






Speaker Anonymization Using Orthogonal Householder Neural Network

Xiaoxiao Miao , Member, IEEE, Xin Wang , Member, IEEE, Erica Cooper , Member, IEEE, Junichi Yamagishi , Senior Member, IEEE, and Natalia Tomashenko , Member, IEEE

Abstract—Speaker anonymization aims to conceal a speaker’s identity while preserving content information in speech. Current mainstream neural-network speaker anonymization systems disentangle speech into prosody-related, content, and speaker representations. The speaker representation is then anonymized by a selection-based speaker anonymizer that uses a mean vector over a set of randomly selected speaker vectors from an external pool of English speakers. However, the resulting anonymized vectors are subject to severe privacy leakage against powerful attackers, reduction in speaker diversity, and language mismatch problems for unseen-language speaker anonymization. To generate diverse, language-neutral speaker vectors, this article proposes an anonymizer based on an orthogonal Householder neural network (OHNN). Specifically, the OHNN acts like a rotation to transform the original speaker vectors into anonymized speaker vectors, which are constrained to follow the distribution over the original speaker vector space. A basic classification loss is introduced to ensure that anonymized speaker vectors from different speakers have unique speaker identities. To further protect speaker identities, an improved classification loss and similarity loss are used to push original-anonymized sample pairs away from each other. Experiments on VoicePrivacy Challenge datasets in English and the AISHELL-3 dataset in Mandarin demonstrate the proposed anonymizer’s effectiveness.

Index Terms—Speaker anonymization, selection-based anonymizer, orthogonal Householder neural network anonymizer, weighted additive angular softmax.

I. INTRODUCTION

SPEECH technology enables machines to recognize, analyze, and understand human speech, which facilitates human-machine communication and offers great convenience in our daily lives. Despite its prominent advantages, it suffers from voice privacy leakage, which allows for intrusion upon or tampering with a speaker’s private information. For instance, by

Manuscript received 3 February 2023; revised 13 July 2023; accepted 30 August 2023. Date of publication 8 September 2023; date of current version 20 October 2023. This work was supported in part by JST CREST under Grants JPMJCR18A6 and JPMJCR20D3, in part by MEXT KAKENHI under Grants 21K17775, 21H04906, 21K11951, and 22K21319, and in part by the VoicePersonal Project under Grant ANR-18-JSTS-0001. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zhizheng Wu. (Corresponding author: Xiaoxiao Miao.)

Xiaoxiao Miao, Xin Wang, Erica Cooper, and Junichi Yamagishi are with the National Institute of Informatics, Tokyo 101-8340, Japan (e-mail: xiaoxiao-miao@nii.ac.jp; wangxin@nii.ac.jp; ecooper@nii.ac.jp; jyamagis@nii.ac.jp).

Natalia Tomashenko is with the Laboratoire Informatique d’Avignon (LIA), Avignon University, 84911 Avignon, France (e-mail: natalia.tomashenko@univ-avignon.fr).

Digital Object Identifier 10.1109/TASLP.2023.3313429

using advanced speaker [1], [2], dialect [3], [4], pathological condition [5], [6], or other types of speech attribute recognition systems, attributes such as a speaker’s identity, geographical origin, and health status can easily be captured from speech recordings. Moreover, advanced speech synthesis techniques enable resynthesis, cloning, or conversion of a speaker’s identity information to access personal voice-controlled devices [7], [8], [9]. In this article, we are especially interested in speaker anonymization, which is a user-centric voice privacy solution to conceal a speaker’s identity without degrading intelligibility and naturalness [10], [11], [12]. This task was standardized by the VoicePrivacy Challenge (VPC) committee [11], [12], [13], which held challenges in 2020 and 2022, to advance the development of voice privacy preservation techniques.

Several approaches to protect speaker privacy are based on digital signal processing (DSP) methods [11], [12], [14], [15], [16], [17], [18], which modify instantaneous speech characteristics such as the pitch, spectral envelope, and time scaling. State-of-the-art anonymization approaches have borrowed ideas from neural speech conversion and synthesis, mainly focusing on disentangled latent representation learning [10], [19], [20], [21], [22], [23], [24], [25] via two hypotheses. The first is that speech can be explicitly decomposed into content, speaker identity, and prosodic (intonation, stress, and rhythm) representations. Here, the speaker identity is a statistical time-invariant representation throughout an utterance, whereas content and prosodic information vary over time. The second hypothesis is that a speaker’s identity representation carries most of his or her private information. Thus, generated speech using original content, prosodic, and anonymized speaker representations can suppress the original identity information (privacy) while maintaining intelligibility and naturalness (utility).

A general framework for disentanglement-based speaker anonymization involves the following components.

Fine-grained disentangled representation extraction from original speech: Here, extraction entails three aspects: (i) Content feature extraction. Low-dimensional phonetic bottleneck features are typically extracted from an intermediate layer of a language-specific automatic speech recognition neural acoustic model (ASR AM) [26], [27]. This type of content encoder is trained in a supervised manner using transcribed English training data. As the objective is to obtain accurate linguistic representations, the effectiveness is severely limited when applied to a different language. Content encoders based on self-supervised learning (SSL) can overcome this limitation thanks

to being trained in a self-supervised manner using unlabeled training data. Specifically, they can provide general content representations not dependent on the language, thus enabling robust anonymization of speech data even for unseen languages. (ii) Prosody-related feature extraction to obtain the fundamental frequency, i.e., F0. (iii) Speaker embedding extraction. A speaker vector is extracted either from an automatic speaker verification (ASV) system based on a time-delay neural network (TDNN) [28], or from a more effective ASV system based on emphasized channel attention, propagation and aggregation in TDNN (ECAPA-TDNN) [29].

Speaker representation anonymization: The core idea of a speaker vector anonymizer is to hide original speaker information while preserving the diversity among different speakers. A widely used anonymizer is based on the selection and averaging of speaker vectors [23], [30]. Given a large set of speaker vectors, the anonymizer finds the N farthest candidate vectors away from an input original vector. It then randomly selects $N^* < N$ vectors among the N farthest ones and utilizes their average as a pseudo-speaker vector to replace the original speaker vector. The large set of speaker vectors, called an external pool, has to be loaded by the anonymizer during anonymization.

Anonymized speech synthesis: An anonymized speaker vector with the original fundamental frequency and content features is passed to a speech waveform generation model to synthesize high-quality anonymized speech. The speech synthesis model can be a traditional text-to-speech pipeline model—a speech synthesis acoustic model (SS AM) and a neural source filter-(NSF-) based vocoder [31]—or a unified HiFi-GAN [32].

Despite confirmation of this approach’s effectiveness [11], [12], [33], there remains much room for improvement for different attack scenarios and unseen language anonymization. Previous works [11], [12], [23], [33] have suggested that the most significant performance bottleneck for the current mainstream approach is the selection-based speaker anonymizer, whose performance significantly depends on the distribution of the external pool and how pseudo-speakers are selected from the pool. (i) For English speaker anonymization [11], [12], [13], the performance of speaker verifiability has gradually decreased against more powerful attackers. Additionally, voice distinctiveness is significantly degraded by anonymization. (ii) For unseen-language (e.g., Mandarin) speaker anonymization, pseudo-speaker representations are generated from an external English speaker vector pool, and the resulting language mismatch increases the character error rate (CER) [33], [34].

Following this pipeline of disentanglement-based anonymization, with special consideration of the selection-based approach’s problems, we propose a novel speaker anonymization system (SAS) based on an orthogonal Householder neural network (OHNN). As shown in the lower part of Fig. 1, the OHNN-based anonymizer generates distinctive anonymized speaker vectors that can protect privacy under all attack scenarios and can successfully be adapted to unseen-language speaker anonymization without severe language mismatch. Specifically, original speaker vectors are rotated to anonymized ones by an OHNN, which is a linear transformation with orthogonality. This module ensures that the anonymized speaker

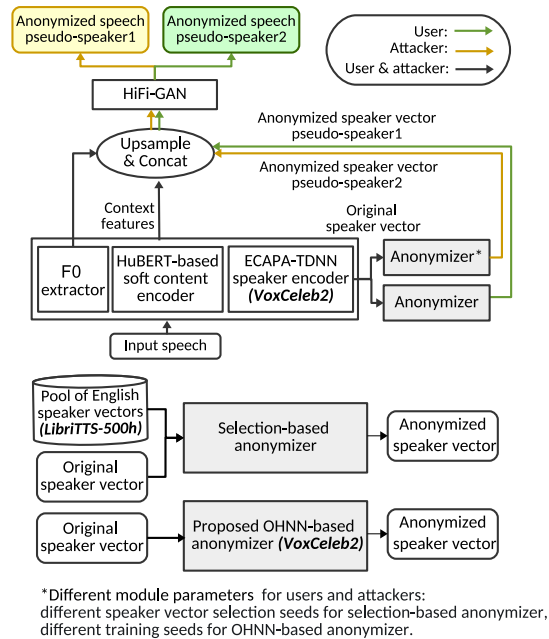


Fig. 1. Architecture of an SSL-based SAS, with selection- and OHNN-based anonymizers.

vectors follow the distribution over the original speaker vector space. To discourage overlap between anonymized speakers and other speakers, we use a classification loss based on an additive angular margin softmax (AAM) and cross-entropy to train the OHNN, and we assign different target class labels to the original and anonymized speaker vectors of different speakers. This encourages the anonymized vectors to not overlap with any other speakers, regardless of whether they are original or anonymized. To further push original-anonymized sample pairs away from each other, an improved classification loss called weighted AAM (w-AAM) and a cosine similarity loss are used.

The main contributions of this work are as follows:

- We propose an OHNN-based anonymizer that transforms original speaker vectors into anonymized ones with carefully designed training constraints. We show empirically that these anonymized speaker vectors are diverse and language-neutral.
- We visualize the cosine similarities between pairs of speaker vectors extracted from the generated speech of users and different attackers. These generated speech are obtained using the commonly used selection-based anonymizer and our OHNN-based anonymizer. The results show that our proposed method effectively reduces the privacy leakage against different attackers and improves the diversity of anonymized speakers. We conducted experiments on VPC English datasets and the AISHELL-3 Mandarin datasets. Our findings show that the proposed model can be successfully adapted to both a matched language condition (i.e., English) and a mismatched language condition where the target language (Mandarin) is not included in the training database. The proposed anonymizer achieved a competitive performance under all

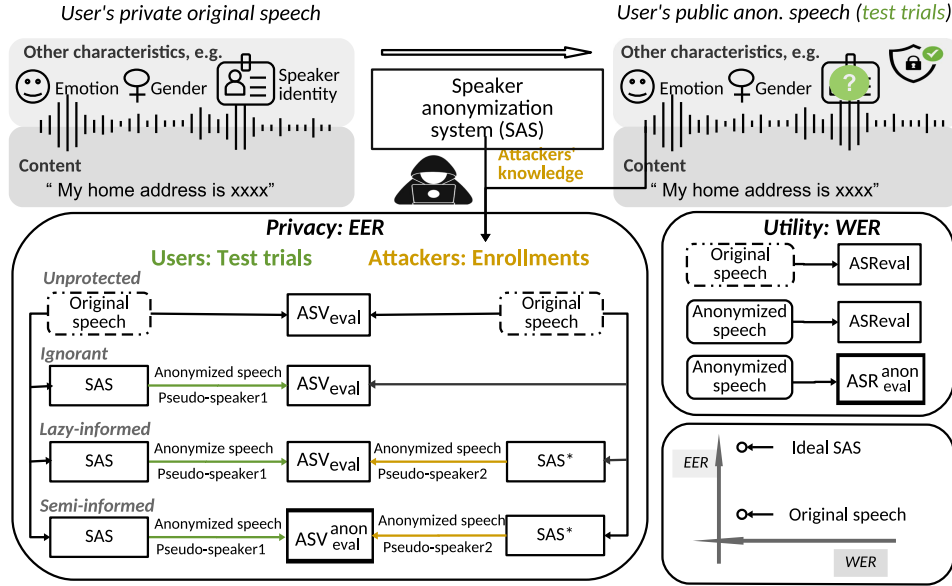


Fig. 2. Speaker anonymization task. A user anonymizes original speech to hide his or her identity before publication, and attackers use biometric (ASV) technology and knowledge of the anonymization method to re-identify the original speaker’s identity.

attack scenarios in terms of privacy and utility metrics. Under the *Semi-informed* condition, our proposed methods achieved better results for English speaker anonymization than all the submissions to VPC2022 [24], [25], [35], [36], [37], [38].

II. RELATED WORK

In this section, we introduce the VPC’s official design, which provides the setting for this study, including definitions of specific goals, attack models, and objective evaluation metrics. We also overview existing speaker anonymization approaches and their limitations.

A. The VoicePrivacy Challenges

The VPC formulates the speaker anonymization task as a game between users and attackers, as shown in Fig. 2. A user publishes anonymized data, called *test trials*, after applying an SAS to his or her original private speech. According to the VPC evaluation plan [13], an SAS should:

- output an anonymized speech waveform;
- conceal the speaker’s identity from different attackers;
- keep content and other paralinguistic attributes unchanged to maintain intelligibility and naturalness;
- ensure all test trials from the same speaker are attributed to the same pseudo-speaker, while test trials from different speakers have different pseudo-speakers¹

1) Attack Models and Objective Evaluation Metrics:

¹This is called speaker-level anonymization. A different approach known as utterance-level anonymization assigns different pseudo-identities to different utterances of the same original speaker. In this work, we follow the VPC protocol and utilize speaker-level anonymization.

a) *Privacy metric*: To assess the ability to protect a speaker’s identity in different scenarios, the ASV performance in terms of the equal error rate (EER) is computed as the primary privacy metric by using language-matched ASV evaluation models. This metric is calculated under the four attack models shown in the lower left of Fig. 2. The attackers are assumed to have access to a few original or anonymized utterances for each speaker, called *enrollment* utterances, and to have different levels of knowledge about the SAS:

- *Unprotected*: No anonymization is applied, and attackers verify the original test trials against the original enrollment data by using an ASV system trained on the original dataset, denoted ASV_{eval} .
- *Ignorant*: Attackers are unaware of the anonymization strategy used for the test trial utterances; instead, they use the original enrollment data and ASV_{eval} to infer a speaker’s identity.
- *Lazy-informed*: Attackers use a similar SAS without accurate parameters to anonymize their enrollment data, and they use ASV_{eval} to detect a speaker’s identity.
- *Semi-informed*: The only difference from *Lazy-informed* is that the attackers use ASV_{eval}^{anon} , a more powerful version trained on anonymized speech, to reduce the mismatch between the original and anonymized speech and infer the speaker’s identity.

b) *Primary utility metric*: To assess how well speech content is preserved in anonymized speech, the ASR performance in terms of the word error rate (WER) is computed as a primary utility metric by using language-matched ASR evaluation models. As illustrated in the lower right of Fig. 2, two ASR models are trained in the same way to decode the anonymized data: ASR_{eval} , trained on the original data, and ASR_{eval}^{anon} , trained on the anonymized data. This enables exploration of whether

speech content can be maintained better by simply retraining with similarly anonymized data.

c) Secondary utility metric: To assess and visualize the preservation of voice distinctiveness, the gain of voice distinctiveness metric, G_{VD} [39], [40], is computed. Precisely, $M = (M(i, j))_{1 \leq i \leq N, 1 \leq j \leq N}$ is a voice similarity matrix for N speakers, where the similarity value $M(i, j)$ for speakers i and j is formulated as follows:

$$M(i, j) = \text{sigmoid} \left(\frac{1}{n_i n_j} \sum_{\substack{1 \leq k \leq n_i \text{ and } 1 \leq l \leq n_j \\ k \neq l \text{ if } i=j}} \text{LLR}(x_k^{(i)}, x_l^{(j)}) \right), \quad (1)$$

Here, n_i and n_j are the numbers of utterances for each speaker; and $\text{LLR}(x_k^{(i)}, x_l^{(j)})$ is the log-likelihood ratio obtained by comparing the k -th utterance of the i -th speaker with the l -th utterance of the j -th speaker. These LLR scores are computed by probabilistic linear discriminant analysis (PLDA) [41] of the ASV_{eval} model trained on the original data.

Three matrices are constructed from the original (o) and anonymized (a) data: M_{oo} from the original data, M_{oa} from the original and anonymized data, and M_{aa} from the anonymized data. The diagonal dominance $D_{\text{diag}}(M)$ is computed as the absolute difference between the mean values of diagonal and off-diagonal elements:

$$D_{\text{diag}}(M) = \left| \sum_{1 \leq i \leq N} \frac{M(i, i)}{N} - \sum_{\substack{1 \leq j \leq N \text{ and } 1 \leq k \leq N \\ j \neq k}} \frac{M(j, k)}{N(N-1)} \right|. \quad (2)$$

Next, G_{VD} [39] is defined as the diagonal dominance ratio of the two matrices:

$$G_{VD} = 10 \log_{10} \frac{D_{\text{diag}}(M_{aa})}{D_{\text{diag}}(M_{oo})}, \quad (3)$$

Here, a gain of $G_{VD} = 0$ dB indicates that voice distinctiveness is preserved on average after anonymization, while a gain above or below 0 dB corresponds respectively to an average increase or decrease in voice distinctiveness.

An ideal anonymization system should achieve high EERs (close to 50%) in the *Ignorant*, *Lazy-informed*, and *Semi-informed* scenarios to protect the speaker's information. In addition, the WER should be as low as for the original speech, and G_{VD} should be close to 0 dB to preserve voice distinctiveness.

B. Existing Speaker Anonymization Approaches

1) Digital Signal Processing (DSP) Methods: A simple approach [14] that does not require training data is to change speaker attributes with distortion of the spectral envelope by using McAdams coefficients [42] to randomly shift the positions of formant frequencies. Widening of formant peaks [15] further distorts the spectral envelope. Data-driven formant modification can also be applied by using the formant statistics of desired speakers [16] or time-scale algorithms [18]. Phonetically controllable anonymization [17] modifies a speaker's vocal tract and voice source features, with a focus on F0 trajectories. Although

these methods perceptually manipulate the speech signal, previous works have indicated that powerful attackers can effortlessly recover speaker identities [11], [12], [43].

2) Disentangled Representation Methods: A typical approach based on disentangled representation learning, called x-vector based anonymization, is used as the primary baseline in the VPC [10], [11], [12], [13]. It extracts speaker representations and linguistic features by using a pretrained TDNN-based ASV system [28] and ASR AM based on a factorized time-delay neural network (TDNN-F), respectively. Then, to hide the original speaker's information, a selection-based speaker anonymizer [30] replaces the original x-vector with the mean vector of a set of randomly selected speaker vectors from an external pool of English speakers. Specifically, given a centroid of source speaker vectors from one speaker, the cosine distance is used to find the 200 farthest centroids in an external speaker vector pool, and 100 of those are randomly selected and averaged to obtain an anonymized speaker vector [30]. Finally, an SS AM generates mel-filterbank features from the anonymized pseudo x-vector, F0, and linguistic features, and an NSF-based waveform generator synthesizes anonymized speech.

Because this disentanglement-based method is more effective at protecting speaker identities than the DSP-based methods discussed in Section II-B1 [12], [43], most speaker anonymization studies have followed a similar framework. Improvements mainly come from two sources:

Improved speech disentanglement: Some works [44], [45], [46] have argued that the disentangled linguistic information extracted from the language-specific ASR AM and F0 still contain speaker information. Accordingly, they modify the F0 and linguistic information to remove the residual speaker identity.

Improved speaker vector anonymization: Other researchers have modified the original x-vector in ways that increase the privacy protection ability. Perero-Codosero et al. [47] transformed an original x-vector to an anonymized one by using an autoencoder with an adversarial training strategy to suppress speaker, gender, and accent information. This requires labels for the speaker identity, gender, and nationality. Turner et al. [48] sampled anonymized x-vectors from a Gaussian mixture model in a space reduced by principal component analysis (PCA) over an external pool of speakers, which preserves the distributional properties of the original x-vectors. There have been recent attempts to generate a target pseudo-speaker for speaker anonymization in the systems submitted to the VoicePrivacy Challenge 2022. For example, Meyer et al. [24] utilized a generative adversarial network to generate artificial speaker embeddings, where the anonymization stage requires a manual search to find vectors that are dissimilar to the anonymized one. Yao et al. [25] proposed using a look-up table (LUT)-based method to generate pseudo-speaker embeddings, along with an average of randomly selected speaker embeddings from the real speakers. However, it suffers from limited variability in the anonymized voices. Chen et al. [35] proposed a method for distorting an input speech signal by adding adversarial noise designed to hide the original speaker identity.

Most of the existing approaches are limited in two aspects. First, they use an ASR-based content extractor that requires large

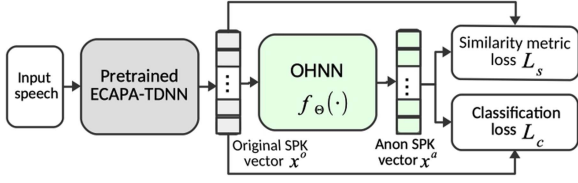


Fig. 3. Framework for an OHNN-based anonymizer. The \mathbf{x}^o are original speaker representations extracted from a pretrained ECAPA-TDNN, which then pass through a transfer module $f(\cdot)$ to produce the corresponding anonymized speaker representations \mathbf{x}^a . The \mathbf{x}^o and \mathbf{x}^a are trained as different speakers. After training, \mathbf{x}^a is used as a pseudo-speaker vector to synthesize speech.

amounts of transcribed English training data. Such an ASR-based content extractor is ineffective for speaker anonymization in unseen languages. Our previous work alleviates this issue by using an SSL-based content extractor [33]. As shown in Fig. 1, this SSL-based SAS consists of a HuBERT-based soft content encoder [49], an ECAPA-TDNN speaker encoder [29], an F0 extractor, and a HiFi-GAN decoder [32]. It does not require text transcriptions or any other language-specific resources, and it has demonstrated the ability to anonymize speech data with reasonable performance even if the data is in a language not included in the training data. However, it suffers from a remaining limitation of selection-based anonymizers according to previous results [11], [12], [13], [33], [34]: the distribution of the external speaker pool significantly affects anonymized speakers, and the averaging of vectors from the speaker pool reduces voice distinctiveness.

III. PROPOSED OHNN-BASED ANONYMIZER

To mitigate the problems with existing approaches, we propose the OHNN-based anonymizer shown in Fig. 3. Hence, this section formulates speaker anonymization as a constrained optimization problem, describes a general form of the proposed anonymizer, and explains the implementation details.

A. Problem Formulation

The training set $\{(\mathbf{x}_i^o, y_i^o)\}_{i=1}^M$ comprises M speaker vector \mathbf{x}_i^o and the corresponding speaker label y_i^o . The speaker vector $\mathbf{x}_i^o \in \mathbb{R}^d$ is a d -dimensional segment-level speaker embedding obtained from an ECAPA-TDNN pretrained on the original audio waveform. \mathbf{x}_i^o follows an unknown distribution $\mathbf{x}_i^o \sim p_{\mathbf{x}^o}$.

Anonymized speaker vectors $\mathbf{x}_i^a \in \mathbb{R}^d$ are obtained by transforming \mathbf{x}_i^o with a function $f_\Theta: \mathbb{R}^d \rightarrow \mathbb{R}^d$, written as follows:

$$\mathbf{x}^a = f_\Theta(\mathbf{x}^o). \quad (4)$$

Accordingly, the anonymized speaker vectors follow another distribution $\mathbf{x}_i^a \sim p_{\mathbf{x}^a}$ or $\mathbf{x}_i^a \sim p_{f_\Theta(\mathbf{x}^o)}$.

An ideal speaker anonymization method should meet at least three constraints:

- **Speaker privacy protection:** \mathbf{x}_i^o and \mathbf{x}_i^a are dissimilar to hide the original speaker identity. More specifically, in the context of VPC, \mathbf{x}_i^o and \mathbf{x}_i^a are dissimilar to the extent that the anonymized speech generated using \mathbf{x}_i^a is recognized as being a different speaker by the attackers' ASV.

- **Speaker diversity:** \mathbf{x}_i^a has a unique speaker identity y_i^a to maintain the diversity of anonymized speech across different speakers.
- **Distribution similarity:** $\mathbf{x}_i^a \sim p_{\mathbf{x}^a}$ satisfies the same distribution as \mathbf{x}_i^o to maintain the naturalness of the original speech.

The above constraints can be formulated as an optimization problem:

$$(\Theta, \Psi)^* = \arg \min_{\Theta, \Psi} \mathbb{E}_{\{\mathbf{x}^o, y^o\} \in D} [\lambda \mathcal{L}_s(\mathbf{x}^o, f_\Theta(\mathbf{x}^o)) + \mathcal{L}_c(y^o, g_\Psi(\mathbf{x}^o); y^a, g_\Psi(f_\Theta(\mathbf{x}^o)))] , \quad (5)$$

$$\text{s.t. } \mathcal{D}(p_{\mathbf{x}^o}, p_{f_\Theta(\mathbf{x}^o)}) < \epsilon, \quad (6)$$

where λ is a hyperparameter to balance the multi-objective function. \mathcal{L}_s is a similarity metric to optimize Θ by minimizing the similarity of the original-anonymized pair, which ideally makes the original and anonymized speech be recognized as different speakers by the attackers' ASV.

Next, $g_\Psi(\cdot)$ denotes the classifier layer, and \mathcal{L}_c is its classification loss function to optimize Θ and Ψ by minimizing the discrepancy between the sets of desired outputs, y^o, y^a , and predicted outputs, $g_\Psi(\mathbf{x}^o), g_\Psi(\mathbf{x}^a)$. The outputs may be defined for a multi-speaker classification task in which the original and corresponding anonymized speaker vectors are intentionally treated as different target speaker classes. This means that all speaker vectors after anonymization are treated as different classes, as well as different classes from the original speakers to maintain speaker diversity.

Finally, $\mathcal{D}(p_{\mathbf{x}^o}, p_{f_\Theta(\mathbf{x}^o)})$ is the divergence between distributions of \mathbf{x} included in a training database before and after anonymization. This term ensures similarity between the distributions of the anonymized and original speaker vectors, with some tolerance ϵ . The Kullback–Leibler divergence (KLD) or other types of divergence are applicable.

B. General Form of Proposed Anonymizer

Finding a direct solution of (5) and (6) for an arbitrarily designed DNN-based f_Θ is difficult. Here, we propose an anonymizer that, with a few assumptions, always satisfies the constraint in (6) regardless of the value of Θ . In such a case, Θ and Ψ can be optimized via (5) and a conventional gradient descent method.

Let $\boldsymbol{\mu}_{\mathbf{x}^o} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}_{\mathbf{x}^o} \in \mathbb{R}^{d \times d}$ be the mean and covariance matrix of $p_{\mathbf{x}^o}$, respectively. Our proposed anonymizer $f_\Theta(\cdot)$ can be written as follows:

$$\mathbf{x}^a = f_\Theta(\mathbf{x}^o) = \mathbf{L}_{\mathbf{x}^o}^{-1} \mathbf{W} \mathbf{L}_{\mathbf{x}^o} (\mathbf{x}^o - \boldsymbol{\mu}_{\mathbf{x}^o}) + \boldsymbol{\mu}_{\mathbf{x}^o}, \quad (7)$$

where $\mathbf{L}_{\mathbf{x}^o}$ is a whitening matrix² that satisfies $\mathbf{L}_{\mathbf{x}^o}^{-1} \boldsymbol{\Sigma}_{\mathbf{x}^o}^{-1} \mathbf{L}_{\mathbf{x}^o}^{-1 \top} = \boldsymbol{\Sigma}_{\mathbf{x}^o}$, and $\mathbf{W} \in \mathbb{R}^{d \times d}$ is an orthogonal matrix that satisfies $\mathbf{W} \mathbf{W}^\top = \mathbf{W}^\top \mathbf{W} = \mathbf{I}$. While $\boldsymbol{\mu}_{\mathbf{x}^o}$ and $\mathbf{L}_{\mathbf{x}^o}$ are determined by the data distribution, the values of \mathbf{W} are learned via (5).

Before introducing the parameterization and optimization of \mathbf{W} , we show that the proposed anonymizer satisfies

² $\mathbf{L}_{\mathbf{x}^o}$ is a whitening matrix. It can be derived from $\boldsymbol{\Sigma}_{\mathbf{x}^o}$ by a matrix decomposition method used in, e.g., PCA or Cholesky whitening [50].

$\mathcal{D}(p_{\mathbf{x}^o}, p_{f_{\Theta}(\mathbf{x}^o)}) = 0$ given that $p_{\mathbf{x}^o}$ is a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}^o}, \boldsymbol{\Sigma}_{\mathbf{x}^o})$ ³ We first decompose (7) into three steps:

- Centering and whitening: $\tilde{\mathbf{x}}^o = \mathbf{L}_{\mathbf{x}^o}(\mathbf{x}^o - \boldsymbol{\mu}_{\mathbf{x}^o})$,
- Rotation: $\tilde{\mathbf{x}}^a = \mathbf{W}\tilde{\mathbf{x}}^o$,
- De-whitening and de-centering: $\mathbf{x}^a = \mathbf{L}_{\mathbf{x}^o}^{-1}\tilde{\mathbf{x}}^a + \boldsymbol{\mu}_{\mathbf{x}^o}$.

The centered and whitened speaker vector $\tilde{\mathbf{x}}^o$ obviously follows a normal distribution $\tilde{\mathbf{x}}^o \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. As \mathbf{W} is an orthogonal matrix, $\tilde{\mathbf{x}}^a$ also follows a normal distribution $\mathcal{N}(\mathbf{W}\mathbf{0}, \mathbf{W}\mathbf{W}^T) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. Through the affine transformation in the last step, we know that $\mathbf{x}^a \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}^o}, \mathbf{L}_{\mathbf{x}^o}^{-1}\mathbf{L}_{\mathbf{x}^o}^{-1T}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}^o}, \boldsymbol{\Sigma}_{\mathbf{x}^o})$. Hence, the defined anonymizer does not change the distribution, i.e., $\mathcal{D}(p_{\mathbf{x}^o}, p_{f_{\Theta}(\mathbf{x}^o)}) = 0$.

The above explanation also reveals the core idea of our proposed anonymizer: while it does not change the overall distribution, each speaker vector is rotated through an orthogonal transformation. The anonymized \mathbf{x}^a is guaranteed to be different from the original \mathbf{x}^o as long as $\mathbf{W} \neq \mathbf{I}$. While an infinite number of orthogonal matrices can be applied for rotation, the optimal \mathbf{W} with respect to the criterion in (5) must be estimated through an optimization process.

In real applications, $\boldsymbol{\mu}_{\mathbf{x}^o}$ and $\boldsymbol{\Sigma}_{\mathbf{x}^o}$ of the test set data are unknown. They can be estimated by collecting multiple samples from the test domain if it is possible. Otherwise, we can either use the statistics from the training set or make some simplifications. Through preliminary experiments, we found an effective, simplified form:

$$\mathbf{x}^a = f_{\Theta}(\mathbf{x}^o) = \mathbf{W}(\mathbf{x}^o - \boldsymbol{\mu}_{\mathbf{x}^o}^{\text{train}}) + \boldsymbol{\mu}_{\mathbf{x}^o}^{\text{train}}, \quad (8)$$

where $\boldsymbol{\mu}_{\mathbf{x}^o}^{\text{train}}$ is the mean of the speaker vectors in the training set, and $\boldsymbol{\Sigma}_{\mathbf{x}^o}$ is assumed to be an identity matrix.

C. Rotation Matrix Using Householder Reflection

We now need a specific way to parameterize \mathbf{W} to guarantee that the learned \mathbf{W} through gradient descent is orthogonal. While many methods can be used, we found that one based on a Householder reflection [53] is efficient for DNNs. Without loss of generality, assume that \mathbf{W} is a product of multiple orthogonal matrices:

$$\mathbf{W} = \mathbf{W}_1 \mathbf{W}_l \dots \mathbf{W}_L, \quad (9)$$

where each matrix $\mathbf{W}_l \in \mathbb{R}^{d \times d}$ is given by

$$\mathbf{W}_l = \mathbf{H}_{q_l} \mathbf{H}_{q_l-1} \dots \mathbf{H}_1, \quad q_l \leq d, \quad (10)$$

Here, each sub-matrix \mathbf{H}_{q_l} is constructed with a Householder reflection [53] given a non-zero vector $\mathbf{v}_{q_l} \in \mathbb{R}^d$ as follows:

$$\mathbf{H}_{q_l} = \mathbf{I} - \frac{2}{\mathbf{v}_{q_l}^T \mathbf{v}_{q_l}} \mathbf{v}_{q_l} \mathbf{v}_{q_l}^T. \quad (11)$$

The resulting \mathbf{H} is known to be an orthogonal matrix for any non-zero vector \mathbf{v}_{q_l} , i.e., $\mathbf{H}^T \mathbf{H} = \mathbf{H} \mathbf{H}^T = \mathbf{I}$ and $\mathbf{H} \neq$

³Being Gaussian is a desirable but not absolutely required condition to ensure $\mathcal{D}(p_{\mathbf{x}^o}, p_{f_{\Theta}(\mathbf{x}^o)}) = 0$, but many types of speaker vectors can be assumed to follow a multivariate Gaussian distribution in the high dimensional space. One example is the length-normalized i-vector [51]. Another example is the ECAPA-TDNN speaker vectors, which can be well modeled using PLDA with Gaussian distributions [52].

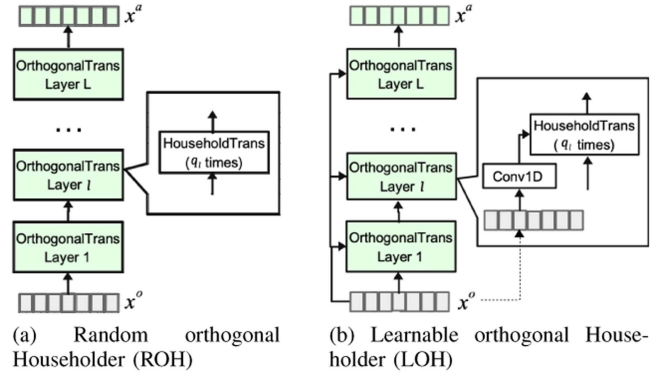


Fig. 4. Two types of OHNN-based anonymizers.

$\mathbf{I}, \forall \mathbf{v}_{q_l} \neq \mathbf{0}$. Accordingly, \mathbf{W}_l and \mathbf{W} are orthogonal and guaranteed not to be the identity matrix.

Equations (9)–(11) allow us to parameterize \mathbf{W} as $\{\dots, \mathbf{v}_{q_l}, \dots\}$. We further propose two implementations, which differ in how they compute \mathbf{v} :

- 1) *Random orthogonal Householder (ROH) reflection*: \mathbf{v} is treated as a learnable free parameter, i.e., $\Theta = \{\dots, \mathbf{v}_{q_l}, \dots\}$, and each \mathbf{v} is randomly initialized and optimized using (5). The anonymization process is illustrated in Fig. 4(a).
- 2) *Learnable orthogonal Householder (LOH) reflection*: Each \mathbf{v} is transformed from a small NN given the input \mathbf{x}^o . In such a case, Θ is the set of the trainable weights in a set of small NNs. Fig. 4(b) illustrates an implementation in which each DNN has a single 1D convolution layer with 192 output channels and a kernel size of 3.

While both implementations ensure that the transformation matrix \mathbf{W} is orthogonal, the first approach assumes a global transformation for all the input speaker vectors. In contrast, the latter approach assumes that the transformation matrix varies according to the input.

D. Loss Functions

Before delving into the details of the loss functions, we describe how to build batch data for an OHNN-based anonymizer. Let N be the batch size and C be the number of original speakers. Each mini-batch comprises $N/2$ original samples: $[\mathbf{x}^o, \mathbf{y}^o] = \{(\mathbf{x}_i^o, y_i^o)\}_{i=1}^{N/2}$, where $y_i^o \in [1, C]$ and $N/2$ corresponding anonymized samples, and $[\mathbf{x}^a, \mathbf{y}^a] = \{(\mathbf{x}_i^a, y_i^a + C)\}_{i=(N/2)+1}^N$. Therefore, the number of speakers is $2C$ during the training of an OHNN-based anonymizer.

We now explain the loss functions for learning the best values of Θ and Ψ as defined in (5). For the classification loss \mathcal{L}_c , we first consider the widely used AAM softmax loss [54], [55]:

$$\mathcal{L}_c = \mathcal{L}_{\text{AAM-softmax}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{\|\mathbf{w}_{y_i}\| \cdot \|\mathbf{x}_i\| \cdot \cos(\theta_{y_i, i+m_1})}}{Z}, \quad (12)$$

where

$$Z = e^{\|\mathbf{w}_{y_i}\| \cdot \|\mathbf{x}_i\| \cdot \cos(\theta_{y_i, i+m_1})} + \sum_{j=1, j \neq y_i}^{2C} e^{\|\mathbf{w}_j\| \cdot \|\mathbf{x}_i\| \cdot \cos(\theta_{j, i})},$$

\mathbf{w}_j is the j -th column of the weight in the fully-connected layer before the softmax layer, where $\mathbf{w} \in \mathbb{R}^{d \times 2C}$; and $\theta_{y_i, i}$ is the angle between \mathbf{x}_i and the target class's weight vector \mathbf{w}_{y_i} . After fixing the weight $\|\mathbf{w}_{y_i}\| = 1$ by ℓ_2 -normalization and rescaling $\|\mathbf{x}_i\|$ to s to ensure that the gradient is not too small during training, we can write (12) as

$$\mathcal{L}_c = \mathcal{L}_{\text{AAM-softmax}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{s(\cos(\theta_{y_i, i} + m_1))}}{Z}, \quad (13)$$

where $Z = e^{s(\cos(\theta_{y_i, i} + m_1))} + \sum_{j=1, j \neq y_i}^{2C} e^{s(\cos(\theta_{j, i}))}$. Since the target label y_i varies across the original and anonymized speakers, the classification loss $\mathcal{L}_{\text{AAM-softmax}}$ encourages the OHNN-based anonymizer to produce anonymized vectors that are varied for different speakers and distinct from original speaker vectors.

To further improve the discrepancy for original-anonymized (or anonymized-original) pair samples, we add an extra margin penalty m_2 into the AAM softmax loss. The approach is called weighted additive angular margin (w-AAM) softmax. Let $i \in [1, N]$ be the index of the original (or anonymized) sample in a mini-batch based on this batch data construction method. The corresponding anonymized (or original) sample is indexed by $(i + N/2) \% N$, where $\%$ denotes the modulo operation. The proposed w-AAM-based loss function $\mathcal{L}_{\text{w-AAM}}$ is similar to $\mathcal{L}_{\text{w-AAM}}$ except that the factor Z is defined as

$$Z = e^{s(\cos(\theta_{y_i, i} + m_1))} + e^{s(\cos(\theta_{y_{(i+N/2)\%N}, i} - m_2))} + \sum_{j=1, j \neq i, j \neq (i+N/2)\%N}^{2C} e^{s(\cos(\theta_{j, i}))}. \quad (14)$$

In our experiments, we set $m_1 = m_2 = 0.2$, $s = 30$ and compared the performance with settings of $\mathcal{L}_c = \mathcal{L}_{\text{AAM}}$ and $\mathcal{L}_c = \mathcal{L}_{\text{w-AAM}}$.

For the similarity metric \mathcal{L}_s , we choose the cosine similarity⁴ given by $\mathcal{L}_s(\mathbf{x}_i^o, \mathbf{x}_i^a) = \max(0, \cos(\mathbf{x}_i^o, \mathbf{x}_i^a) - m)$, we set the margin $m = 0$. The cosine similarity is a reasonable choice because it is closer to what most ASV systems use for scoring the similarity between speaker vectors. As the anonymizers are trained to minimize the cosine similarity between original and anonymized speaker vectors, the anonymized speech is expected to be judged as a different speaker by the attacker ASV, hence protecting the speaker's identity.

IV. EVALUATION

To evaluate the effectiveness of the SSL-based SAS using the proposed OHNN-based anonymizer under all the attack scenarios for English speaker anonymization, we followed the VPC evaluation plan [11], [12], [13] described in Section II. Then, we conducted anonymization experiments under a language-mismatched condition, using Mandarin data as the non-included language in the training database. The purpose of these experiments was to determine whether that the proposed OHNN-based anonymizer, which eliminates the need for an English

speaker pool, can effectively reduce the language mismatch present in anonymized speaker representations. As a result, better speech content preservation is achieved for Mandarin speaker anonymization.

A. Speaker Anonymization Dataset and Experimental Setup

1) *Dataset*: The SSL-based SAS was built using the following VPC standard datasets [11]: an ECAPA-TDNN speaker encoder trained on the *VoxCeleb-2* [56]; a HuBERT-based soft content encoder finetuned from a pretrained HuBERT Base model⁵ on *LibriTTS-train-clean-100* [57] to capture content representations; and a HiFi-GAN model trained on *LibriTTS-train-clean-100* [57].

Unlike the selection-based anonymizer, which relies on an additional multi-speaker English dataset (*LibriTTS-train-other-500*) containing data from 1,160 speakers as the external pool, the OHNN-based anonymizers reuse a multi-speaker multi-language dataset (*VoxCeleb-2*), that is used to train the ECAPA-TDNN of the SSL-based SAS [33]. This large-scale dataset contains over 1 million utterances by 5,994 speakers of 145 different nationalities.

English speaker anonymization was evaluated on the official VPC development and test sets [11], [12], [13]. These two sets contain English utterances by several female and male speakers from the *LibriSpeech* and *VCTK* [58] corpora. For the *Ignorant* and *Lazy-informed* conditions, we used the language-matched ASV_{eval} system provided by the VPC [11], [12], [13]. It was trained on the original *LibriSpeech-train-clean-360* English dataset. For the *Semi-informed* condition, we trained $ASV_{\text{eval}}^{\text{anon}}$ system in the same way as ASV_{eval} , but with anonymized speech data. Likewise, ASR_{eval} and $ASR_{\text{eval}}^{\text{anon}}$ were trained with the same original and anonymized speech data, respectively.

The same anonymization systems used for English speakers were directly adopted for Mandarin speaker anonymization without training or fine-tuning on Mandarin data. The evaluation for Mandarin was conducted on a test set sampled from a 20-hour, multi-speaker Mandarin corpus called *AISHELL-3* [59]. The test set contains 4,267 utterances by 44 speakers. We split the utterances into test trial (88 utterances) and enrollment (4,179 utterances) subsets, which were used to produce 10,120 enrollment-test pairs for ASV evaluation, including 2,200 same-speaker and 7,920 different-speaker pairs. The ASV evaluation model $ASV_{\text{eval}}^{\text{mand}}$ under the *Lazy-informed* condition was an ECAPA-TDNN trained on the Mandarin datasets *CN-Celeb-1 & 2* [60], [61]. The ASV evaluation model under the *Semi-informed* condition called $ASV_{\text{eval}}^{\text{anon-mand}}$ was fine-tuned from $ASV_{\text{eval}}^{\text{mand}}$ using anonymized utterances from 285 speakers in the interview, speech and live broadcasting genres of *CN-Celeb-1 & 2*. The ASR evaluation model $ASR_{\text{eval}}^{\text{mand}}$ was a publicly available ASR Transformer [62] trained on a 150-hour Mandarin ASR dataset, *AISHELL-1* [63].

2) *Experimental Setup*: Table I lists notations for the different speaker anonymization approaches that we examined. **B1.a**, **B1.b**, and **B2** are the baseline systems from VPC 2022 [13].

⁴<https://pytorch.org/docs/stable/generated/torch.nn.CosineEmbeddingLoss.html>

⁵<https://github.com/pytorch/fairseq/tree/main/examples/hubert>

TABLE I
NOTATIONS FOR THE EVALUATED SPEAKER ANONYMIZATION METHODS

	Notation	Content encoder	Speaker encoder	Syn. model	Speaker anon.
Disentangle	B1.a [13]	F-TDNN	TDNN	SS-AM+NSF	Select.
	B1.b [13]	F-TDNN	TDNN	HiFi-GAN+NSF	Select.
	S-Select [33]	SSL	ECAPA	HiFi-GAN	Select.
	S-ROH	SSL	ECAPA	HiFi-GAN	ROH
	S-LOH	SSL	ECAPA	HiFi-GAN	LOH
DSP	B2 [13]	McAdams coefficients-based			

S-Select denotes the SSL-based SAS using a selection-based anonymizer. **S-ROH** denotes a system obtained by replacing the selection-based anonymizer of **S-Select** with a random OH (ROH) anonymizer and keeping other components unchanged. Likewise, **S-LOH** indicates the use of a learnable OH (LOH) anonymizer. Noted that, hereafter, **S-ROH*** and **S-LOH*** refer to models trained with the w-AAM and cosine similarity losses.

For **S-Select**, the YAAPT algorithm [64] is used to extract the F0. The ECAPA-TDNN with 512 channels in the convolution frame layers [29] provides 192-dimensional speaker identity representations. The HuBERT-based soft content encoder [49] takes the CNN encoder and the first and sixth transformer layers of the pretrained HuBERT base model as a backbone. It down-samples a raw audio signal into a 768-dimensional continuous representation, which is then mapped to a 200-dimensional vector by one projection layer to predict discrete speech units. These speech units are obtained by discretizing the intermediate 768-dimensional representations via k -means clustering⁶ [65], [66]. The training procedures are detailed in [33]. For the selection-based anonymizer, attackers had different random seeds from users when randomly choosing 100 speaker vectors from the 200 farthest ones; thus, the attackers had different pseudo-speaker vectors.

The OHNN-based anonymizer accepts 192-dimensional speaker representations extracted from a pretrained ECAPA-TDNN, which was the same here as the ECAPA-TDNN of the SSL-based SAS. We followed the VPC evaluation plan, in which attackers in the *Lazy-informed* and *Semi-informed* scenarios have partial knowledge of the speaker anonymizer. They are assumed to know the training dataset, structure, loss functions, and other training parameters of the user’s OHNN-based anonymizer, except for the training seed to initialize the training weights. Specifically, the training seeds were 50 and 1986 for users and attackers, respectively.⁷ Using knowledge of the OHNN-based anonymizer, an attacker trains a new anonymizer to anonymize speech. All the OHNN-based anonymizers were trained with a cyclical learning rate [67], which varied between $1e-8$ and $1e-3$, and the Adam optimizer [68] by using the SpeechBrain [62] toolkit based on PyTorch [69]. The number of iterations of one cycle was set to 130 k. We fixed $d = 192$, $L = 12$ for both the ROH and LOH anonymizers, but we use

⁶https://github.com/pytorch/fairseq/tree/main/examples/textless_nlp/gslm/speech2unit

⁷The training seeds can be any values as long as they are different for users and attackers.

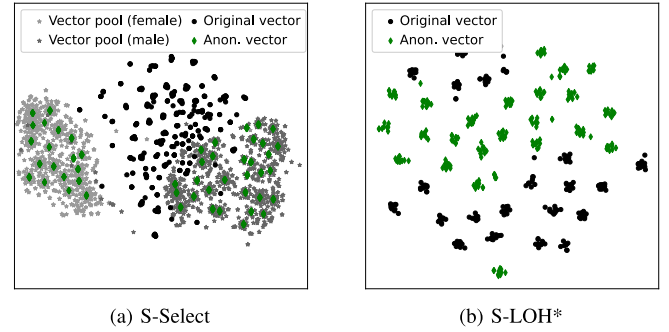


Fig. 5. Visualization of original and anonymized speaker vectors generated by the **S-Select** and **S-LOH*** anonymizers.

$q_l = 192$ and $q_l = 50$ for the ROH and LOH training, respectively. The hyperparameter λ in (5) was set to 20^8 .

B. Speaker Anonymization Experiments in English

For the English experiments, first, we explored the difference between selection- and OHNN-based anonymizers by comparing the performance of **S-Select** and **S-LOH***. Then, we investigated different configurations for the OHNN-based anonymizer, including the losses and whether to explicitly use speaker information to optimize the Householder transformation. Finally, we compared SSL-based speaker anonymization using an OHNN-based anonymizer with other approaches, including the disentanglement- and DSP-based approaches.

1) *Comparison of Selection- and OHNN-Based Anonymizers*: In the first experiments, we visualized the original and anonymized speech generated by **S-Select** and **S-LOH*** in terms of speaker embeddings, the cosine similarity of the speech pairs, and voice distinctiveness.

Original and anonymized speaker embeddings: To show the difference between the **S-Select** and **S-LOH*** anonymizers, we first applied t-distributed stochastic neighbor embedding (t-SNE) [70] to visualize the original and anonymized embeddings. The results are shown in Fig. 5. The speaker embeddings were extracted from 50 speakers in the *VoxCeleb-2* training set, which are shown in different colors, and 10 utterances were randomly selected from each speaker. Clearly, the anonymized speaker vectors generated by **S-Select** were heavily dependent on the distribution of an external pool, whereas **S-LOH*** generated distinctive anonymized speaker vectors that followed the distribution of the original speaker vector space.

Cosine similarity distribution on speech pairs: Fig. 6 plots the cosine similarities between pairs of speaker vectors extracted from generated speech for all the test sets of *LibriSpeech* and *VCTK* on speech pairs provided by [12]. Depending on the attack condition, the speech can be original or anonymized generated by **S-Select** or **S-LOH***. For the *Unprotected* condition, shown on the left side of Fig. 6, the positive cosine similarity distributions (green) are close to 1, and the negative distributions (yellow) are close to 0, which indicates that the speaker vectors

⁸Audio samples are available at <https://github.com/nii-yamagishilab/SSL-SAS>

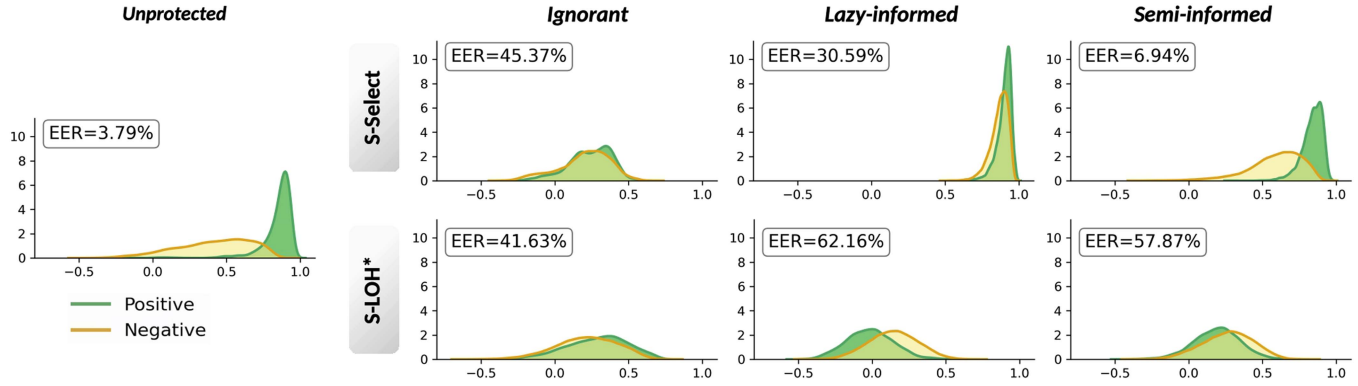


Fig. 6. Cosine similarities between pairs of the speaker vectors extracted from the generated speech of users and different attackers. Positive: paired utterances from the same speaker. Negative: paired utterances from different speakers.

of the original speech were highly discriminative. To protect speaker privacy, an ideal SAS should push the positive score distributions toward the negative ones regardless of the attacker type.

On the right side of Fig. 6, the top part shows the score distributions for three attacker conditions with **S-Select**. There are much bigger overlaps of the positive and negative distributions for the *Ignorant* condition than for the *Unprotected* condition, which means that **S-Select** achieved reasonable speaker privacy performance under the *Ignorant* condition. Unfortunately, the overlaps are smaller for the *Lazy-informed* and *Semi-informed* conditions. This reveals the reason for the significant speaker privacy leakage under more powerful attack conditions. Moreover, most of the cosine similarity scores are very close to 1, which may pose a risk of reducing the diversity of the anonymized speakers.

The bottom right of Fig. 6 shows the score distributions for three attacker conditions with **S-LOH***. The overlaps of the positive and negative distributions are well magnified under all the attack scenarios. This verifies the effectiveness of our OHNN-based anonymizer in ensuring that the attackers cannot gain significant speaker privacy information from users. Furthermore, most of the cosine similarity scores are far from 1, indicating the diversity of the anonymized speakers.

Comparison of gain of voice distinctiveness (G_{VD}): Fig. 7 shows voice similarity matrices obtained for **S-Select** and **S-LOH***. The upper-left submatrix of each matrix M is M_{oo} , and the distinct diagonal reflects the high voice distinctiveness within the original speech. The upper-right (or lower-left) submatrix M_{oa} reflects the voice similarity between the original and the anonymized speech, such that the diagonal disappears when they differ. The lower-right submatrix M_{aa} reflects the voice similarity within the anonymized speech, where a dominant diagonal appears if the anonymized speakers remain distinguishable [39]. There is a very weak dominant diagonal in M_{aa} for **S-Select**, indicating that voice distinctiveness was lost among the anonymized speakers. In contrast, the matrices for **S-LOH*** exhibit distinct diagonals in M_{aa} , indicating that voice distinctiveness was preserved after anonymization.

In general, the **S-LOH*** anonymizer met the three constraints described in Section III-A: good privacy protection, voice

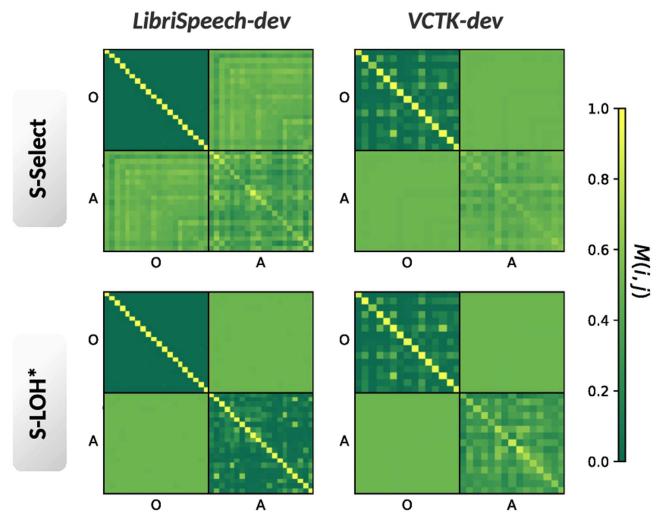


Fig. 7. Voice similarity matrices for **S-Select** and **S-LOH*** on the female speakers in the LibriSpeech-dev and VCTK-dev datasets. The global matrix M for each system comprises three submatrices M_{oo} , M_{oa} , and M_{aa} defined in Section II-A1 via $M = \begin{bmatrix} M_{oo} & M_{oa} \\ M_{oa} & M_{aa} \end{bmatrix}$.

distinctiveness, and naturalness of the speaker vector space from the above analysis and visualization.

2) *Effects of Various Components for Proposed OHNN-Based Anonymizer:* The proposed OHNN-based anonymizer has two novel components: the loss functions and the Householder transformations. Table II summarizes the average EERs and WERs⁹ under all attack scenarios using two OHNN-based anonymizers with different losses. In the table, \uparrow indicates a better performance with higher values, while \downarrow indicates a better performance with lower values.

Effect of the different losses: For the proposed OHNN-based anonymizer, w-AAM+cos performed better than AAM+cos in terms of the EER under most attacker conditions. This was because the introduced margin of w-AAM expands the inter-class variance of original-anonymized pairs, thus increasing the dissimilarity.

⁹The EER weights and detailed results for each subset are given in Appendix A. Due to limited space, other results are moved to the appendix of the article on Arxiv.

TABLE II
AVERAGE EER (%), WER (%), AND $G_{VD}(dB)$ ON THE VPC ENGLISH DEVELOPMENT(DEV) AND TEST SETS

	Original		OHNN-based anonymizer							
			S-ROH				S-LOH			
			AAM+cos		w-AAM+cos		AAM+cos		w-AAM+cos	
	dev	test	dev	test	dev	test	dev	test	dev	test
<i>Ignorant</i> by $ASV_{eval} \uparrow$	3.54	3.79	43.28	45.09	47.60	49.83	45.19	42.17	45.94	41.63
<i>Lazy-informed</i> $ASV_{eval} \uparrow$	3.54	3.79	40.20	47.37	41.69	45.16	47.49	49.62	59.31	62.16
<i>Semi-informed</i> $ASV_{eval} \uparrow$	-	-	7.75	42.41	41.62	41.88	41.79	40.66	60.12	57.87
WER by $ASR_{eval} \downarrow$	7.30	8.48	9.31	10.20	9.32	10.24	9.31	10.22	9.28	10.20
WER by $ASR_{anon} \downarrow$	-	-	7.52	7.72	7.67	7.84	7.47	7.99	7.52	7.94
$G_{VD} \uparrow$	0	0	-1.86	-1.59	-1.92	-1.64	-3.89	-3.55	-2.52	-2.25

The speaker vectors were anonymized by an OHNN-based anonymizer with AAM+cos or w-AAM+cos. \uparrow indicates better performance with higher values, while \downarrow indicates better performance with lower values.

TABLE III
AVERAGE EER (%), WER (%), AND $G_{VD}(dB)$ ON THE VPC ENGLISH DEV AND TEST SETS WHEN PROCESSED BY VARIOUS SPEAKER ANONYMIZATION SYSTEMS

	DSP		Selection-based anonymizer								OHNN-based anonymizer			
			Original		B2 [13]		B1.a [13]		B1.b [13]		S-Select [33]		S-ROH*	
	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test
<i>Ignorant</i> by $ASV_{eval} \uparrow$	3.54	3.79	37.01	38.29	53.14	50.29	53.91	52.14	48.23	45.37	47.60	49.83	45.94	41.63
<i>Lazy-informed</i> $ASV_{eval} \uparrow$	3.54	3.79	43.80	45.04	32.12	32.82	27.39	27.51	29.29	30.59	41.69	45.16	59.31	62.16
<i>Semi-informed</i> $ASV_{eval} \uparrow$	-	-	6.53	7.77	11.74	11.81	9.93	9.18	7.75	6.94	41.42	41.86	60.12	57.87
WER by $ASR_{eval} \downarrow$	7.30	8.48	17.15	18.52	10.88	10.98	10.69	10.84	8.73	9.77	9.32	10.24	9.28	10.20
WER by $ASR_{anon} \downarrow$	-	-	8.04	9.03	7.94	8.29	7.59	7.56	7.74	8.44	7.67	7.84	7.52	7.94
$G_{VD} \uparrow$	0	0	-1.72	-1.63	-9.17	-10.15	-6.44	-6.44	-7.90	-7.99	-1.92	-1.64	-2.52	-2.25

* indicates that w-AAM+cos was used for training S-ROH and S-LOH.

Effect of different Householder transformations: Clearly, the LOH anonymizers generally achieved better EERs than the ROH did. This result supports the view that, instead of using a global transformation for ROH, the LOH is more flexible because it learns from the speaker embeddings and thus brings more discriminative information.

For the WERs, those computed by ASR_{eval}^{anon} were consistently lower than those of ASR_{eval} for all systems. This implies that such utility degradation due to OHNN-based anonymizers can easily be offset by training ASR evaluation models on similar anonymized data. Meanwhile, all the OHNN-based anonymizers achieved similar WERs with ASR_{eval} or ASR_{eval}^{anon} , which confirms that the orthogonality of ROH and LOH did not change the distributions of the original and anonymized speaker vectors.

3) *Comparison of Various SASs Using Different Anonymizers: Primary privacy and utility evaluation:* Table III lists the average EER and WER results for various SASs under all scenarios. To anonymize the speaker representations, B2 randomly alters the formant position, B1.a, B1.b, and S-Select used the selection-based anonymizer, while S-ROH* and S-LOH* used the OHNN-based anonymizer.

First, we examine the results with the selection-based anonymizer. Using the selection-based anonymizer, the EERs of S-Select, B1.a and B1.b decreased by around 30% under the *Lazy-informed* condition and 7%–9% under the *Semi-informed* condition, indicating severe speaker privacy leakage.

Next, we examine the results with the proposed OHNN-based anonymizer integrated into different configurations. First, S-ROH* and S-LOH* could protect speaker information almost

as well as the VPC baselines (B1.a and B1.b) could when facing the *Ignorant* attacker. Moreover, for the *Lazy-informed* and *Semi-informed* attackers, it comfortably outperformed all the baseline systems, achieving over 40% EER. Second, among all the methods, S-ROH* and S-LOH* preserved speech content the best with ASR_{eval}^{anon} , achieving even lower WERs than for original speech on average.

Another interesting observation is that, while B2, B1.a, B1.b, and S-select are effective for protecting user privacy under the *Ignorant* condition, the utility performance in terms of WER and G_{VD} is worse than that of the OHNN-based anonymizers. This suggests that the baseline methods sacrifice utility to achieve a high privacy protection performance. Our proposed methods achieve a good balance between improving both privacy and utility metrics under various attack scenarios.

Secondary utility evaluation: The bottom of Table III lists the results for the average gain of voice distinctiveness, G_{VD} . They indicate that our proposed S-ROH* and S-LOH* achieved much better preservation of voice distinctiveness than the SASs using the selection-based anonymizer. The G_{VD} results of the S-select and OHNN-based anonymizers again confirm the findings described in Section IV-B1. **MOS prediction:** To further analyze the effectiveness of our proposed models, we utilize a recently proposed mean opinion score (MOS) prediction network [71] to estimate the perceived naturalness as another utility metric. Box plots of the predicted MOS scores are shown in Fig. 9. The results demonstrate that S-Select has a higher naturalness than B2 and B1.a. After replacing the selection-based anonymizer

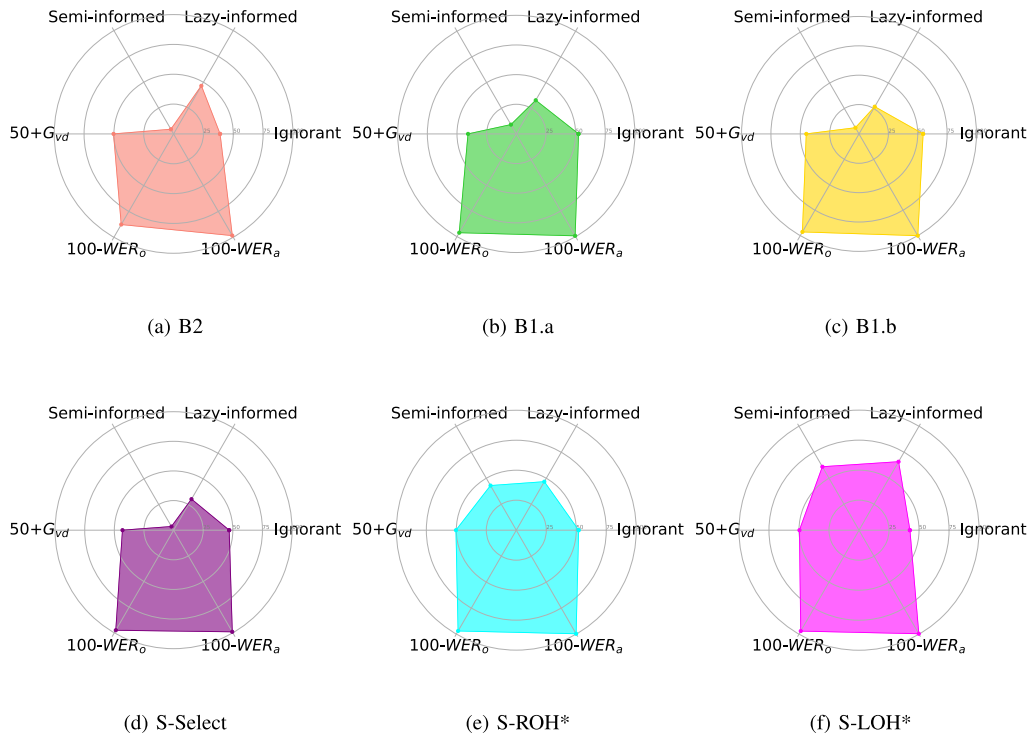


Fig. 8. Radar charts for each system on English speaker anonymization. All values are rescaled to [0, 100].

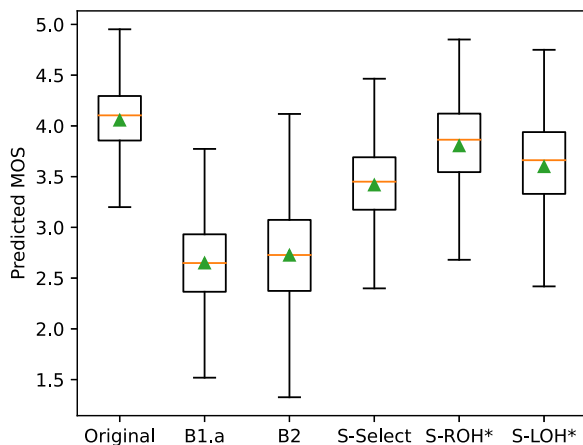


Fig. 9. Box plots on predicted naturalness scores of anonymized speech from experimental systems. Triangle symbols and the bar marks in the boxes represent mean and median scores, respectively.

with the OHNN-based anonymizers **S-ROH*** and **S-LOH***, we see a further improvement in naturalness.

Note that we used predicted MOS rather than human perception-based MOS obtained through listening tests in light of time and cost limits. The predicted MOS is reasonably well-aligned with human perception [71]. In Fig. 9, we can see that the ranking of the predicted MOS of the original, **B1.a**, and **B2** are consistent with those from the listening test done by the VPC [12].

Overall performance: As there are multiple metrics for evaluating the model performance, we summarize the results using

a radar chart for each system in Fig. 8. Each radar chart covers the EER values under the *Ignorant*, *Lazy-informed*, and *Semi-informed* conditions, WER_o by ASR_{eval} , WER_a by ASR_{anon}^{eval} , and G_{VD} . Note that the chart shows $100 - WER$, so the higher the better. Accordingly, a larger shaded area in the radar plot indicates a better overall performance. It is evident that the proposed **S-ROH*** and **S-LOH*** achieve larger shaded areas than the other systems, which performed particularly worse under the challenging semi-informed condition.

C. Speaker Anonymization Experiments in Mandarin

Table IV lists the EERs and CERs for the Mandarin test dataset. The first observation is that baselines **B1.a** and **B2** obtained EERs higher than 30% under the three conditions, but the CERs were higher than 60%. These results indicate that both systems achieved a high level of speaker identity protection by heavily distorting the speech contents. In particular, the results of **B1.a** suggest that it was inappropriate to use the ASR AM trained on the English data to extract speech content from the Mandarin data. The second observation is that the trends for the **S-Select** and OHNN-based anonymizers with different losses were remarkably similar to those observed on the English test sets.

The proposed OHNN-based anonymizers obtained ASV EERs higher than 30% under all evaluation conditions, and the CERs were lower than those of other systems. Compared to the baselines, the proposed systems adequately protected the speaker information without heavily sacrificing the speech contents. Compared to the selection-based system, the

TABLE IV
EER (%) AND CER (%) ON MANDARIN DATA WITH $ASV_{\text{eval}}^{\text{MAND}}$, $ASV_{\text{eval}}^{\text{ANON}^{\text{MAND}}}$, AND $ASR_{\text{eval}}^{\text{MAND}}$

	Original	DSP	Selection-based anonymizer		OHNN-based anonymizer			
		B2	B1.a	S-select	S-ROH		S-LOH	
					AAM+cos	w-AAM+cos	AAM+cos	w-AAM+cos
<i>Ignorant</i> by $ASV_{\text{eval}}^{\text{mand}} \uparrow$	2.04	35.50	44.54	37.90	32.09	34.82	33.28	33.27
<i>Lazy-informed</i> by $ASV_{\text{eval}}^{\text{mand}} \uparrow$	2.04	36.31	41.54	22.58	34.04	34.54	39.54	47.49
<i>Semi-informed</i> by $ASV_{\text{eval}}^{\text{anon}^{\text{mand}}} \uparrow$	-	42.73	42.44	19.82	42.81	41.72	40.72	48.99
CER by $ASR_{\text{eval}}^{\text{mand}} \downarrow$	10.36	61.90	68.67	18.92	17.15	17.28	17.20	17.90

A higher EER indicates better privacy, while a lower CER indicates better intelligibility.

proposed system[s?] achieved a lower CER while obtaining much higher ASV EERs, particularly in the most challenging *Lazy-informed* and *Semi-informed* scenarios. In particular, the CER on the anonymized speech decreased to less than 18% with the OHNN-based anonymizers, suggesting improved utility. One possible reason for the decreased CER when using OHNN-based anonymizers is that this mismatch was mitigated by the OHNN-based anonymizers trained using *VoxCeleb 2*, which contains large-scale, multi-speaker, and multi-language data.

V. CONCLUSION

This article has proposed a novel OHNN-based speaker anonymization approach that rotates original speaker vectors into anonymized ones with a distribution following the original speaker vector space. Towards good privacy protection and voice distinctiveness, AAM/w-AAM and cosine similarity loss functions were introduced to encourage the generation of distinctive anonymized speaker vectors. Experiments on English VPC datasets demonstrated that the proposed model protects speaker privacy while maintaining speech content: it achieved competitive performance under all attack scenarios in terms of privacy and utility metrics. Comparison of the cosine similarities between pairs of speaker vectors extracted from the generated speech with a commonly used selection-based anonymizer and the OHNN-based anonymizer further verified that our proposed method can effectively reduce privacy leakage when facing different attackers, while improving the diversity of anonymized speakers. Experiments on the Mandarin *AISHELL-3* datasets demonstrated that our OHNN-based anonymizer is more robust to the language mismatch scenario than the selection-based methods and can be adopted for this unseen-language anonymization task directly.

To further improve the privacy protection performance under various attack scenarios, our future work will investigate the training loss. One potential direction is to optimize the distance between the original and anonymized speaker vectors by integrating a proxy ASV evaluation model into the training process i.e., using an ASV to measure \mathcal{L}_s in (5) on original and anonymized speech waveforms. Such a training scheme is closer to how attackers infringe on the speaker's identity. Additionally, we are considering extending the OHNN-based anonymizer to protect other personal attributes such as age, gender, emotion,

and dialect. We previously proposed a system for concealing the gender of a speaker [72], and we feel the framework can be extended to other attributes as well. Our goal is to achieve controllable voice privacy protection that enables users to customize and control the anonymization process according to their specific privacy needs.

APPENDIX A DETAILED RESULTS

TABLE V
IGNORANT EER WITH AAM+COS, W-AAM+COS OF ROH AND LOH

Dataset	Gender	Weight	EER, %					
			S-Select	ROH		LOH		
				AAM+cos	wAAM+cos	AAM+cos	wAAM+cos	
LibriSpeech-dev	female	0.25	47.44	39.77	44.74	46.73	49.72	
	male	0.25	46.72	45.81	49.22	47.05	45.34	
VCTK-dev (diff.)	female	0.20	52.11	41.55	45.54	44.97	45.48	
	male	0.20	48.04	44.07	49.53	42.18	42.68	
VCTK-dev (com.)	female	0.05	47.97	45.93	47.67	42.44	45.64	
	male	0.05	45.30	49.29	54.42	43.87	45.30	
Weighted average dev			48.23	43.28	47.60	45.19	45.94	
LibriSpeech-test	female	0.25	41.24	35.77	40.51	37.77	41.61	
	male	0.25	42.54	48.78	51.22	43.43	39.42	
VCTK-test (diff.)	female	0.20	50.31	48.46	51.08	43.42	41.87	
	male	0.20	48.11	47.47	52.30	45.01	43.17	
VCTK-test (com.)	female	0.05	47.40	45.38	48.84	41.91	45.66	
	male	0.05	47.46	50.00	55.65	41.81	41.81	
Weighted average test			45.37	45.09	49.83	42.17	41.63	

TABLE VI
LAZY-INFORMED EER WITH AAM+COS, W-AAM+COS OF ROH AND LOH

Dataset	Gender	Weight	EER, %					
			S-Select	ROH		LOH		
				AAM+cos	wAAM+cos	AAM+cos	wAAM+cos	
LibriSpeech-dev	female	0.25	29.55	42.19	43.18	51.42	57.24	
	male	0.25	34.78	42.70	43.94	44.88	62.89	
VCTK-dev (diff.)	female	0.20	24.09	29.65	32.57	38.12	58.28	
	male	0.20	29.48	45.01	46.45	52.51	57.37	
VCTK-dev (com.)	female	0.05	20.35	35.17	37.50	47.38	57.56	
	male	0.05	29.63	45.87	44.73	58.40	66.53	
Weighted average dev			29.29	40.20	41.69	47.49	59.31	
LibriSpeech-test	female	0.25	29.74	40.88	40.51	53.10	64.78	
	male	0.25	33.18	47.88	43.65	54.34	67.04	
VCTK-test (diff.)	female	0.20	29.32	42.39	42.70	45.11	54.53	
	male	0.20	30.71	59.36	54.31	44.14	61.48	
VCTK-test (com.)	female	0.05	27.75	41.91	42.49	49.71	57.51	
	male	0.05	29.38	54.80	51.98	48.59	62.71	
Weighted average test			30.59	47.37	45.16	49.62	62.16	

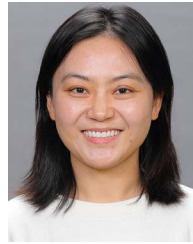
TABLE VII
SEMI-INFORMED EER WITH AAM+COS, W-AAM+COS OF ROH AND LOH

Dataset	Gender	Weight	EER,%				
			ROH		LOH		
			AAM +cos	wAAM +cos	AAM +cos	wAAM +cos	
LibriSpeech-dev	female	0.25	11.65	44.89	45.88	51.14	60.65
	male	0.25	4.96	40.06	43.01	36.18	61.49
VCTK-dev (diff.)	female	0.20	7.52	30.38	33.58	35.26	57.33
	male	0.20	6.79	43.82	42.18	44.07	61.14
VCTK-dev (com.)	female	0.05	7.55	33.14	36.92	36.05	52.91
	male	0.05	7.12	43.59	44.16	45.87	64.96
Weighted average dev			7.75	39.91	41.42	41.79	60.12
LibriSpeech-test	female	0.25	4.92	34.85	37.23	41.61	59.85
	male	0.25	2.89	42.34	44.10	45.66	59.69
VCTK-test (diff.)	female	0.20	13.48	41.77	40.74	35.44	48.77
	male	0.20	7.75	51.49	45.75	38.17	60.73
VCTK-test (com.)	female	0.05	10.40	41.62	39.02	38.44	55.34
	male	0.05	4.52	47.74	46.05	44.07	66.38
Weighted average test			6.94	42.41	41.88	40.66	57.87

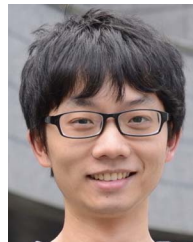
REFERENCES

- D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Commun.*, vol. 17, no. 1/2, pp. 91–108, 1995.
- X. Miao, I. McLoughlin, W. Wang, and P. Zhang, "D-MONA: A dilated mixed-order non-local attention network for speaker and language recognition," *Neural Netw.*, vol. 139, pp. 201–211, 2021.
- A. Ali et al., "The mgb-5 challenge: Recognition and dialect identification of dialectal arabic speech," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop.*, 2019, pp. 1026–1033.
- X. Miao, I. McLoughlin, and Y. Song, "Variance normalised features for language and dialect discrimination," *Circuits, Syst., Signal Process.*, vol. 40, no. 7, pp. 3621–3638, 2021.
- A. A. Dibazar, S. Narayanan, and T. W. Berger, "Feature analysis for automatic detection of pathological speech," in *Proc. IEEE 2nd Joint 24th Annu. Conf. Annu. Fall Meeting Biome. Eng. Soc. Eng. Med. Biol.*, 2002, vol. 1, pp. 182–183.
- B. Schuller et al., "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. Interspeech, 14th Annu. Conf. Int. Speech Commun. Assoc.*, 2013, pp. 148–152.
- V. Vestman, T. Kinnunen, R. G. Hautamäki, and M. Sahidullah, "Voice mimicry attacks assisted by automatic speaker verification," *Comput. Speech Lang.*, vol. 59, pp. 36–54, 2020.
- Q. Jin, A. R. Toth, A. W. Black, and T. Schultz, "Is voice transformation a threat to speaker identification?," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2008, pp. 4845–4848.
- R. K. Das, X. Tian, T. Kinnunen, and H. Li, "The attacker's perspective on automatic speaker verification: An overview," in *Proc. Interspeech2020*, pp. 4213–4217, doi: 10.21437/Interspeech.2020-1052.
- F. Fang et al., "Speaker anonymization using x-vector and neural waveform models," in *Proc. 10th ISCA Speech Synth. Workshop*, 2019, vol. 9, pp. 155–160.
- N. Tomashenko et al., "Introducing the VoicePrivacy initiative," in *Proc. Interspeech*, 2020, pp. 1693–1697.
- N. Tomashenko et al., "The VoicePrivacy 2020 challenge: Results and findings," *Comput. Speech Lang.*, vol. 74, 2022, Art. no. 101362.
- N. Tomashenko et al., "The VoicePrivacy 2022 challenge evaluation plan," 2022, *arXiv:2203.12468*.
- J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, "Speaker anonymisation using the McAdams coefficient," in *Proc. Interspeech*, 2021, pp. 1099–1103.
- P. Gupta, G. P. Prajapati, S. Singh, M. R. Kamble, and H. A. Patil, "Design of voice privacy system using linear prediction," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2020, pp. 543–549.
- S. P. Dubagunta, R. V. Son, and M. M. Doss, "Adjustable deterministic pseudonymisation of speech: Idiapi-ntki's submission to VoicePrivacy 2020 challenge," 2020. [Online]. Available: <https://www.voiceprivacychallenge.org/docs/Idiap-NKI.pdf>
- L. Tavi, T. Kinnunen, and R. G. Hautamäki, "Improving speaker de-identification with functional data analysis of f0 trajectories," *Speech Commun.*, vol. 140, pp. 1–10, 2022.
- C. O. Mawalim, S. Okada, and M. Unoki, "Speaker anonymization by pitch shifting based on timescale modification," in *Proc. 2nd Symp. Secur. Privacy Speech Commun.*, 2022, pp. 35–42, doi: 10.21437/SPSC.2022-7.
- 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.*, 2009, pp. 3909–3912.
- J. Qian, H. Du, J. Hou, L. Chen, T. Jung, and X.-Y. Li, "Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity," in *Proc. 16th ACM Conf. Embedded Netw. Sensor Syst.*, 2018, pp. 82–94.
- C.-Y. Huang, Y. Y. Lin, H.-Y. Lee, and L.-S. Lee, "Defending your voice: Adversarial attack on voice conversion," in *Proc. Spoken Lang. Technol. Workshop*, 2021, pp. 552–559.
- C. Magarinos, P. Lopez-Otero, L. Docio-Fernandez, E. Rodriguez-Banga, D. Erro, and C. Garcia-Mateo, "Reversible speaker de-identification using pre-trained transformation functions," *Comput. Speech Lang.*, vol. 46, pp. 36–52, 2017.
- B. M.L. Srivastava et al., "Privacy and utility of x-vector based speaker anonymization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2383–2395, 2022.
- S. Meyer, P. Tilli, P. Denisov, F. Lux, J. Koch, and N. T. Vu, "Anonymizing speech with generative adversarial networks to preserve speaker privacy," in *Proc. Spoken Lang. Technol. Workshop*, 2022, pp. 912–919.
- J. Yao, Q. Wang, L. Zhang, P. Guo, Y. Liang, and L. Xie, "NWPU-ASLP system for the VoicePrivacy 2022 challenge," in *Proc. 2nd Symp. Secur. Privacy Speech Commun.*, 2022.
- D. Povey et al., "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proc. Interspeech*, 2018, pp. 3743–3747.
- V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. Interspeech*, 2015, pp. 3214–3218.
- D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5329–5333.
- B. Desplanques, J. Thienpondt, and K. Demuyne, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.
- B. M.L. Srivastava et al., "Design choices for x-vector based speaker anonymization," in *Proc. Interspeech*, 2020, pp. 1713–1717.
- X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 5916–5920.
- J. Kong, J. Kim, and J. Bae, "Hifi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 17022–17033.
- X. Miao, X. Wang, E. Cooper, J. Yamagishi, and N. Tomashenko, "Language-independent speaker anonymization approach using self-supervised pre-trained models," in *Proc. Speaker Lang. Recognit. Workshop*, 2022, pp. 279–286.
- X. Miao, X. Wang, E. Cooper, J. Yamagishi, and N. Tomashenko, "Analyzing language-independent speaker anonymization framework under unseen conditions," in *Proc. Interspeech2022*, pp. 4426–4430.
- X. Chen et al., "System description for voice privacy challenge 2022," in *Proc. 2nd Symp. Secur. Privacy Speech Commun.*, 2022.
- U. E. Gaznepoglu, A. Leschanowsky, and N. Peters, "VoicePrivacy 2022 system description: Speaker anonymization with feature-matched f0 trajectories," in *Proc. 2nd Symp. Secur. Privacy Speech Commun.*, 2022.
- C. O. Mawalim, S. Okada, and M. Unoki, "Speaker anonymization by pitch shifting based on time-scale modification," in *Proc. 2nd Symp. Secur. Privacy Speech Commun.*, 2022, pp. 35–42.
- R. Khamsehashari et al., "Voice privacy - leveraging multi-scale blocks with ECAPA-TDNN SE-Res2NeXt extension for speaker anonymization," in *Proc. 2nd Symp. Secur. Privacy Speech Commun.*, 2022, pp. 43–48.
- P.-G. Noé, J.-F. Bonastre, D. Matrouf, N. Tomashenko, A. Nautsch, and N. Evans, "Speech pseudonymisation assessment using voice similarity matrices," in *Proc. Interspeech*, 2020, pp. 1718–1722.
- P.-G. Noé et al., "Towards a unified assessment framework of speech pseudonymisation," *Comput. Speech Lang.*, vol. 72, 2022, Art. no. 101299.
- S. Ioffe, "Probabilistic linear discriminant analysis," in *Proc. Comput. Vis.-ECCV 9th Eur. Conf. Comput. Vis.*, 2006, pp. 531–542.
- S. E. McAdams, *Spectral Fusion, Spectral Parsing and the Formation of Auditory Images*. Stanford, CA, USA: Stanford Univ., 1984.
- B. M. L. Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, "Evaluating voice conversion-based privacy protection against informed attackers," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 2802–2806.

- [44] C. O. Mawalim, K. Galajit, J. Karnjana, S. Kidani, and M. Unoki, "Speaker anonymization by modifying fundamental frequency and x-vector singular value," *Comput. Speech Lang.*, vol. 73, 2022, Art. no. 101326.
- [45] C. Pierre, A. Larcher, and D. Juvet, "Are disentangled representations all you need to build speaker anonymization systems?," in *Proc. Interspeech*, 2022, pp. 2793–2797.
- [46] A. S. Shamsabadi et al., "Differentially private speaker anonymization," *Proc. Privacy Enhancing Technol.*, vol. 2023, no. 1, Jan. 2023. [Online]. Available: <https://hal.inria.fr/hal-03588932>
- [47] J. M. Perero-Codosero, F. M. Espinoza-Cuadros, and L. A. Hernández-Gómez, "X-vector anonymization using autoencoders and adversarial training for preserving speech privacy," *Comput. Speech Lang.*, vol. 74, 2022, Art. no. 101351.
- [48] H. Turner, G. Lovisotto, and I. Martinovic, "Generating identities with mixture models for speaker anonymization," *Comput. Speech Lang.*, vol. 72, 2022, Art. no. 101318.
- [49] B. v. Niekerk, M.-A. Carbonneau, J. Zaïdi, M. Baas, H. Seuté, and H. Kamper, "A comparison of discrete and soft speech units for improved voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 6562–6566.
- [50] A. Kessy, A. Lewin, and K. Strimmer, "Optimal whitening and decorrelation," *Amer. Statistician*, vol. 72, no. 4, pp. 309–314, 2018.
- [51] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech*, 2011, pp. 249–254.
- [52] Q. Wang, K. A. Lee, and T. Liu, "Scoring of large-margin embeddings for speaker verification: Cosine or PLDA?," in *Proc. Interspeech*, 2022, pp. 600–604.
- [53] A. S. Householder, "Unitary triangularization of a nonsymmetric matrix," *J. ACM*, vol. 5, no. 4, pp. 339–342, 1958.
- [54] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4690–4699.
- [55] X. Xiang, S. Wang, H. Huang, Y. Qian, and K. Yu, "Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2019, pp. 1652–1656.
- [56] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.
- [57] H. Zen et al., "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," 2019, *arXiv:1904.02882*.
- [58] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019. [Online]. Available: <https://datashare.is.ed.ac.uk/handle/10283/3443>
- [59] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "AISHELL-3: A multi-speaker mandarin TTS corpus," in *Proc. Interspeech*, 2021, pp. 2756–2760.
- [60] L. Li et al., "CN-Celeb: Multi-genre speaker recognition," *Speech Commun.*, vol. 137, pp. 77–91, 2022.
- [61] Y. Fan et al., "CN-Celeb: A challenging chinese speaker recognition dataset," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7604–7608.
- [62] M. Ravanelli et al., "SpeechBrain: A general-purpose speech toolkit," 2021, *arXiv:2106.04624*.
- [63] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Proc. 20th Conf. Oriental Chapter Int. Coordinating Committee Speech Databases Speech I/O Syst. Assessment*, 2017, pp. 1–5.
- [64] K. Kasi and S. A. Zahorian, "Yet another algorithm for pitch tracking," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002, vol. 1, pp. 1–361–1–364.
- [65] A. Polyak et al., "Speech resynthesis from discrete disentangled self-supervised representations," in *Proc. Interspeech*, 2021, pp. 3615–3619.
- [66] K. Lakhotia et al., "On generative spoken language modeling from raw audio," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 1336–1354, 2021.
- [67] L. N. Smith, "Cyclical learning rates for training neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2017, pp. 464–472.
- [68] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
- [69] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32.
- [70] L. V. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.
- [71] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of MOS prediction networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 8442–8446.
- [72] P.-G. Noé, X. Miao, X. Wang, J. Yamagishi, J.-F. Bonastre, and D. Matrouf, "Hiding speaker's sex in speech using zero-evidence speaker representation in an analysis/synthesis pipeline," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.



Xiaoxiao Miao (Member, IEEE) received the Ph.D. degree from the Institute of Acoustics, Chinese Academy of Sciences/University Chinese Academy of Sciences, Beijing, China, in 2021. She is currently a Project Researcher with the National Institute of Informatics (NII), Tokyo, Japan. Her research interests include speaker and language recognition, speech security, and machine learning. She is a co-organizer of the latest VoicePrivacy challenge.



Xin Wang (Member, IEEE) received the B.E. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2012, and the M.S. degree from the University of Science and Technology of China, Hefei, China, in 2015, and the Ph.D. degree from SOKENDAI/NII, Japan, in 2018. He is currently a Project Assistant Professor with the National Institute of Informatics (NII), Tokyo, Japan. His research interests include statistical speech synthesis, speech security, and machine learning. He is a co-organizer of the latest ASVspoof and VoicePrivacy challenges.



Erica Cooper (Member, IEEE) received the B.Sc. and M.Eng. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2009 and 2010, respectively, and the Ph.D. degree in computer science from Columbia University, New York, NY, USA, in 2019. She is currently a Project Assistant Professor with the National Institute of Informatics, Tokyo, Japan. Her research interests include statistical machine learning and speech synthesis. Dr. Cooper's awards include the 3rd Prize in the CSAW

Voice Biometrics and Speech Synthesis Competition, the Computer Science Service Award from Columbia University, and the Best Poster Award in the Speech Processing Courses in Crete.



Junichi Yamagishi (Senior Member, IEEE) received the Ph.D. degree from the Tokyo Institute of Technology, Tokyo, Japan, in 2006. From 2007 to 2013, he was a Research Fellow with the Centre for Speech Technology Research (CSTR) at the University of Edinburgh, Edinburgh, U.K. He was appointed as an Associate Professor with the National Institute of Informatics (NII), Tokyo, Japan, in 2013. He is currently a Professor with NII. His research interests include speech processing, machine learning, signal processing, biometrics, digital media cloning, and

media forensics. He was an co-organizer for the bi-annual ASVspoof Challenge and the bi-annual Voice Conversion Challenge. He was also a member of the IEEE Speech and Language Technical Committee during 2013–2019, an Associate Editor of the IEEE/ACM TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING during 2014–2017, a Chairperson of ISCA SynSIG during 2017–2021, and the Senior Area Editor of the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING during 2019–2023. He is currently a PI of JST-CREST and ANR supported VoicePersonae Project.



Natalia Tomashenko (Member, IEEE) received the Ph.D. degree in computer science from the University of Le Mans, Le Mans, France. She is currently a Researcher with the University of Avignon, Avignon, France. Her research interests include statistical machine learning for speech and language processing with application to automatic speech and speaker recognition, spoken language understanding, machine translation, and speech privacy. She is an organizer of the VoicePrivacy challenge.