Computer Speech and Language 73 (2022) 101326

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Speaker anonymization by modifying fundamental frequency and x-vector singular value

Candy Olivia Mawalim [a,*], Kasorn Galajit [a,b], Jessada Karnjana [b], Shunsuke Kidani [a], Masashi Unoki [a]

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

ARTICLE INFO

Keywords:
Speaker anonymization
X-vector singular value
Fundamental frequency
Clustering
Subjective evaluation

ABSTRACT

Speaker anonymization is a method of protecting voice privacy by concealing individual speaker characteristics while preserving linguistic information. The VoicePrivacy Challenge 2020 was initiated to generalize the task of speaker anonymization. In the challenge, two frameworks for speaker anonymization were introduced; in this study, we propose a method of improving the primary framework by modifying the state-of-the-art speaker individuality feature (namely, x-vector) in a neural waveform speech synthesis model. Our proposed method is constructed based on x-vector singular value modification with a clustering model. We also propose a technique of modifying the fundamental frequency and speech duration to enhance the anonymization performance. To evaluate our method, we carried out objective and subjective tests. The overall objective test results show that our proposed method improves the anonymization performance in terms of the speaker verifiability, whereas the subjective evaluation results show improvement in terms of the speaker dissimilarity. The intelligibility and naturalness of the anonymized speech with speech prosody modification were slightly reduced (less than 5% of word error rate) compared to the results obtained by the baseline system.

1. Introduction

Various forms of speech are utilized throughout social media. Advanced speech technology, such as voice conversion techniques and speech synthesis, can synthesize or clone speech entirely as a human voice (Fang et al., 2018; Wang et al., 2019). Distributing users' speech publicly on a social network without privacy measures affects the security of speech technology and privacy protection. For example, consider an automatic speaker verification (ASV) system that analyzes speech samples to authenticate a user's access to sensitive information (Irum and Salman, 2019): Without protection, speech samples on the internet could be used for theft of personally identifiable information, fraud, and/or authentication of the ASV system for criminal purposes (Das and Prasanna, 2018; Vestman et al., 2020). Therefore, there must be a solution to the emerging threat of unauthenticated speech signals, such as synthesizing, cloning, and speech conversion (Das et al., 2020). This solution can be achieved by speaker de-identification or anonymization that can conceal personal information in speech signals (Fang et al., 2019).

Several methods have been proposed to solve this problem (Abou-Zleikha et al., 2015; Fang et al., 2019; Jin et al., 2008, 2009; Magariños et al., 2017; Pobar and Ipsic, 2014; Sathiyamurthi and Ramakrishnan, 2017). Cryptography is a conventional method of providing security by concealing the speech signals of both content and personal information (Sathiyamurthi and Ramakrishnan,

* Corresponding author.
E-mail addresses: candyolivia@jaist.ac.jp (C.O. Mawalim), kasorn@jaist.ac.jp (K. Galajit), jessada.karnjana@nectec.or.th (J. Karnjana), kidani@jaist.ac.jp (S. Kidani), unoki@jaist.ac.jp (M. Unoki).

https://doi.org/10.1016/j.csl.2021.101326
Received 8 January 2021; Received in revised form 3 September 2021; Accepted 3 November 2021
Available online 27 November 2021

2017). However, cryptography does not protect a speech signal once the content is decrypted. Previously, voice transformation was utilized to suppress speaker identity for anonymization purposes (Jin et al., 2008, 2009). Kaldi phone, a diphone-based syntactic source speech, was transformed to attack the speaker identification system, successfully fooling the Gaussian mixture model (GMM)-based speaker identification system (Jin et al., 2008). Subsequently, de-identification of online speakers was feasible with a voice transformation method that de-identifies speaker using GMM mapping and harmonic-stochastic models (Pobar and Ipsic, 2014). Next, a voice transformation technique that uses a target's natural speech instead of a synthetic voice was developed to conceal speaker identity (Abou-Zleikha et al., 2015). Cepstral frequency warping is another alternative approach implemented with an amplitude scaling technique to transform speech and hide identity (Magariños et al., 2017).

Recently, an anonymization method based on a neural source-filter (NSF) model was proposed by Fang et al. (2019). This method separates the speaker identity and the linguistic content from the input speech. An x-vector that refers to the speaker identity was modified to hide personal information before resynthesizing the speech data. In the VoicePrivacy Challenge 2020 (Tomashenko et al., 2020b), this method was introduced as the primary baseline system because the x-vector could effectively encode speaker identity as a feature in a speaker verification system (Snyder et al., 2018). Another baseline introduced in the VoicePrivacy Challenge 2020 used the McAdams coefficient (McAdams, 1984) to transform the spectral envelope of speech signals to achieve speaker de-identification (Patino et al., 2020). The objective evaluation results showed the primary baseline based on the NSF model hid speaker information better than the second baseline based on the McAdams coefficient (Tomashenko et al., 2020b), confirming that the x-vector is a practical feature to represent speaker identity information.

In this study, we extend our prior work Mawalim et al. (2020) that aimed to improve the primary baseline system by modifying the x-vector singular value for speaker anonymization. Our preceding experimental results in Mawalim et al. (2020) showed that our method improves the anonymization rate and is comparable with the baseline system. There is also room for improvement that we intend to present in this study. First, we thoroughly analyze the effectiveness of x-vector modification with singular value decomposition (SVD) by considering various singular value thresholds. We predict that modifying the significant elements represented in an x-vector singular value (SV) can fulfill the speaker-to-speaker correspondence requirement in an anonymization system. Second, despite using a regression model as in our prior work Mawalim et al. (2020), we construct a clustering model for selecting a set of x-vectors for generating the pseudo-target x-vector. Third, we modify acoustic features such as fundamental frequency (F_0) and speech duration to improve our method. The F_0 and speech duration are strongly related to the perception of speaker individuality (Akagi and Ienaga, 1997; Dellwo et al., 2007), so modifying these features should de-identify speaker individuality. To evaluate the performance and effectiveness of our method, we conduct an objective evaluation that follows the VoicePrivacy Challenge 2020 (Tomashenko et al., 2020b), and we propose a more reliable subjective evaluation for assessing the privacy and utility-related metrics in a speaker anonymization system.

The rest of this paper is organized as follows. Section 2 describes the general speaker anonymization system, including the baseline system and its evaluation methods. Section 3 presents our method. Section 4 details the experimental setting, evaluation, and results. Section 5 discusses our evaluation results and remaining limitations. Section 6 concludes the paper and discusses future work.

2. Speaker anonymization

2.1. Definition

Speaker anonymization (also known as de-identification) is a method of protecting voice privacy. It works by concealing the personally identifiable information of uttered speech without degrading the linguistic information (Fang et al., 2019). The VoicePrivacy Challenge 2020 was organized as an initiative to generalize the task and metrics of speaker anonymization (Tomashenko et al., 2020a).

A speaker anonymization system must meet four requirements in accordance with the VoicePrivacy Challenge 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.

Several open-source corpora are introduced in the VoicePrivacy Challenge 2020 to develop a speaker anonymization system, as follows:

(a) LibriSpeech (Panayotov et al., 2015), a corpus of English read speech designed for automatic speech recognition (ASR). This corpus contains a total of approximately 1000 h of 16 kHz speech.
(b) LibriTTS (Zen et al., 2019), a corpus of approximately 585 h of 24 kHz speech that derived from LibriSpeech corpus and designed for text-to-speech (TTS).
(c) VCTK (Veaux et al., 2017), a corpus of approximately 44 h of 48 kHz English read speech spoken by 109 native speakers with various accents and initially designed for TTS.
(d) VoxCeleb-1,2 (Chung et al., 2018; Nagrani et al., 2017), an audiovisual corpus designed for speaker verification research. This corpus contains approximately 2770 h of 16 kHz speech spoken by 7360 speakers in various accents and languages.

These corpora were divided into several subsets for training, development, and evaluation. The detail description and statistics of these subsets was explained in the VoicePrivacy Challenge 2020's evaluation plan (Tomashenko et al., 2020b).

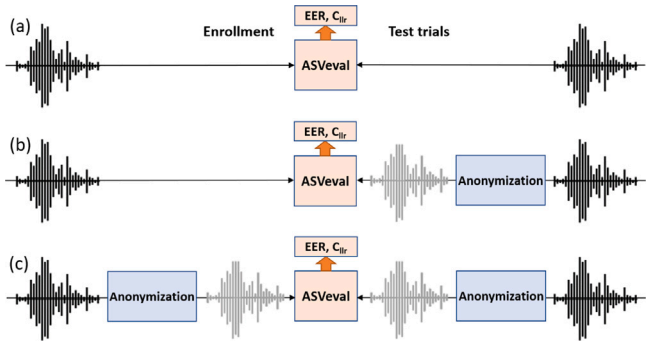

Fig. 1. ASV evaluation for (a) clean trial and enrollment (o-o), (b) anonymized trial and clean enrollment (o-a), and (c) anonymized trial and enrollment (a-a) (Tomashenko et al., 2020b).

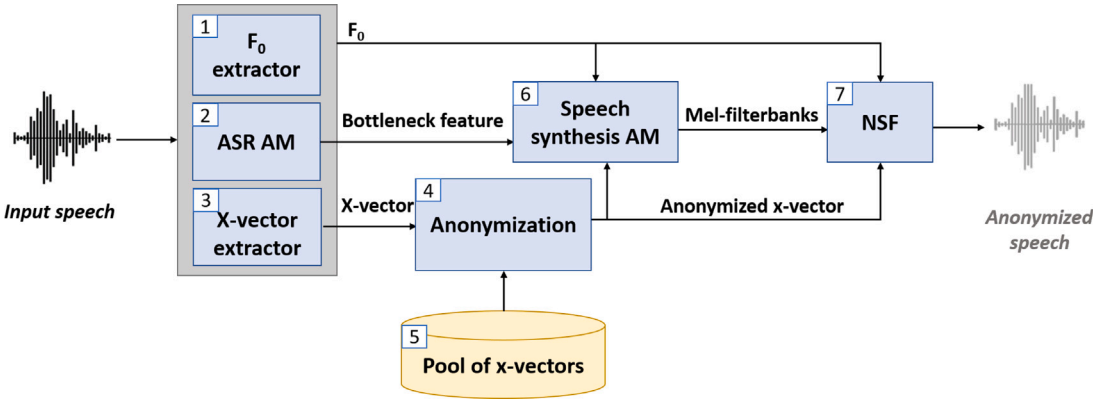

Fig. 2. Schematic diagram of baseline speaker anonymization system (Tomashenko et al., 2020b).

2.2. Evaluation metrics

The privacy and utility metrics should objectively assess an anonymization system based on the requirements mentioned in Section 2.1. The privacy metric should measure the speaker verifiability, whereas the utility metric should measure the ability to preserve the linguistic content. An ASV system is deployed to assess the speaker verifiability metric and an ASR system is deployed to assess utility metric (Tomashenko et al., 2020b). Both ASV and ASR systems for assessing an anonymization system (hereafter, we refer to these systems as ASVeval and ASReval) are trained on a subset of the LibriSpeech dataset (LibriSpeech-train-clean-360) using a Kaldi toolkit (Povey et al., 2011).

An evaluation using ASVeval is conducted utilizing probabilistic linear discriminant analysis (PLDA) on the x-vector (state-of-the-art speaker embedding) (Snyder et al., 2018), under the three conditions shown in Fig. 1. In ASVeval, the equal error rate (EER) and log-likelihood-ratio cost function (C_{llr} and C_{llr}^{min}, proposed in Brümmer and du Preez (2006)), are computed as the objective verifiability metrics. On the other hand, the evaluation using ASReval is conducted based on a factorized time delay neural network (TDNN-F) acoustic model (AM) (Fang et al., 2019; Peddinti et al., 2015) and a trigram language model using a Kaldi recipe for a LibriSpeech dataset. The word error rate (WER) is computed to identify the intelligibility of the anonymized speech in comparison with the original speech only in the trial.

2.3. Baseline systems

In the VoicePrivacy Challenge 2020, two anonymization techniques were introduced as the baseline systems (Tomashenko et al., 2020b). The primary baseline (B1) system was developed using x-vectors and an NSF model (Fang et al., 2019). The second baseline (B2) system was developed based on linear prediction analysis using McAdams coefficient (McAdams, 1984). In this paper, we focus on developing an anonymization system based on the B1.

The B1 system is primarily built on the idea of separating linguistic content and speaker individuality features from the input speech. The anonymized speech is then synthesized by the extracted linguistic content (to preserve the linguistic information) and the modified speaker individuality feature. Fig. 2 shows the block diagram of the B1 system, which consists of seven components: an F_0 extractor, an ASR AM, an x-vector extractor, an anonymization model, a pool of x-vectors, a speech synthesis AM, and an NSF model. The anonymization process is subdivided into the following three main steps:

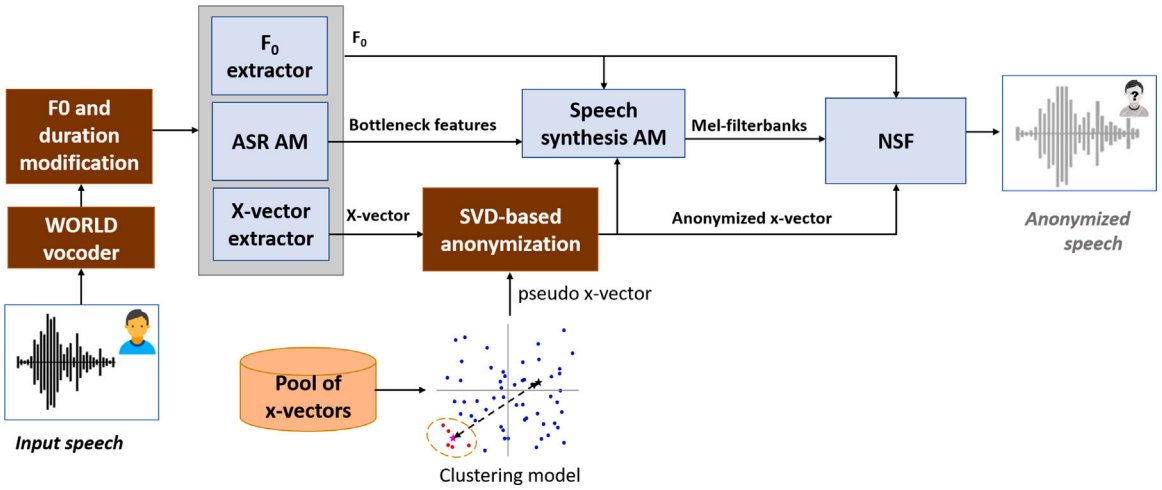

Fig. 3. Schematic diagram of proposed speaker anonymization system.

(i) Feature extraction: extraction of the F_0, a bottleneck feature (as linguistic feature representation using an ASR acoustic model (AM) model Fang et al., 2019; Peddinti et al., 2015), and a speaker individuality feature (x-vector based on Snyder et al., 2018);
(ii) X-vector anonymization: modification of the extracted x-vector by averaging a set of candidate x-vectors from the pool of x-vectors; and
(iii) Speech synthesis: speech synthesis using the F_0, the bottleneck features, and the modified/anonymized x-vector based on the speech synthesis AM (Fang et al., 2019) and NSF (Wang et al., 2019) models.

The Kaldi toolkit (Povey et al., 2011) is used in the feature extraction step. The YAAPT algorithm is used as the F_0 extractor. Subsequently, an ASR AM model is built based on factorial time delay neural network (TDNN-F) model architecture (Fang et al., 2019; Peddinti et al., 2015) and trained using the training data of the LibriSpeech dataset (Panayotov et al., 2015) to extract the bottleneck feature. The output of an x-vector extractor constructed using a time delay neural network (TDNN) model (Snyder et al., 2018) and trained using the VoxCeleb-1,2 dataset is used to represent the speaker individuality feature.

In the x-vector anonymization step, the x-vector of a given input speaker is modified by a new pseudo x-vector obtained by averaging a set of candidate x-vectors determined by a given similarity distance range. The candidate x-vectors belong to the pool of x-vectors extracted from the train-other-500 subset of the LibriTTS dataset (Zen et al., 2019). The cosine similarity, or probabilistic linear discriminant analysis (PLDA), is used as the similarity distance measure. A smaller set of x-vectors is randomly chosen from a set of most farthest x-vectors as the candidate x-vectors.

As the last step, the anonymized speech is resynthesized using a speech synthesis AM model and an NSF model. Both models were trained using the train-clean-100 of the LibriTTS dataset. The speech synthesis AM model was constructed based on an autoregressive network (Fang et al., 2019). This model transforms the input F_0, bottleneck features, and anonymized x-vector into Mel-filterbanks features. Subsequently, the NSF model (Wang et al., 2019) is used to generate the anonymized speech from the F_0, Mel-filterbanks features, and the anonymized x-vector.

3. Proposed method

This section demonstrates our contributions based on the B1 system. Our hypothesis is that modifying the F_0 and the x-vector anonymization model by using SVD could improve the verifiability performance of a speaker anonymization system. We apply the modification to components 4 and 5 in Fig. 2. Fig. 3 shows the schematic diagram of our proposed method.

3.1. Modification of F_0 and speech duration

F_0 contours and their dynamics are strongly related to speaker individuality (Akagi and Ienaga, 1997; Dellwo et al., 2007) because F_0 is an important physical factor that affects pitch perception. Furthermore, it accommodates the perception of several kinds of paralinguistic and prosodic information (Gussenhoven et al., 1997; Hirose and Kawanami, 2002). For instance, we could classify a speaker's gender solely by using the F_0 as the feature because the F_0 of female speech is generally higher than that of male speech.

In this study, we modify the F_0 using the mean F_0 information of adult female and male speakers from a previous study Traunmüller and Eriksson (1995) by using WORLD vocoder (Morise et al., 2016). We classify the speech into a high F_0 and a low F_0 by comparing the mean F_0 of the input utterance and the mean F_0 regarding gender. Accordingly, we convert the low F_0 into a high

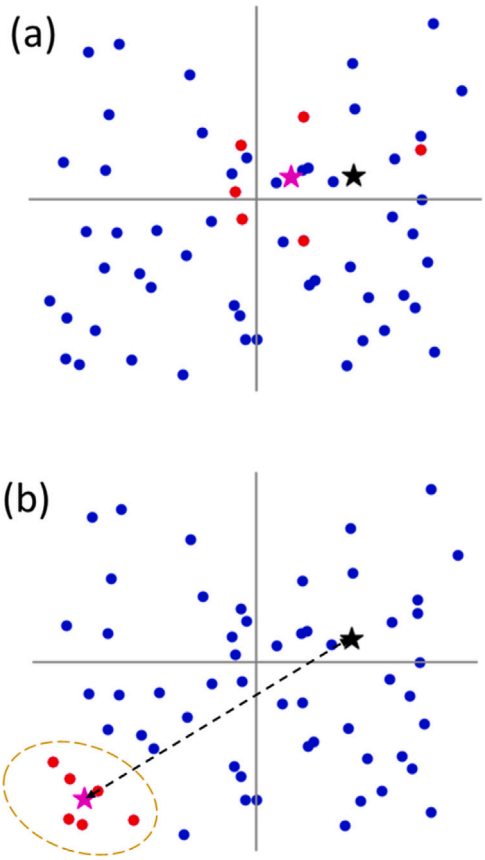

Fig. 4. Illustration of x-vector selection algorithm using: (a) random selection and (b) clustering-based selection. Round blue markers indicate set of x-vector candidates, round red markers indicate chosen x-vector candidates, black star markers indicate given input x-vectors, and magenta star markers indicate chosen pseudo-target x-vectors. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

F_0 by 1.5 times and vice versa (as shown in Eq. (1)). Factor 1.5 was chosen because our preliminary experiment showed that it is the largest factor that outcome the possible F_0 range (Traunmüller and Eriksson, 1995) in the dataset.

$$f_0'(n) = \begin{cases} 1.5 \times f_0(n), & \overline{f_0}(n) \leq \overline{F_0} \\ f_0(n)/1.5, & \overline{f_0}(n) > \overline{F_0} \end{cases}, \tag{1}$$

where m indicates the time frame, $f_0(n)$ and $f_0'(n)$ are the original F_0 and the modified F_0 in the time domain, respectively. The $\overline{f_0}(n)$ is the mean F_0 value of original F_0, whereas $\overline{F_0}$ is the mean female or male F_0 value based on Traunmüller and Eriksson (1995).

In addition to F_0 modification, we carry out speech duration modification. Duration is a speech property relevant to expressing "stress" in speaking. Consequently, the speaking rate varies from speaker to speaker (Dellwo et al., 2007). Speaking rates have been reported to significantly affect speaker verification system performance (Das et al., 2018). In this study, we lengthen speech by increasing the frame duration by 1.2 times because the mismatched speech tempo could be minimized by this factor (minimizing the possible distortion caused by this modification) (Das et al., 2018).

3.2. Pseudo-target generation

In contrast to a voice conversion system, the speaker target is unknown in an anonymization system, so a target anonymized speaker (pseudo-speaker) must be determined. The x-vector of an input speaker in the B1 system was modified as a speaker individuality feature using a selection algorithm on a pool of x-vectors (as explained in Section 2.3) (Tomashenko et al., 2020b). This selection algorithm was utilized by randomly choosing 100 x-vectors from a set of 200 x-vectors obtained from speakers who were the furthest distance from the input speaker. The distance was determined using the PLDA.

As Fig. 4 (a) shows, the average x-vector from the randomly selected subset from the furthest x-vectors set can cause the input speaker's x-vector to be given nearby. To reduce occurrences of this problem, we constructed a gender-dependent clustering model based on k-means as the selection algorithm for a set of the furthest x-vectors of the same-gender utterances. K-means clustering is commonly used because of its simple implementation and because it scales to large datasets and guarantees convergence (Bottou

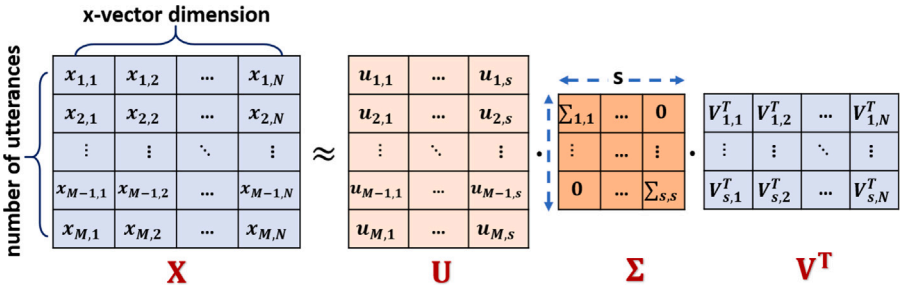

Fig. 5. Modification of x-vector SVs (Mawalim et al., 2020). The $x_{i,j}$ refers to the element of matrix **X** in row i and column j. Similarly, $u_{i,j}$, $\Sigma_{i,j}$, and $V^T_{i,j}$ are the elements of matrix **U**, **Σ**, and $\mathbf{V^T}$ in row i and column j, respectively. The $\mathbf{V^T}$ is the transpose matrix of **V**. The **s** determines the number of singular values.

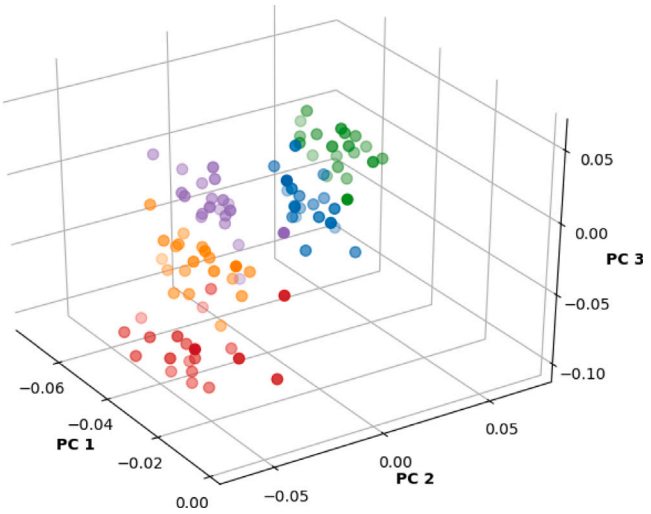

Fig. 6. Principal components (PCs) of x-vectors from five speakers in VCTK development dataset for enrollment in 3D space. Colors represent speaker labels (e.g., round orange markers represent class of x-vectors of speaker with ID label "p234"). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and Bengio, 1994). The pseudo x-vector was then determined using the mean of the set x-vectors. Fig. 4 (b) illustrates the selection algorithm of our method.

3.3. SVD-based x-vector anonymization

For the x-vector anonymization technique, we applied one of the matrix factorization concepts in linear algebra, namely, SVD (Golub and Reinsch, 1970). SVD is widely used for dimension reduction applications (e.g., data compression and denoising) because it provides a more stable matrix decomposition than the other methods (Goodfellow et al., 2016; Sahidullah and Kinnunen, 2016). The SVD technique decomposes a given input matrix into its constituent elements based on the polar decomposition. Mathematically, the SVD is expressed by the following equation.

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}, \tag{2}$$

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

The x-vectors are extracted from a variety of utterances spoke by a speaker which is not equivalent to each other. However, a PLDA classifier distinguishes which speaker the x-vectors originated from Snyder et al. (2018). Principal component analysis shows that the distribution of a single speaker's x-vectors are clearly clustered close together (as shown in Fig. 6). Considering those preliminary studies, x-vector anonymization by SVD could capture the eigenstructure and result in a better representation of intra-speaker information. Thus, modifying the SV of the x-vectors matrix could satisfy the speaker-to-speaker correspondence requirement (the x-vectors of a given speaker should not be similar to the other speakers).

Our SVD-based x-vector technique (Mawalim et al., 2020) is conducted in the following three steps (shown in Fig. 7):

i. *Pseudo-target x-vector matrix formation*

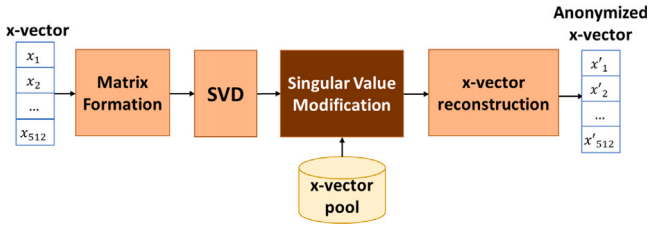

Fig. 7. Schematic diagram of x-vector modification by SVD (Mawalim et al., 2020). x_i and the x'_i are the ith element of input x-vector and anonymized x-vector, respectively.

After the pseudo-target x-vectors of a given speaker from all available train utterances were chosen using the clustering model, we concatenated those x-vectors into a matrix ($\mathbf{X}$) that had an $M \times N$ dimension, where M is the total available utterances and N is the dimension of the x-vector (512).

ii. *SV decomposition and modification*
The pseudo-target x-vector matrix was decomposed by SVD (as shown in Eq. (2)) into two singular matrices ($\mathbf{U}$ and $\mathbf{V}$) and a diagonal singular values matrix ($\mathbf{\Sigma}$). In this approach, $\mathbf{U}$ could be interpreted as the utterance-to-concept similarity matrix and $\mathbf{V}$ as the x-vector-to-concept similarity matrix. $\mathbf{\Sigma}$ represents the strength of each concept involved. The anonymization was conducted by controlling the dimension of $\mathbf{\Sigma}$ using a threshold parameter ($\mathbf{s}$) to obtain more general constituent elements of the x-vector. Fig. 5 shows the x-vector anonymization by SV modification.

iii. *Anonymized x-vector reconstruction*
After the SV modification, we reconstructed the modified matrix using $\mathbf{U}$, $\mathbf{V}$, and the modified $\mathbf{\Sigma}$. The anonymized x-vector of the given utterance was then extracted accordingly.

4. Experiments

The experiments were entirely based on the protocols and datasets provided in the VoicePrivacy Challenge 2020[1] (Tomashenko et al., 2020b). In this Section, we provide the specific description of our method, including the datasets we used, our experiments, and our evaluation settings.

4.1. Datasets

We conducted our experiments using four publicly open-source corpora as described in Section 4 of the VoicePrivacy Challenge 2020's evaluation plan (Tomashenko et al., 2020b): LibriSpeech (libri) (Panayotov et al., 2015), LibriTTS (Zen et al., 2019), the voice cloning toolkit (VCTK) (Veaux et al., 2017), and VoxCeleb-1,2 (Chung et al., 2018; Nagrani et al., 2017). Each corpus was split into training, development, and testing data. Additionally, "common part" and "different part" subsets of trial utterances were constructed specifically for the VCTK dataset to evaluate speaker verifiability regardless of text-dependency. The common part consisted of the utterances that were identical for all the speakers, and the different part consisted of the distinct utterances for all the speakers.

We utilized the available training subsets from LibriTTS (train-other-500 and train-clean-100) (Zen et al., 2019), comprised of approximately 1400 speakers and 240,000 total utterances. Table 1 shows the statistics of these datasets. We evaluated our method with both development and test data of libri (Panayotov et al., 2015) and VCTK (Veaux et al., 2017) in ASVeval and ASReval.

4.2. Experimental setting

The experiments were conducted using the Kaldi toolkit (Povey et al., 2011) for the main anonymization framework, WORLD for F_0 modification (Morise et al., 2016), and the scikit-learn software (Pedregosa et al., 2011) for the k-means clustering model. First, we analyzed the input signal using WORLD to obtain the F_0, the aperiodicity, and the spectral envelope. Subsequently, we modified the F_0 based on Eq. (1) and the frame duration before re-synthesizing the speech.

The output of the resynthesized speech was given input to the NSF-based anonymization system. In the x-vector anonymization block (sub-element 4 in Fig. 2), we utilized the pseudo-target generation and x-vector SV modification as explained in Section 3. We conducted our method's entire process separately for each gender.

In pseudo-target generation, we chose the 200 furthest x-vectors using PLDA and clustered those x-vectors into 50 groups by the k-means algorithm. Finally, the pseudo x-vector of a given speaker was determined by the centroid furthest from the corresponding x-vector. In these experiments, we also analyzed the effect of controlling the SV threshold parameter. The threshold values are 0.9 and 0.8.

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

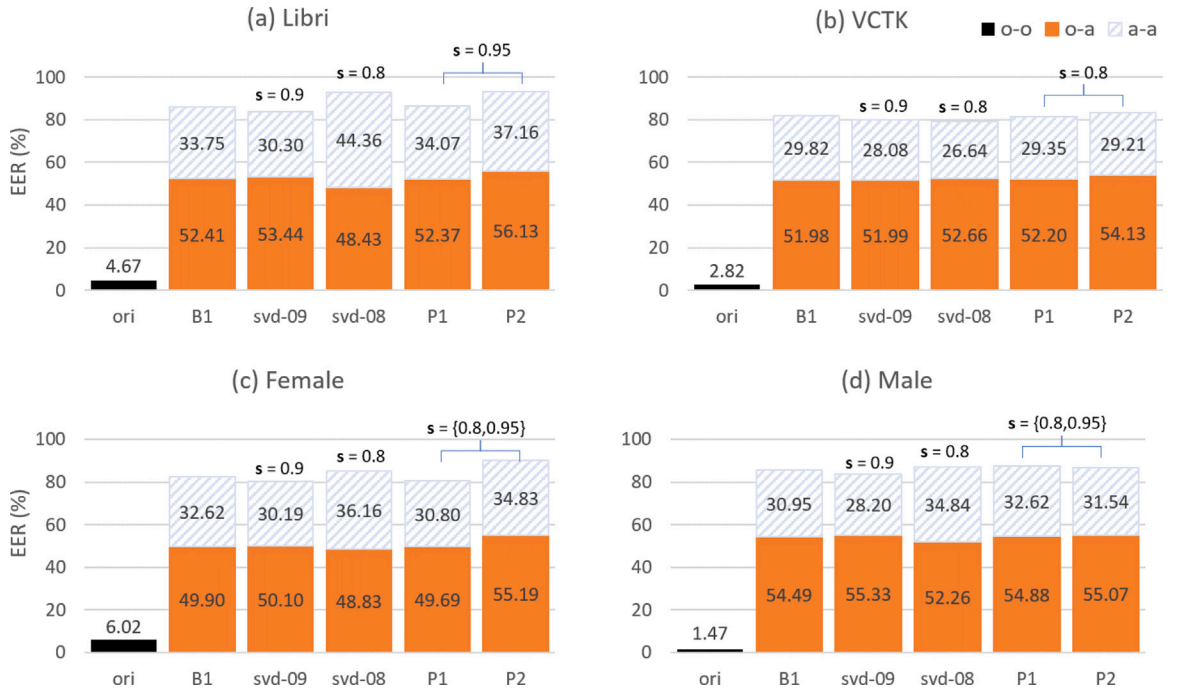

Fig. 8. Average ASVeval results from controlling SV threshold using k-means clustering and modification of F_0 and duration. Original speech as "ori" denotes ASVeval results using both original enrollment and trials (o-o). "B1" denotes results of ASVeval using primary baseline model (Tomashenko et al., 2020b). "svd-09" and "svd-08" denote ASVeval results by x-vector SV modification with thresholds (**s**) 0.9 and 0.8, respectively. "P1" denotes ASVeval results obtained by x-vector SV modification with k-means clustering, whereas "P2" denotes results with additional F_0 and speech duration modification. Orange bars represent results in pairs of original enrollment and anonymized trials (o-a). Gray shaded bars represent results in pairs of anonymized enrollment and anonymized trials (a-a). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Training data for pool of x-vectors.

Subset	Female	Male	Total	#Utter
train-clean-100	123	124	247	33,236
train-other-500	560	600	1160	205,044

4.3. Evaluation

We evaluated our method with both objective and subjective tests. The evaluation was conducted based on a comparative study of the B1 method and our method.

4.3.1. Objective test

The general procedure of the objective test was based on the VoicePrivacy Challenge 2020 (Tomashenko et al., 2020b) and investigated three points. First, we investigated how effectively we could control the SV threshold. Second, we investigated how effectively we could select pseudo x-vectors using k-means clustering. Third, we investigated how the anonymization system performs by modifying the F_0 and speech duration.

Fig. 8 compares the average results of the corresponding development and test datasets of the speaker verifiability assessment using ASVeval with the B1 system with those of our method. These results were obtained by averaging the EER results of development and test datasets. The subset of the common part is not available in the LibriSpeech dataset; therefore, the results in Fig. 8 were derived from all the results, excluding the subset of the common part of the VCTK dataset. To determine the effectiveness of the x-vector SV modification, we conducted the experiments without using the random selection provided in the baseline (only the mean value of the 200 furthest x-vectors). The results shown in Fig. 8 indicate results comparable to the B1. By controlling the threshold, we could slightly improve speaker verifiability.

The corresponding results of ASReval are provided in Fig. 9. The ASReval performance with x-vector SV modification was slightly better than the baseline, except for the LibriSpeech dataset with a 0.8 threshold (svd-08). Although the ASVeval results from using svd-08 in the a-a case were improved for the LibriSpeech dataset, the intelligibility in terms of ASReval degraded significantly. We predict that this occurred because the LibriSpeech dataset is a clean dataset. The dimension reduction modification on the SV could distort the constituent elements of the x-vectors. To compensate for this degradation, we used the 0.95 threshold parameter for the

Table 2
Detailed ASVeval results using only x-vector SV modification with 0.95 threshold for LibriSpeech and 0.8 threshold for VCTK (SV Modif), our P1 method, and our P2 method.

Dataset	Gen	Anonymization		SV Modif			SV Modif + k-means (P1)			SV Modif + k-means + F_0 (P2)		
		Enroll	Trial	EER (%)	C_{llr}^{min}	C_{llr}	EER (%)	C_{llr}^{min}	C_{llr}	EER (%)	C_{llr}^{min}	C_{llr}
Libri (dev)	F	ori	ori	8.67	0.30	42.86			=			
			anon	51.99	1.00	147.21	50.57	0.998	145.131	**55.82**	1.00	**156.61**
		anon		32.95	0.86	14.25	35.37	0.88	14.694	**38.35**	0.92	**27.00**
	M	ori	ori	1.24	0.03	14.25			=			
			anon	58.70	1.00	170.42	57.76	0.999	169.887	**60.4**	1.00	**174.72**
		anon		28.88	0.78	18.43	34.01	0.861	24.696	**34.32**	0.87	**29.11**
Libri (test)	F	ori	ori	7.66	0.18	26.79			=			
			anon	48.72	1.00	151.98	48.36	0.996	152.426	**55.29**	1.00	**153.69**
		anon		28.65	0.78	12.73	31.02	0.819	15.449	**39.23**	0.91	**36.93**
	M	ori	ori	1.11	0.04	15.30			=			
			anon	54.34	1.00	168.93	52.78	**0.999**	169.064	**53.01**	1.00	**168.19**
		anon		30.73	0.81	24.20	35.86	0.903	34.784	**36.75**	**0.90**	**39.89**
VCTK common (dev)	F	ori	ori	2.62	0.09	0.87			=			
			anon	50.87	1.00	167.48	49.71	1.00	175.25	**54.07**	1.00	**187.25**
		anon		24.42	0.70	7.12	**26.16**	0.71	6.66	24.71	0.72	**21.19**
	M	ori	ori	1.43	0.05	1.56			=			
			anon	**57.26**	1.00	191.60	55.27	1.00	194.49	56.13	1.00	**207.22**
		anon		25.93	0.71	18.20	**32.76**	0.84	23.69	26.21	0.72	**23.89**
VCTK diff (dev)	F	ori	ori	2.86	0.10	1.13			=			
			anon	50.14	0.99	165.94	51.04	0.99	168.53	**54.86**	1.00	**188.41**
		anon		26.78	0.77	8.72	25.32	0.74	8.13	**26.67**	0.73	**12.88**
	M	ori	ori	1.44	0.05	1.16			=			
			anon	**55.98**	1.00	166.42	54.39	1.00	168.77	52.06	1.00	**175.87**
		anon		25.31	0.74	18.28	**29.73**	0.82	**22.89**	25.46	0.74	17.49
VCTK common (test)	F	ori	ori	2.89	0.09	0.87			=			
			anon	50.00	1.00	156.09	50.00	1.00	156.09	**56.36**	1.00	**178.95**
		anon		28.61	0.80	8.81	**30.92**	0.83	9.12	30.64	0.83	**21.41**
	M	ori	ori	1.13	0.04	1.04			=			
			anon	**55.65**	1.00	186.48	54.80	1.00	191.83	52.54	1.00	**204.88**
		anon		20.34	0.62	9.79	**30.51**	0.82	20.60	25.42	0.72	**23.73**
VCTK diff (test)	F	ori	ori	4.89	0.17	1.50			=			
			anon	49.64	1.00	142.88	48.77	1.00	148.15	**54.78**	1.00	**161.07**
		anon		32.66	0.87	11.36	31.48	0.84	11.50	**35.08**	0.87	**18.15**
	M	ori	ori	2.07	0.07	1.82			=			
			anon	54.31	1.00	164.68	54.59	1.00	168.63	**54.82**	1.00	**179.63**
		anon		21.81	0.67	13.26	**30.88**	0.84	**22.90**	29.62	0.83	18.26

"Gen" stands for gender (F: female and M: male). "=" stands for the equivalent results to the left columns.

LibriSpeech dataset and 0.8 for the VCTK dataset in the following experiments. In practical use, when we have information about the characteristics of the dataset, we could follow the threshold parameters of proposed methods (e.g., $\mathbf{s} = 0.8$ when the dataset has high variabilities as the VCTK dataset). We suggest the value of 0.9 as the threshold parameter for a completely unknown dataset to maintain a general anonymization performance based on x-vectors SV modification.

The two right bars in Fig. 8 show the comparative average ASVeval results of our method using x-vector SV modification with k-means clustering (P1) and additional F_0 and speech duration modification (P2). Overall results from this figure show no significant difference between the B1 method and the P1 method. However, the additional F_0 and speech duration modification in the P2 method could effectively improve the performance of the anonymization approximately 5% of the EER score in the o-a case and up to approximately 3% in the a-a case. The P2 method clearly yields better results than the B1 method, especially with the LibriSpeech dataset and the female utterances.

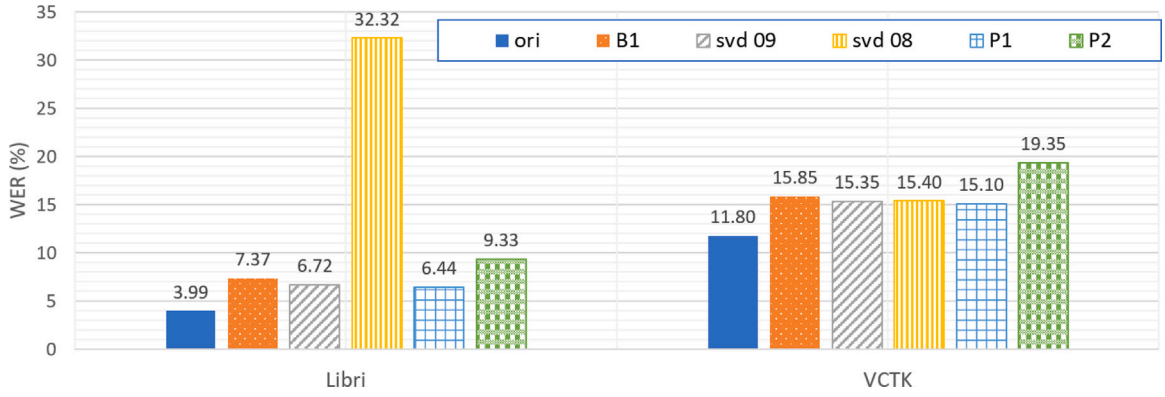

Fig. 9. ASReval results of ori, anonymized speech by B1, controlling SV threshold (svd-09, svd-08), modifying x-vectors SV (**s** = {0.8, 0.95}) with k-means clustering (P1), and modifying F_0, speech duration, and x-vector SV modification (**s** = {0.8, 0.95}) with k-means clustering (P2).

In addition to the EER score, we calculated the log-likelihood-ratio cost function for evaluating the speaker verifiability of the methods using x-vector SV modification with 0.95 threshold for LibriSpeech and 0.8 threshold for VCTK (SV Modif) only, SV Modif with k-means clustering (P1), and SV Modif with k-means and F_0 modification. The details of ASVeval are provided in Table 2. The increasing trend in the EER and C_{llr} scores indicates improvement in the privacy metric of a speaker anonymization system. In terms of the utility metric, the ASReval results of the P1 and P2 systems shown in Fig. 9 are almost similar to the B1 system.

4.3.2. Subjective test

Compared with the B1 method, our P2 method showed distinguishable results for the speaker verifiability assessment using ASVeval and slightly inferior assessment results using ASReval. However, these results cannot sufficiently determine the effectiveness of an anonymization system. For instance, Table 2 shows that the P2 results are not always better than the P1 results. Therefore, we propose a subjective test that considers human hearing perception in the speaker anonymization system assessment.

In the initial state, we focused on the main purpose of an anonymization system, which is to conceal as much personally identifiable information as possible while maintaining the naturalness and intelligibility of the speech. The attack model has not yet been considered in this test. Three metrics were used for the subjective evaluation: speech intelligibility, speech naturalness, and speaker dissimilarity. Since the listeners did not know the context of the spoken utterances, we define "intelligible speech" as speech that contains words that can clearly be heard in the corresponding language. In this experiment, the words are in English. Meanwhile, "natural speech" is the speech most closely perceived as a human voice.

We conducted our subjective evaluation with a listening test divided into two main parts. The first part was measuring the intelligibility and naturalness metrics. A 5-point scale was used for both intelligibility (1-mostly unintelligible, 2-somewhat unintelligible, 3-cannot decide, 4-somewhat intelligible, 5-mostly intelligible) and naturalness (1-mostly unnatural, 2-somewhat unnatural, 3-cannot decide, 4-somewhat natural, 5-mostly natural). The second part was measuring the verifiability metric. We provided paired stimuli (original and anonymized utterances) and asked the participants to determine whether the speakers of those two stimuli were the same. The similarity metric was also a 5-point scale (1-completely similar, 2-mostly similar, 3-somewhat similar, 4-mostly different, and 5-completely different).

Twenty-four participants (thirteen men and eleven women, aged 20–35) were employed in our subjective test. Each participant had a normal hearing ability and was a non-native speaker with a B2 English proficiency level. We conducted a paired-comparison test (with the original utterance provided as reference) to compensate for any bias from the participants as non-native speakers.

The subjective evaluation compared three methods: B1, P1, and P2. Nine stimuli containing both female and male utterances were randomly chosen from both LibriSpeech and VCTK datasets to evaluate speech intelligibility and naturalness (three stimuli from each method). The two stimuli used to compare speaker dissimilarity consisted of 36 pairs (twelve of the same stimuli from each method). These twelve pairs were randomly chosen from the development and test data of the LibriSpeech and VCTK datasets. There was an equal distribution of female and male utterances.

This test was conducted in a standard soundproof room equipped with a computer, an audio interface (Roland OCTA-CAPTURE), and headphones (SENNHEISER HDA 200) to avoid environmental bias. The sound pressure level of the background noise in the room was lower than 28 dB. We also randomized the order of the stimuli and normalized all the sound data of the listening test at the same sound pressure level of −20 decibels relative to full scale (dBFS) and sampled at 16 kHz. Before the experiment, we explained the test to each participant and instructed them to ensure they understood the test and the metrics. During the test, each stimulus was played only once to prevent bias.

The overall results of our subjective evaluation are shown in Fig. 10. The figure shows that the P1 method performed similarly to the B1 method with all the metrics, while the P2 method performed significantly better than the B1 and P1 methods regarding dissimilarity. An apparent limitation of the P2 method is the slight reduction in the intelligibility and naturalness metrics. To verify

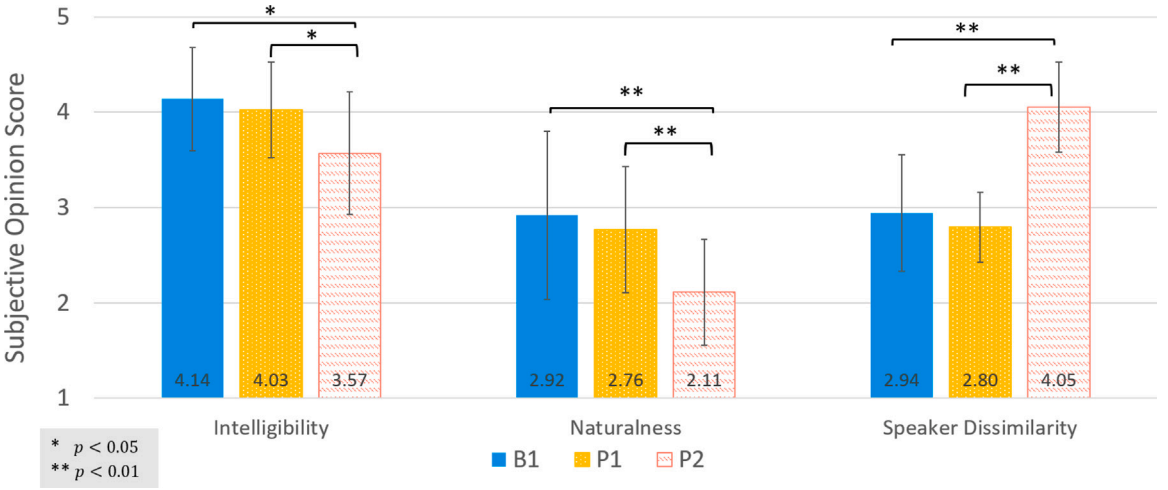

Fig. 10. Overall subjective evaluation results in terms of intelligibility, naturalness, and speaker dissimilarity.

this result, we conducted a single-factor analysis of variance (ANOVA) test using the mean of the total stimuli per method of the subject's rating score, i.e., the mean rating scores from three stimuli for the intelligibility & naturalness and twelve stimuli for the speaker dissimilarity.

The results of the ANOVA test showed significant differences between the three methods (B1, P1, and P2) in speech intelligibility ($F(2, 69) = 3.90, p < 0.05$), naturalness ($F(2, 69) = 6.28, p < 0.01$), and speaker dissimilarity ($F(2, 69) = 23.47, p < 0.01$) between the three compared methods (B1, P1, and P2). Subsequently, we conducted a post-hoc Tukey honestly significant difference test to determine the differences between the two methods for each metric. The results indicated that there is a statistically significant difference between the B1 and the P2 ($p < 0.05$ for speech intelligibility, $p < 0.01$ for naturalness, and $p < 0.01$ for speaker dissimilarity). Similarly, the difference between the P1 and the P2 is also significant ($p < 0.05$ for speaker intelligibility, $p < 0.01$ for naturalness, and $p < 0.01$ for speaker dissimilarity). Meanwhile, there is no significant difference between the B1 and the P1 ($p > 0.05$ for speaker intelligibility, naturalness, and speaker dissimilarity).

Fig. 11 shows the speaker dissimilarity test distribution regarding datasets and gender. The top two figures show the speaker dissimilarity results of the three methods from the LibriSpeech dataset (left) and the VCTK dataset (right). The bottom two figures show the speaker dissimilarity results of the three methods from female utterances (left) and male utterances (right). These figures are consistent with the overall results in Fig. 8 that denote how the results of the P1 are relatively similar to the B1, whereas speaker dissimilarity significantly improved using the P2 method.

A demo page of the output anonymized speech of this system is available publicly as a reference.[2]

5. Discussion

We proposed techniques to improve the primary baseline system introduced in the VoicePrivacy Challenge 2020. An ablation test was conducted to determine the effectiveness of each method and its combinations. The detailed results were excluded to condense the excess results obtained from our experiments. Based on our experimental results shown in Section 4.3, in this section, we discuss the effectiveness of each technique in our methods, evaluation design & metrics for speaker anonymization, and we discuss the limitations in the current methods and evaluation protocols.

Four key questions about our current study are the following:

1. **How effective is modifying x-vector SV for speaker anonymization with a k-means clustering model?**
   In this study, we conducted experiments using several SV thresholds for anonymizing x-vectors. The average results are shown in Fig. 8 for speaker verifiability and in Fig. 9 for speech intelligibility. These results suggest that anonymization by modifying x-vector SV could achieve a performance comparable to the primary baseline. Speaker verifiability slightly improved when the threshold of the SV was reduced to 0.8, especially in the a-a scenario. Unfortunately, this improvement significantly distorted speech intelligibility for the LibriSpeech dataset.
   The trade-off between verifiability and intelligibility occurred from using SVD. The SVD technique is used to capture the intra-speaker characteristics. Consequently, the optimal SV threshold for the LibriSpeech dataset is higher than for the VCTK dataset because the VCTK dataset contains more variation (in accents, etc.) than the LibriSpeech dataset. To control the trade-off between speaker verifiability and speech intelligibility, we selected the best threshold for each dataset.

[2] http://www.jaist.ac.jp/~s1920436/anon/demo.html

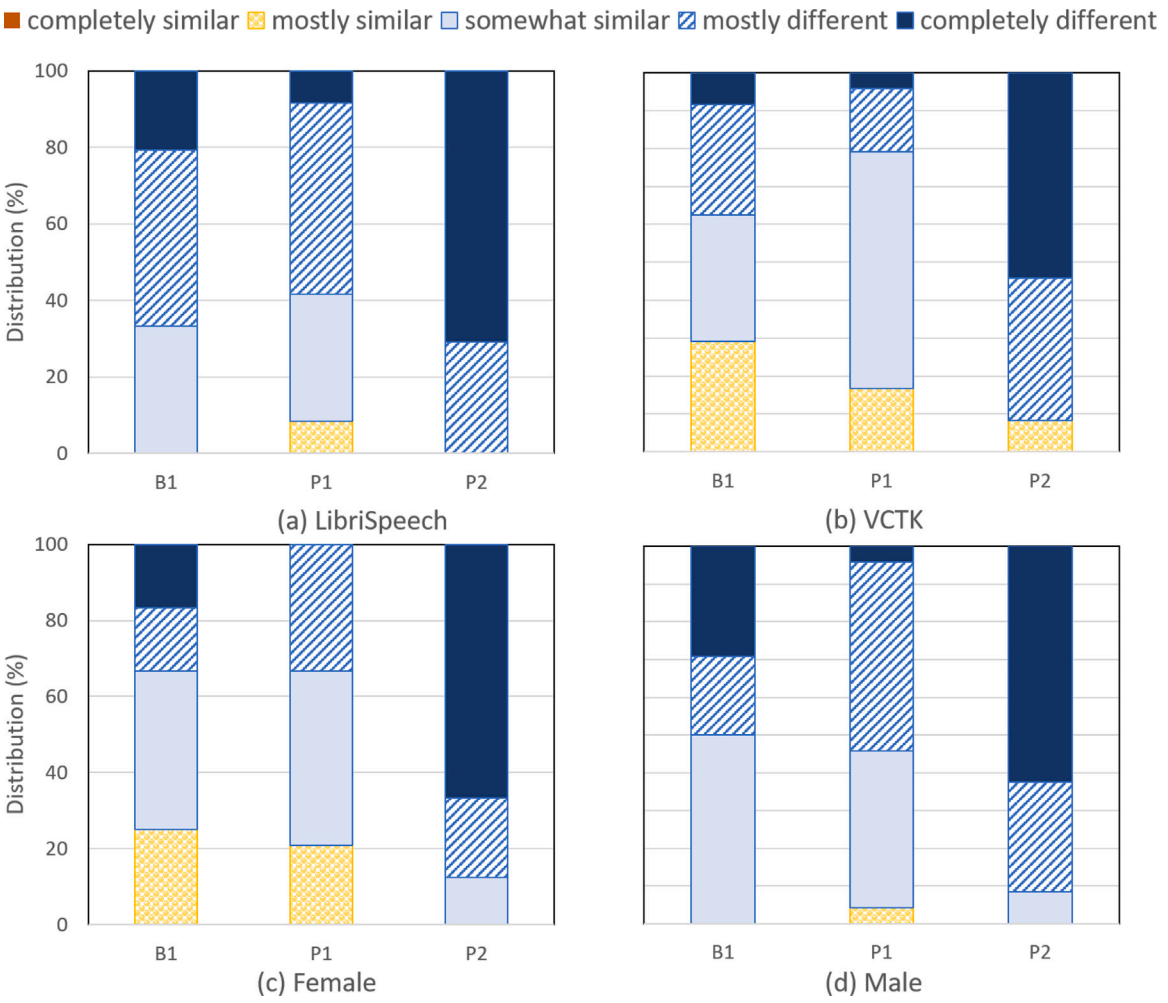

Fig. 11. Subjective evaluation results of speaker dissimilarity in utterances from (a) LibriSpeech dataset, (b) VCTK dataset, (c) female speakers, and (d) male speakers.

Table 3
System description of related system anonymization methods.

System	Description
B1 (Tomashenko et al., 2020b)	Primary baseline (x-vector anonymization)
B2 (Tomashenko et al., 2020b)	Secondary baseline (McAdams coefficient modification)
P1	Proposed method 1 (B1 using SVD and k-means clustering)
P2	Proposed method 2 (P1 + modification on F_0 and speech duration)
O1 (Turner et al., 2020)	B1 using cosine distances, GMM for sampling vectors in a PCA-reduced space
S2 (Espinoza-Cuadros et al., 2020)	B1 using doman-adversarial training autoencoders
K2 (Han et al., 2020)	Anonymization using x-vectors and SS models, voice-indistinguishability metric, Griffin–Lim algorithm based waveform vocoder
I1 (Dubagunta et al., 2020)	Modification on formants, F_0, and speaking rate
C1 (Champion et al., 2021)	B1 + F_0 modification

In our methods, we used the k-means clustering model to choose the pseudo-target x-vector. This differs from our prior work Mawalim et al. (2020) (labeled as "A2" in Figs. 12 and 14) in which we constructed a regression model that chose the pseudo-target x-vector. The contribution of this clustering model is that it improves the speaker verifiability in the a-a scenario without degrading the intelligibility achieved by our prior work (compare A2 and P1 in Fig. 12).

2. **How effective is modifying speech prosody, including the F_0 and speech duration?**
The evaluation results show the effectiveness of modifying the F_0 and speech duration for speaker anonymization. The objective evaluation results of P1 and P2 (shown in Fig. 8) show a slightly better performance of P2 in speaker verifiability,

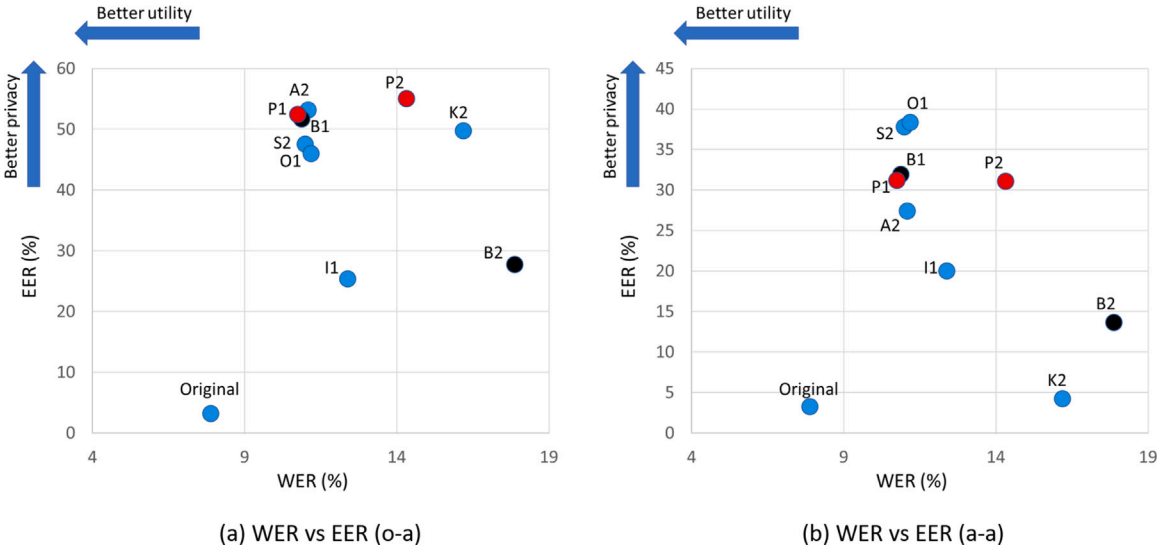

Fig. 12. Mean WER versus mean EER over all LibriSpeech and VCTK datasets in (o-a) and (a-a) scenarios obtained from various systems proposed in VoicePrivacy Challenge 2020. Black dot refers to results obtained by baseline system. Red dot refers to results obtained by our proposed system. Blue dot refers to results obtained by other systems proposed in VoicePrivacy Challenge 2020. Table 3 describes each system. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

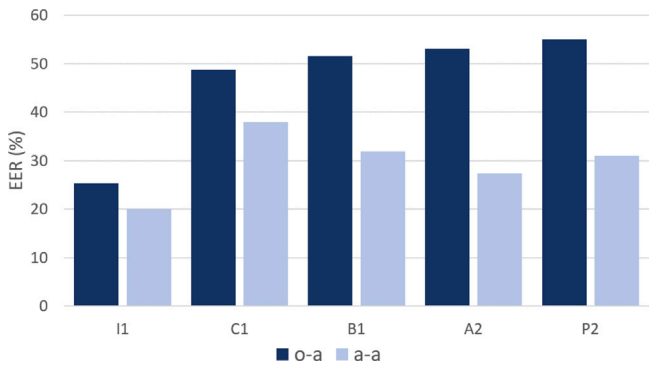

Fig. 13. Mean EER values over LibriSpeech (test set) in (o-a) and (a-a) scenarios obtained by systems related to modifying speech prosody. Table 3 describes each method.

especially with the LibriSpeech dataset and female utterances. Contrary to the results of ASVeval, P2 caused slightly reduced speech intelligibility. Additionally, the subjective evaluation results of P1 and P2 showed more significant differences in speaker verifiability. Unfortunately, P2 caused slightly more perceivable distortion than P1 in terms of utility (intelligibility and naturalness).

To further investigate the effect of speech prosody modification, we conducted comparative analysis of related speaker anonymization systems. Fig. 13 shows the ASVeval results of five systems labeled I1 Dubagunta et al. (2020), C1 Champion et al. (2021), B1, A2 Mawalim et al. (2020), and P2 (detailed in Table 3). We could not obtain full results for all the systems; therefore, the results shown are only for LibriSpeech (test data). The results of I1 (Dubagunta et al., 2020) in Figs. 13 and 14 show that modifying formants, the F_0, and the speaking rate alone improves the anonymization performance (compared with the original in Fig. 14). It supported the previous studies Akagi and Ienaga (1997) and Dellwo et al. (2007) that described the strong relationship between speaker individuality and the F_0 & speech duration. Unfortunately, modifying the speech prosody degraded the intelligibility.

Combining the speech prosody modification with the main framework in the primary baseline (systems C1, A2, and P2) improves speaker anonymization. A study by Champion et al. (2021) investigated the effect of F_0 modification in the primary baseline system across gender. Our current work did not focus on cross-gender modification, so we only compared their results obtained from same-gender modification labeled C1. Although the C1 reduced the speaker verifiability performance in the

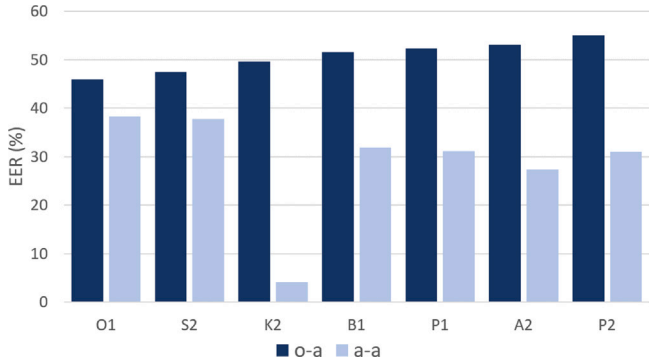

Fig. 14. Mean EER values over all LibriSpeech and VCTK datasets in (o-a) and (a-a) scenarios obtained by systems related to x-vector anonymization. Table 3 describes each method.

o-a scenario compared with B1, it significantly improved the performance in the a-a scenario. The ASReval results showed a similar tendency to other speech prosody modification systems in the slight reduction in WER that occurred (Champion et al., 2021). Due to incomplete results, the mapping of WER and EER results across all datasets for the C1 could not be included in Fig. 12.

In contrast to the C1, we carried out experiments using WORLD for the F_0 modification (systems A2 and P2). In A2 (our prior work Mawalim et al. (2020)), we used the estimated F_0 obtained from the SWIPE algorithm in WORLD in the primary baseline framework with SVD-based modification. Overall, the A2 performed slightly better than the B1 in the o-a scenario but not as well in the a-a scenario. A more significant improvement could be achieved by modifying the estimated F_0 and speech duration (method P2), especially in the o-a scenario (as shown in Fig. 13). Despite improving privacy, we could see drawbacks similar to those that occurred due to the speech prosody modification in Fig. 12.

In summary, all systems related to modifying speech prosody can significantly improve privacy, but they also slightly degrade utility. In addition, the P2 performed better in ASVeval for female utterances than for male utterances. We predict that these results occur due to the different F_0 range between female and male speakers (Traunmüller and Eriksson, 1995) and the linear transformation utilized. An effective transformation of the F_0 for anonymization is gender-dependent because female speakers have a wider F_0 range than male speakers.

3. **How does the performance of the proposed method compare with existing speaker anonymization systems?**
   Fig. 12 shows the scatter plot of the mean WER and mean EER results of several existing systems in both scenarios (o-a and a-a). The mean WER and mean EER were calculated by averaging the results obtained from all the LibriSpeech and VCTK datasets (development and test sets). In the VoicePrivacy Challenge 2020 results, one criteria for determining better privacy is a higher EER. Subsequently, better utility is determined by having a lower WER. Although solely using these two metrics might be considered an oversimplified evaluation, the results in Fig. 12 gives insight into comparing the existing systems. For example, one of the overall conclusions provided in the VoicePrivacy Challenge 2020 results indicated that the systems based on x-vector anonymization (B1, P1, P2, O1 Turner et al. (2020), S2 Espinoza-Cuadros et al. (2020), and K2 Han et al. (2020)) could perform better than the ones based on signal-processing methods (B2 and I1 Dubagunta et al. (2020)).
   The overall results show that some systems (A2, P1, O1, and S2) have nearly similar results as the primary baseline system (B1), especially in the o-a scenario. Our prior work (A2) was less effective than B1 in the a-a scenario. However, the results of other participants that also used the primary baseline framework (O1 and S2) showed improved privacy in the a-a scenario (mean EER increased around 5%), but the mean EER reduced around 5% to 10% for the o-a scenario. Subsequently, our P2 method, slightly improved in the o-a scenario but had a result similar to B1 regarding privacy. Unfortunately, it increased the mean WER by approximately 3.5% in comparison with B1. The B2, I1, and K2 systems not based on B1 were less effective than B1 in both privacy and utility. Fig. 14 compares ASVeval results of anonymization systems that use x-vectors. Although there is a relatively slight degradation in intelligibility, our P2 method achieved the highest mean EER in the o-a scenario.
4. **How reliable and significant is the subjective evaluation for the speaker anonymization system?**
   There are limitations to the existing objective evaluation. One such limitation is that speaker verifiability is evaluated using x-vector embedding (Snyder et al., 2018) in the ASVeval. Although this system performed the best in terms of the EER for the VoxCeleb dataset, Hautamäki and Kinnunen's study (Hautamäki and Kinnunen, 2020) indicates that an x-vector embedding system that only uses the Mel-spectrogram as its input is not robust with intra-speaker variations. Reportedly, this primarily results from the mismatch between the F_0 mean and the associated formant frequencies. Consequently, a better objective evaluation for assessing the verifiability of a speaker anonymization system must be considered because it should distinguish the uniqueness of anonymized speech for each speaker. Another limitation is that it is quite difficult to decide which system is the most effective in all cases regardless of the datasets and genders using only the objective evaluation results (such as the results shown in Figs. 8 and 9).

Accordingly, we proposed a subjective evaluation that can more reliably determine the effectiveness of different speaker anonymization systems. This subjective evaluation differs from the one introduced in the VoicePrivacy Challenge 2020 (Tomashenko et al., 2020b). Even though native speakers have a better understanding of their mother language, we considered the difficulty in gathering an adequate number of native speakers as suggested for the challenge's subjective evaluation. Collecting evaluation results via the internet could be an alternative solution to deal with this problem; however, there will be biases due to different environments, equipment, etc. Furthermore, instead of using a 10-point scale opinion score, we used a 5-point scale based on psychological studies that suggested higher reliability regarding the response rate and quality (Buttle, 1996; Eli and Cox, 1980). The 10-point scale opinion score is too difficult and could increase the "frustration level" even for a native speaker (Buttle, 1996). Furthermore, Eli P. Cox's 1980 study on the optimal number of alternatives for a scale suggested that an odd number of alternatives is preferable to enable a neutral response (Eli and Cox, 1980).

To improve the reliability of our subjective evaluation, we anticipated bias based on environment, equipment, understanding, and/or human perceptual phenomena in the hearing system. We also compensated for any potential misunderstandings from non-native speakers by using pair-comparison even though we verified the participants' English skills (detailed in Section 4.3.2). We also conducted ANOVA tests to analyze significant differences in the systems.

Figs. 10 and 11 show our subjective evaluation results, comparing B1 with our P1 and P2 systems. Unfortunately, the results obtained using the P1 method are not significantly different from the results obtained using the B1 method. However, combining all the techniques proposed in this study (P2) could improve the speaker dissimilarity significantly compared with the B1 method regardless of the dataset and gender.

There are several limitations related to existing techniques and evaluations. Per the summary in the VoicePrivacy Challenge 2020, x-vector-anonymization methods could be more effective than signal-processing methods proposed by other participants. However, we predicted a potential bias from the use of x-vectors in the objective evaluation by the ASV system. Even if there is a slightly better performance using any x-vector modification technique, the output is greatly affected by the NSF model. We determined this through our subjective evaluation, which showed very similar results between the B1 and the P1.

Regarding evaluation limitations, it could be argued that the current evaluations remain insufficient for assessing a speaker anonymization system. Although the metrics used in the current evaluations could give useful information, there are many critical points that are not captured by those metrics. Thus, it is quite difficult to conduct comparative studies solely using those metrics because the results are inconsistent. For instance, the objective evaluation results obtained by a system proposed in Han et al. (2020) are the opposite of a subjective evaluation in terms of privacy metrics. Furthermore, the difficulty in assessing the quality of a speaker anonymization system should be considered. For example, the degradation of anonymized speech quality could cause improve the EER despite the poor performance of the anonymization task. Therefore, we attempted to provide a considerably more consistent and reliable subjective evaluation. However, there are limitations, especially regarding the attack models.

6. Conclusion and future work

We proposed speaker anonymization methods that modify the F_0, speech duration, and x-vectors SV with a k-means clustering model. The SV of the x-vector modification was performed to fulfill the speaker-to-speaker correspondence requirement of an anonymization system. The experiments were conducted with various SV thresholds to investigate its effectiveness depending on the dataset. The F_0 and speech duration were modified based on their high correlation with speaker individuality. To evaluate our methods, we followed the objective evaluation suggested in the VoicePrivacy Challenge 2020 and conducted a more reliable subjective evaluation to compensate for limitations of the objective evaluation. The objective evaluation results demonstrated that combining all the sub-components of the P2 method improves speaker verifiability compared with the B1, especially for the LibriSpeech dataset and female utterances (up to 5%). The subjective evaluation results also showed the P2 effectively improved speaker dissimilarity, with more than 90% of respondents indicating that output sounded mostly different regardless of the dataset and gender. However, due to the manipulation of speech prosody, the P2 slightly reduced the intelligibility and naturalness compared with the results obtained by the B1.

As future work, we will improve our method by studying how to effectively modify the F_0 and other features related to speaker individuality. We intend to study an analysis-synthesis model that can better conceal personally identifiable information because it greatly affects anonymization performance. Furthermore, we plan to improve our evaluation methods, especially the subjective evaluation, considering the attack models described in the VoicePrivacy Challenge 2020 (Tomashenko et al., 2020b).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by a Grant-in-Aid for Scientific Research (B), Japan (No. 17H01761) and JSPS, Japan KAKENHI Grant (No. 20J20580). This work was also supported by Fund for the Promotion of Joint International Research (Fostering Joint International Research (B)), Japan (20KK0233), JSPS-NSFC Bilateral Joint Research Projects/Seminars, Japan (JSJSBP120197416), and KDDI foundation, Japan (Research Grant Program).

References

Abou-Zleikha, M., Tan, Z.-H., Christensen, M.G., Jensen, S.H., 2015. A discriminative approach for speaker selection in speaker de-identification systems. In: 23rd European Signal Processing Conference, EUSIPCO 2015. IEEE, Nice, France, pp. 2102–2106. http://dx.doi.org/10.1109/EUSIPCO.2015.7362755.

Akagi, M., Ienaga, T., 1997. Speaker individuality in fundamental frequency contours and its control. J. Acoust. Soc. Japan E 18 (2), 73–80. http://dx.doi.org/10.1250/ast.18.73.

Bottou, L., Bengio, Y., 1994. Convergence properties of the k-means algorithms. In: Advances in Neural Information Processing Systems 7, NIPS Conference, Denver, Colorado, USA, 1994. MIT Press, pp. 585–592.

Brümmer, N., du Preez, J.A., 2006. Application-independent evaluation of speaker detection. Comput. Speech Lang. 20 (2–3), 230–275. http://dx.doi.org/10.1016/j.csl.2005.08.001.

Buttle, F., 1996. SERVQUAL: Review, critique, research agenda. Comput. Speech Lang. 30, 8–32. http://dx.doi.org/10.1108/03090569610105762.

Champion, P., Jouvet, D., Larcher, A., A Study of F0 modification for X-vector based speech pseudonymization across gender. In: PPAI 2021 - the Second AAAI Workshop on Privacy-Preserving Artificial Intelligence, Virtual, China. URL: https://hal.archives-ouvertes.fr/hal-02995862.

Chung, J.S., Nagrani, A., Zisserman, A., 2018. VoxCeleb2: Deep speaker recognition. In: Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018. ISCA, pp. 1086–1090. http://dx.doi.org/10.21437/Interspeech.2018-1929.

Das, R.K., Prasanna, S.M., 2018. Speaker verification from short utterance perspective: A review. IETE Tech. Rev. 35 (6), 599–617.

Das, R.K., Sharma, B., Prasanna, S.R.M., 2018. Significance of duration modification for speaker verification under mismatch speech tempo condition. Int. J. Speech Technol. 21 (3), 401–408. http://dx.doi.org/10.1007/s10772-017-9474-5.

Das, R.K., Tian, X., Kinnunen, T., Li, H., 2020. The attacker's perspective on automatic speaker verification: An overview. In: Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China. ISCA, pp. 4213–4217. http://dx.doi.org/10.21437/Interspeech.2020-1052.

Dellwo, V., Huckvale, M.A., Ashby, M., 2007. How is individuality expressed in voice? An introduction to speech production and description for speaker classification. In: Speaker Classification I: Fundamentals, Features, and Methods. In: Lecture Notes in Computer Science, vol. 4343, Springer, pp. 1–20. http://dx.doi.org/10.1007/978-3-540-74200-5_1.

Dubagunta, S.P., van Son, R.J.J.H., Magimai, M., 2020. Adjustable deterministic pseudonymisation of speech: Idiap-NKI's submission to VoicePrivacy 2020 Challenge.

Eli, P., Cox, I., 1980. The optimal number of response alternatives for a scale: A review. J. Mar. Res. 17 (4), 407–422. http://dx.doi.org/10.1177/002224378001700401.

Espinoza-Cuadros, F.M., Perero-Codosero, J.M., Antón-Martín, J., Gómez, L.A.H., 2020. Speaker de-identification system using autoencoders and adversarial training. CoRR arXiv:2011.04696.

Fang, F., Wang, X., Yamagishi, J., Echizen, I., Todisco, M., Evans, N.W.D., Bonastre, J.-F., 2019. Speaker anonymization using x-vector and neural waveform models. CoRR arXiv:1905.13561.

Fang, F., Yamagishi, J., Echizen, I., Lorenzo-Trueba, J., 2018. 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, pp. 5279–5283. http://dx.doi.org/10.1109/ICASSP.2018.8462342.

Golub, G.H., Reinsch, C., 1970. Singular value decomposition and least squares solutions. Numer. Math. 14 (5), 403–420. http://dx.doi.org/10.1007/BF02163027.

Goodfellow, I.J., Bengio, Y., Courville, A.C., 2016. Deep Learning. In: Adaptive computation and machine learning, MIT Press.

Gussenhoven, C., Repp, B., Rietveld, A., Rump, H., Terken, J., 1997. The perceptual prominence of fundamental frequency peaks. J. Acoust. Soc. Am. 102, 3009–3022. http://dx.doi.org/10.1121/1.420355.

Han, Y., Cao, Y., Li, S., Ma, Q., Yoshikawa, M., 2020. Voice-indistinguishability - protecting voiceprint with differential privacy under an untrusted server. In: CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA. ACM, pp. 2125–2127. http://dx.doi.org/10.1145/3372297.3420025.

Hautamäki, R.G., Kinnunen, T., 2020. Why did the x-vector system miss a target speaker? Impact of acoustic mismatch upon target score on voxceleb data. In: Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China. ISCA, pp. 4313–4317. http://dx.doi.org/10.21437/Interspeech.2020-2715.

Hirose, K., Kawanami, H., 2002. Temporal rate change of dialogue speech in prosodic units as compared to read speech. Speech Commun. 36 (1–2), 97–111. http://dx.doi.org/10.1016/S0167-6393(01)00028-0.

Irum, A., Salman, A., 2019. Speaker verification using deep neural networks: A review. Int. J. Mach. Learn. Comput. 9 (1).

Jin, Q., Toth, A.R., Black, A.W., Schultz, T., 2008. 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, pp. 4845–4848. http://dx.doi.org/10.1109/ICASSP.2008.4518742.

Jin, Q., Toth, A.R., Schultz, T., Black, A.W., 2009. 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, pp. 529–533. http://dx.doi.org/10.1109/ASRU.2009.5373356.

Magariños, C., Lopez-Otero, P., Fernández, L.D., Banga, E.R., Erro, D., García-Mateo, C., 2017. Reversible speaker de-identification using pre-trained transformation functions. Comput. Speech Lang. 46, 36–52. http://dx.doi.org/10.1016/j.csl.2017.05.001.

Mawalim, C.O., Galajit, K., Karnjana, J., Unoki, M., 2020. X-vector singular value modification and statistical-based decomposition with ensemble regression modeling for speaker anonymization system. In: Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China. ISCA, pp. 1703–1707. http://dx.doi.org/10.21437/Interspeech.2020-1887.

McAdams, S., 1984. Spectral fusion, spectral parsing and the formation of auditory images (Ph.D. thesis). Stanford.

Morise, M., Yokomori, F., Ozawa, K., 2016. WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. IEICE Trans. Inf. Syst. E99.D, 1877–1884. http://dx.doi.org/10.1587/transinf.2015EDP7457.

Nagrani, A., Chung, J.S., Zisserman, A., 2017. 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, pp. 2616–2620.

Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015. Librispeech: An ASR corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia. IEEE, pp. 5206–5210. http://dx.doi.org/10.1109/ICASSP.2015.7178964.

Patino, J., Tomashenko, N.A., Todisco, M., Nautsch, A., Evans, N.W.D., 2020. Speaker anonymisation using the mcadams coefficient. CoRR arXiv:2011.01130.

Peddinti, V., Povey, D., Khudanpur, S., 2015. 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, pp. 3214–3218.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Pobar, M., Ipsic, I., 2014. Online speaker de-identification using voice transformation. In: 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2014, Opatija, Croatia. IEEE, pp. 1264–1267. http://dx.doi.org/10.1109/MIPRO.2014.6859761.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K., 2011. The kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Hawaii, US. IEEE Signal Processing Society.

Sahidullah, M., Kinnunen, T., 2016. Local spectral variability features for speaker verification. Digit. Signal Process. 50, 1–11. http://dx.doi.org/10.1016/j.dsp.2015.10.011.

Sathiyamurthi, P., Ramakrishnan, S., 2017. Speech encryption using chaotic shift keying for secured speech communication. EURASIP J. Audio Speech Music Process. 2017, 20. http://dx.doi.org/10.1186/s13636-017-0118-0.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S., 2018. X-vectors: Robust dnn embeddings for speaker recognition.

Tomashenko, N.A., Srivastava, B.M.L., Wang, X., Vincent, E., Nautsch, A., Yamagishi, J., Evans, N.W.D., Patino, J., Bonastre, J.-F., Noé, P.-G., Todisco, M., 2020a. Introducing the voiceprivacy initiative. In: Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China. ISCA, pp. 1693–1697. http://dx.doi.org/10.21437/Interspeech.2020-1333.

Tomashenko, N., Srivastava, B.M.L., Wang, X., Vincent, E., Nautsch, A., Yamagishi, J., Evans, N., Patino, J., Bonastre, J.-F., Noé, P.-G., Todisco, M., 2020b. The voiceprivacy 2020 challenge evaluation plan. URL: https://www.voiceprivacychallenge.org/docs/VoicePrivacy_2020_Eval_Plan_v1_3.pdf. Visited on 2021-06-10.

Traunmüller, H., Eriksson, A., 1995. The frequency range of the voice fundamental in the speech of male and female adults.

Turner, H., Lovisotto, G., Martinovic, I., 2020. Speaker anonymization with distribution-preserving x-vector generation for the voiceprivacy challenge 2020. CoRR arXiv:2010.13457.

Veaux, C., Yamagishi, J., Macdonald, K., 2017. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit.

Vestman, V., Kinnunen, T., Hautamäki, R.G., Sahidullah, M., 2020. Voice mimicry attacks assisted by automatic speaker verification. Comput. Speech Lang. 59, 36–54. http://dx.doi.org/10.1016/j.csl.2019.05.005.

Wang, X., Takaki, S., Yamagishi, J., 2019. 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, pp. 5916–5920. http://dx.doi.org/10.1109/ICASSP.2019.8682298.

Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R.J., Jia, Y., Chen, Z., Wu, Y., 2019. 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, pp. 1526–1530. http://dx.doi.org/10.21437/Interspeech.2019-2441.