten speakers (five males and five females) with 162 and 54 utterances for training and testing for each speaker, respectively. Because the VCC 2016 was originally designed for voice conversation tasks, it separates the source and the target speakers.

However, for anonymization, we used all speakers for training and testing.

The input speech was down-sampled to 16 kHz using the SoX toolkit [24] and converted to 36-dimensional Mel-frequency cepstral coefficients (MFCCs) with a frame size and frame shift of 64 and 5 ms, respectively. Other signal processing parameters were identical to those presented in [19] where the CycleVAE-GAN was introduced.

For speaker anonymization, we used the CycleVAE-GAN voice conversion algorithm, explained in Section II-B, with the WORLD vocoder for speech synthesis. The CycleVAE-GAN uses gated linear units [25] for the encoder, decoder, and discriminators. The Adam optimizer [26] with a learning rate and batch size of 0.0001 and 16, respectively, was used.

For speaker identification, we built two systems to assure that the resultant speech data are independent of the characteristics of the speaker identification algorithms; a Gaussian mixture model (GMM)-based speaker identification system and a deep neural network (DNN)-based speaker identification system. For the GMM-based system, we used the HTK toolkit [27] to construct the GMM for each speaker with 1024 mixture components. The DNN-based system consisted of four convolutional layers and one softmax layer, as shown in Fig. 4. To maximize the speaker recognition accuracy of the DNN-based speaker identification system, we used a 40-dimensional log-amplitude Mel-spectrogram with the first and the second order time derivatives for its input feature. To compute the speaker probability, we averaged the softmax probabilities of every 64 frames in each given speech. Both GMM and DNN speaker identification systems were trained using the training set of the VCC 2016.

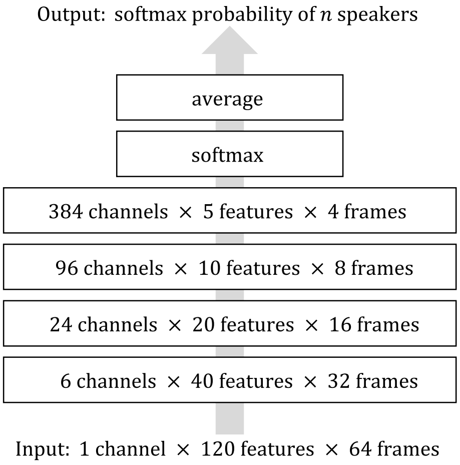

Output: softmax probability of n speakers
average
softmax
384 channels × 5 features × 4 frames
96 channels × 10 features × 8 frames
24 channels × 20 features × 16 frames
6 channels × 40 features × 32 frames
Input: 1 channel × 120 features × 64 frames

FIGURE 4. A DNN-based speaker identification system consisted of four convolutional layers and a fully connected softmax layer.

For speech recognition, we used the Google Cloud STT service (https://cloud.google.com/speech-to-text). Because the VCC 2016 does not have official transcriptions, we manually crafted them.

B. BASELINE RESULTS

Before performing speaker anonymization, we tested the basic performance of the speaker identification systems on the test set of the VCC 2016. Both GMM- and DNN-based speaker identification systems showed an accuracy of 100% for the VCC 2016 test set, indicating that the both systems were trained well.

By measuring the speaker identification accuracy on voice converted data, we can verify whether a voice conversion algorithm can successfully change the speaker identity of a given utterance to another. A voice converted version of the VCC 2016 test set was created using the CycleVAE-GAN: each utterance from a speaker in the test set was converted to rest nine speakers' voice, creating 10 × 9 × 54 utterances. Table 1 shows the accuracies of the speaker identification systems on this voice converted data. The DNN-based system yielded an accuracy of 99.8%, while the GMM-based system produced an accuracy of 94.0%. This implies that the CycleVAE-GAN can successfully modify the identity of a source speaker to that of the target speaker.

TABLE 1. Baseline speaker identification accuracies on the voice converted version of the VCC 2016 test data.

System	Accuracy (%)
GMM	94.1
DNN	99.8

TABLE 2. Baseline speech recognition accuracies on the VCC 2016 test data and the self-converted version of the test data by the CycleVAE-GAN.

Data	Accuracy (%)
Original VCC 2016	88.7
Self-converted VCC 2016	73.4

If the speech recognition accuracy falls dramatically after voice conversion, it can be assumed that the linguistic content of the speech are damaged. To examine the lower bounds of such damages, a self-converted version of the VCC 2016 test set was created where the source and the target speakers were the same during voice conversion: each utterance from a speaker in the test set went through the CycleVAE-GAN with the target speaker identity vector set to the source speaker identity vector. Table 2 summarizes the speech recognition accuracies obtained on the VCC 2016 test set and the self-converted set. It was found that the speech recognition accuracy deteriorated to some extent. This can be partly attributed to the artifacts caused by the WORLD vocoder used to synthesize the output speech.

C. SPEAKER ANONYMIZATION RESULTS

As discussed in Section III, several methods can be used to create anonymous speaker identity vectors that suppress the voice biometrics of the input speaker. We used three uniformly valued anonymous speaker identity vectors, as illustrated in Fig. 2, and three non-uniformly valued anonymous

TABLE 3. Summary of speaker anonymization using various anonymous speaker identity vectors. Lower speaker identification accuracies imply that the speaker anonymization technique is successful in removing the voice biometrics of an original speaker.

Speaker identity vector	Source speaker value	Non-source speaker value	Speaker identification accuracy (%)		Speech recognition accuracy (%)
			GMM	DNN	
a_1	0	$1/(n-1)$	18.8	7.0	69.3
a_2	-1	$2/(n-1)$	0.7	**0.1**	57.0
a_3	$-1/(n-2)$	$1/(n-2)$	11.8	5.7	68.5
a_4	0	Inverse of cosine similarity	15.9	9.2	**72.1**
a_5	0	1 for the 1-farthest speaker and 0 for others	**0.3**	**0.1**	69.3
a_6	0	0.5 for the 2-farthest speakers and 0 for others	20.9	9.6	69.2

speaker identity vectors, as illustrated in Fig. 3. Table 3 summarizes the speaker identification and speech recognition accuracies for various types of anonymous speaker identity vectors. Because our objective is to remove the voice biometrics of the source speaker while retaining the linguistic content of the speech, systems with lower speaker identification accuracies and higher speech recognition accuracies represent better speaker anonymization systems.

It can be observed in Table 3 that all methods successfully suppressed the source speaker identity with some impact on the linguistic content. Considering that the speech recognition accuracy on the self-converted speech data was 73.4%, it is evident that most of the speaker anonymization techniques did not further damage the linguistic content too much. Moreover, it should be noted that because we used ten speakers for speaker identification, random guessing will yield an accuracy of 10%. Therefore, it can be assumed that a speaker identification accuracy of approximately 10% implies that the converted speech is anonymized well. Furthermore, if it is lower than 10%, it implies that the converted speech is explicitly steered to suppress the voice biometrics of the source speaker.

The most obvious case is a_2, which assigns -1 to the source speaker and $2/(n-1)$ to other speakers. It aggressively removes speaker-related information, showing 0.7% of speaker identification accuracy. However, it damaged the linguistic content much, resulting a relative reduction in the speech recognition accuracy of 35.7%.

On the other hand, anonymization using a_4 retained most of the linguistic content. The speech recognition accuracy dropped 18.7% relatively while yielding a speaker identification accuracy of 15.9%.

Converting the identity of the source speaker to the single farthest speaker (a_5) yielded the lowest speaker identification accuracy of 0.3% because it significantly boosts the characteristics of the most different speaker from the source speaker. The linguistic content was moderately maintained: a relative reduction of speech recognition accuracy of 21.9%.

Tables 4 and 5 present the confusion matrices of the GMM- and DNN-based speaker identification systems, respectively, for speeches anonymized using a_4 anonymous speaker identity vectors. After investigating the confusion matrices of the GMM- and the DNN-based systems, it was found that they have quite different characteristics; the GMM-based system tends to classify ambiguous speakers as TF2, TM1, TM2, or TM3, while the DNN-based system tends to classify them as TF1, TF2, TM1, or TM3. That is, TF1 and TM2 show different trends in each system. Since the GMM-based system was better than the DNN-based system

TABLE 4. Confusion matrix of the GMM-based speaker identification system tested on the anonymized speech data using anonymous speaker identity vector a_4.

	SF1	SF2	SF3	SM1	SM2	TF1	TF2	TM1	TM2	TM3
SF1	3	2	2	0	0	0	44	0	2	1
SF2	0	0	2	0	0	0	52	0	0	0
SF3	10	10	5	0	0	0	28	0	0	1
SM1	0	0	0	2	0	0	0	6	4	42
SM2	0	0	0	0	15	0	0	0	34	5
TF1	1	0	47	0	0	0	6	0	0	0
TF2	1	25	0	0	0	0	28	0	0	0
TM1	0	0	0	0	0	0	0	4	40	10
TM2	0	0	0	0	0	0	0	43	0	11
TM3	0	0	0	0	0	0	0	25	0	29

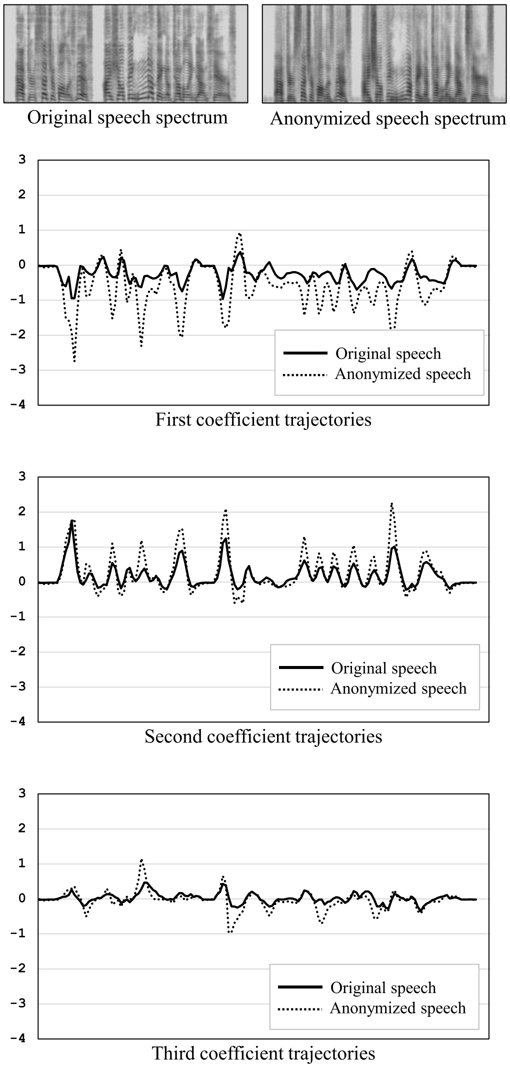
TABLE 5. Confusion matrix of the DNN-based speaker identification system tested on the anonymized speech data using anonymous speaker identity vector a_4.

	SF1	SF2	SF3	SM1	SM2	TF1	TF2	TM1	TM2	TM3
SF1	0	1	5	0	1	1	46	0	0	0
SF2	0	0	0	0	0	54	0	0	0	0
SF3	0	1	0	0	0	36	17	0	0	0
SM1	0	0	0	0	16	0	0	2	0	36
SM2	0	0	0	0	45	0	0	0	0	9
TF1	0	0	54	0	0	0	0	0	0	0
TF2	5	41	3	0	0	0	5	0	0	0
TM1	0	0	0	0	2	0	0	0	13	39
TM2	0	0	0	0	0	0	1	52	0	1
TM3	0	0	0	0	2	0	0	52	0	0

TABLE 6. Speech recognition accuracy on the anonymized speech data using anonymous speaker identity vector a_4.

Speaker	Types of error (number of error words)			Accuracy (%)
	Substitution	Insertion	Deletion	
SF1	109	12	57	67.7
SF2	84	9	16	80.2
SF3	99	10	55	70.2
SM1	72	6	20	82.2
SM2	102	7	24	75.9
TF1	115	8	57	67.3
TF2	79	9	27	79.1
TM1	119	5	38	70.6
TM2	164	8	116	47.7
TM3	85	5	19	80.2
Average	102.8	7.9	42.9	72.1

in speaker identification task for the anonymized speech, the error rates of the GMM-based system were referred above in discussing Table 3.


Original speech spectrum

Anonymized speech spectrum

First coefficient trajectories

Second coefficient trajectories

Third coefficient trajectories

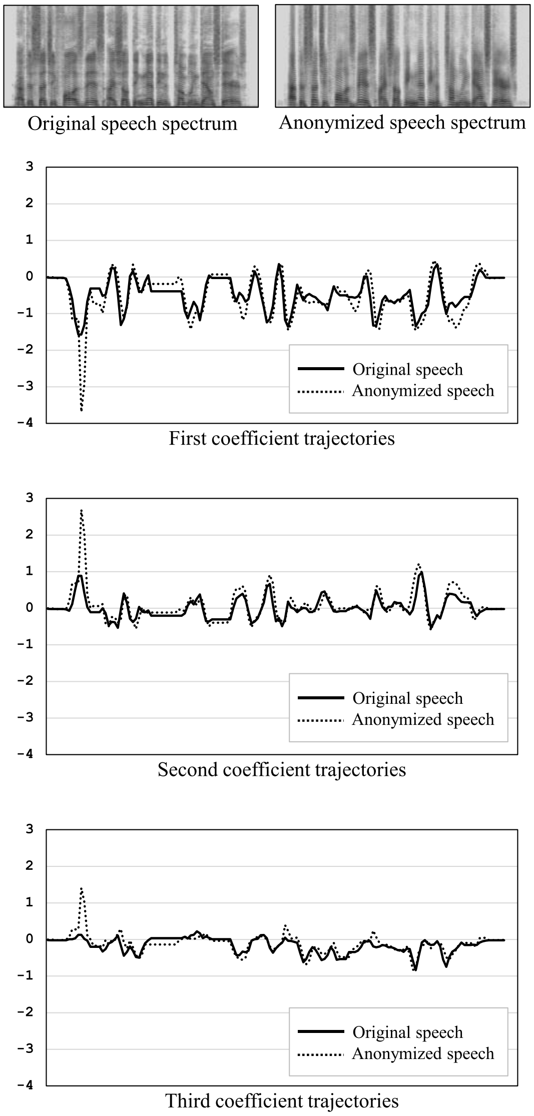
FIGURE 5. The spectrograms and the first three cepstral coefficient trajectories of a sample utterance from speaker SF2 and its anonymized version.

As evident from Table 4, the utterances from speaker SF2 are perfectly anonymized. Fig. 5 shows the spectrograms and the first three cepstral coefficient trajectories of a sample utterance from speaker SF2 and its anonymized version. It can be observed that the trajectory differences are wide. In contrast, 15 utterances out of the 54 utterances from speaker SM2 failed to be anonymized. Fig. 6 shows the spectrograms and cepstral trajectories for speaker SM2. In comparison to Fig. 5, the trajectory differences are relatively narrow.

Table 6 shows the details of the speech recognition accuracies of the same data.

V. CONCLUSION

In this study, we applied voice conversion techniques to achieve speaker anonymization for the security of personal information; the objective of this study was to retain the


Original speech spectrum

Anonymized speech spectrum

First coefficient trajectories

Second coefficient trajectories

Third coefficient trajectories

FIGURE 6. The spectrograms and the first three cepstral coefficient trajectories of a sample utterance from speaker SM2 and its anonymized version.

linguistic content of the given speech while suppressing the voice biometrics of the original speaker. The proposed method modified the conventional one-hot encoded speaker identity vectors to anonymized speaker identity vectors using various methods. The proposed method can anonymize the speech almost perfectly. Some inherent losses were observed in the linguistic content due to the voice conversion process; however, the damage was not severe. Such losses can be partly attributed to the characteristics of the WORLD vocoder used in this study to synthesize the resultant speech. Future works can include other types of vocoders such as the WaveNet [28] or WaveRNN [29].

Because the proposed method modifies only the speaker identity vectors, voice conversion algorithms other than the CycleVAE-GAN can also be used if they utilize speaker identity vectors. The combination of various voice conversion algorithms and vocoders may produce interesting results for speaker anonymization tasks. We plan to investigate this with a larger number of training speakers for the anonymization of the unseen speakers who are not included in the training data.

REFERENCES

[1] S. A. and D. John, "Survey on chatbot design techniques in speech conversation systems," *Int. J. Adv. Comput. Sci. Appl.*, vol. 6, no. 7, pp. 72–80, 2015.

[2] J. G. A. Barbedo, "Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification," *Comput. Electron. Agricult.*, vol. 153, pp. 46–53, Oct. 2018.

[3] X. Zhu, C. Vondrick, D. Ramanan, and C. Fowlkes, "Do we need more training data or better models for object detection?" in *Proc. Procedings Brit. Mach. Vis. Conf.*, Surrey, U.K., 2012.

[4] European Parliament and Council, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 Apr. 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)," *Off. J. Eur. Union*, vol. 59, nos. 1–88, p. 294. 2016.

[5] 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 *Proc. Interspeech*, Graz, Austria, 2019, pp. 3695–3699.

[6] A. Nelus, J. Ebbers, R. Haeb-Umbach, and R. Martin, "Privacy-preserving variational information feature extraction for domestic activity monitoring versus speaker identification," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 3710–3714.

[7] A. Nelus, S. Rech, T. Koppelmann, H. Biermann, and R. Martin, "Privacy-preserving siamese feature extraction for gender recognition versus speaker identification," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 3705–3709.

[8] P. Thaine and G. Penn, "Extracting mel-frequency and bark-frequency cepstral coefficients from encrypted signals," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 3715–3719.

[9] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent, "Privacy-preserving adversarial representation learning in ASR: Reality or illusion?" in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 3700–3704.

[10] A. Nautsch, J. Patino, A. Treiber, T. Stafylakis, P. Mizera, M. Todisco, T. Schneider, and N. Evans, "Privacy-preserving speaker recognition with cohort score normalisation," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 2868–2872.

[11] P. P. Zarazaga, S. Das, T. Bäckström, V. V. R. V., and A. K. Vuppala, "Sound privacy: A conversational speech corpus for quantifying the experience of privacy," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 3720–3724.

[12] Q. Jin, A. R. Toth, T. Schultz, and A. W. Black, "Voice convergin: Speaker de-identification by voice transformation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Taipei, Taiwan, Apr. 2009, pp. 3909–3912.

[13] T. Justin, V. Struc, S. Dobrisek, B. Vesnicer, I. Ipsic, and F. Mihelic, "Speaker de-identification using diphone recognition and speech synthesis," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Ljubljana, Slovenia, May 2015, pp. 1–7.

[14] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[15] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, Jeju, South Korea, Dec. 2016, pp. 1–6.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2014, pp. 2672–2680.

[17] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 3364–3368.

[18] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Non-parallel voice conversion with cyclic variational autoencoder," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 674–678.

[19] D. Yook, S.-G. Leem, K. Lee, and I.-C. Yoo, "Many-to-Many voice conversion using cycle-consistent variational autoencoder with multiple decoders," in *Proc. Odyssey Speaker Lang. Recognit. Workshop*, Nov. 2020, pp. 215–221.

[20] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, vol. 70. Sydney, NSW, Australia, 2017, pp. 214–223.

[21] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[22] 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 *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, Waikoloa, HI, USA, 2011.

[23] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The voice conversion challenge 2016," in *Proc. Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 1632–1636.

[24] C. Bagwell and U. Klauer. *Sox-Sound Exchange*. Accessed: Nov. 1, 2020. [Online]. Available: http://sox.sourceforge/Main/HomePage

[25] Y. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Int. Conf. Mach. Learn.*, vol. 70. Sydney, NSW, Australia, 2017, pp. 933–941.

[26] D. Kingma and L. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, San Diego, CA, USA, 2015.

[27] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, A. Ragni, V. Valtchev, P. Woodland, and C. Zhang, *The HTK book (HTK Version 3.5)*. Cambridge, U.K.: Cambridge University Engineering Department, 2015. [Online]. Available: http://htk.eng.cam.ac.uk/ftp/ software/htkbook-3.5.alpha-1.pdf

[28] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *Proc. ISCA Speech Synth. Workshop*, Sunnyvale, CA, USA, 2016, p. 125.

[29] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," 2018, *arXiv:1802.08435*. [Online]. Available: http://arxiv.org/abs/1802.08435

IN-CHUL YOO (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer science from Korea University, Seoul, South Korea, in 2006, 2008, and 2015, respectively.

He is currently a Research Professor with the Artificial Intelligence Laboratory, Korea University. His research interests include robust speech recognition and speaker recognition.

KEONNYEONG LEE received the B.S. degree in information system engineering from Hansung University, Seoul, South Korea, in 2018, and the M.S. degree in computer science from Korea University, Seoul, in 2020.

He is currently an Associate with ATEC AP Company Ltd., Seongnam-si, Gyeonggi-do, South Korea. This work was done while he was a student with the Artificial Intelligence Laboratory, Korea University. His research interests include voice conversion and speech recognition.

SEONGGYUN LEEM received the B.S. and M.S. degrees in computer science from Korea University, Seoul, South Korea, in 2018 and 2020, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, The University of Texas at Dallas, Richardson, TX, USA. This work was done while he was a Student with the Artificial Intelligence Laboratory, Korea University.

His research interests include speech recognition and deep learning.

HYUNWOO OH received the B.S. degree in computer science from Sangmyung University, Seoul, South Korea, in 2018, and the M.S. degree in computer science from Korea University, Seoul, in 2020.

He is currently an Associate Researcher with the Research Group, FXGear Inc., Seoul. This work was done while he was a Student with the Artificial Intelligence Laboratory, Korea University. His research interests include parallel speech recognition and voice conversion.

BONGGU KO received the B.A. degree in library and information science and the B.S. degree in computer science from Chung-Ang University, Seoul, South Korea, in 2013, and the M.S. degree in computer science from Korea University, Seoul, in 2020.

He is currently a Researcher with the Artificial Intelligence Laboratory, Korea University. His research interests include speech synthesis and descriptive artificial intelligence.

DONGSUK YOOK (Member, IEEE) received the B.S. and M.S. degrees in computer science from Korea University, Seoul, South Korea, in 1990 and 1993, respectively, and the Ph.D. degree in computer science from Rutgers University, Piscataway, NJ, USA, in 1999.

From 1999 to 2001, he worked on speech recognition with the IBM T.J. Watson Research Center, Ossining, NY, USA. He is currently a Professor with the Department of Computer Science and Engineering, Korea University. He is also the Director of the Artificial Intelligence Laboratory, Korea University. His research interests include machine learning and speech processing.