The Speaker and Language Recognition Workshop (Odyssey 2022)
28 June - 1 July 2022, Beijing, China

Language-Independent Speaker Anonymization Approach using Self-Supervised Pre-Trained Models

Xiaoxiao Miao[1], *Xin Wang*[1], *Erica Cooper*[1], *Junichi Yamagishi*[1], *Natalia Tomashenko*[2]

[1]National Institute of Informatics, Japan [2]LIA, University of Avignon, France

xiaoxiaomiao@nii.ac.jp

Abstract

Speaker anonymization aims to protect the privacy of speakers while preserving spoken linguistic information from speech. Current mainstream neural network speaker anonymization systems are complicated, containing an F0 extractor, speaker encoder, automatic speech recognition acoustic model (ASR AM), speech synthesis acoustic model and speech waveform generation model. Moreover, as an ASR AM is language-dependent, trained on English data, it is hard to adapt it into another language. In this paper, we propose a simpler self-supervised learning (SSL)-based method for language-independent speaker anonymization without any explicit language-dependent model, which can be easily used for other languages. Extensive experiments were conducted on the VoicePrivacy Challenge 2020 datasets in English and AISHELL-3 datasets in Mandarin to demonstrate the effectiveness of our proposed SSL-based language-independent speaker anonymization method[1].

1. Introduction

It is well known that speech data contains a plethora of privacy information, such as speaker identity, age, emotion, gender, and so on. The speaker identities of speech recordings without any protection could be resynthesized, cloned, and converted by using advanced speech synthesis (SS) technologies, which may lead to a privacy risk [1]. In fact, speaker anonymization is a way to conceal speaker information while maintaining intelligibility and naturalness as much as possible [2–5].

In recent decades, voice protections have mainly focused on noise addition, voice transformation, conversion, synthesis and disentangled representation learning [3, 6–12]. Efforts to study the trade-off between speaker identity and speech intelligibility by adding noise to test data have also been made. The goal of voice transformation, conversion, synthesis, and disentangled representation learning methods is to synthesize various speakers' voices. This can be achieved by modifying the non-linguistic information in speech, such as F0, energy, or speaker vector, or by producing a new speaker voice with generative models.

To define the speaker anonymization problem accurately and fairly, the VoicePrivacy Challenge (VPC) 2020 [4, 5] has provided common datasets, evaluation metrics (objective and subjective), and two main baselines to suppress the individuality of speech, leaving other speaker attribute information such as age and gender, while preserving the naturalness of speech. The primary baseline of VPC 2020 consists of several components: an F0 extractor, pre-trained time delay neural network (TDNN) x-vector automatic speaker verification (ASV) system [13], pre-trained factorized time delay neural network (TDNN-F) automatic speech recognition acoustic model (ASR AM) [14, 15], speech synthesis acoustic model (SS AM), and neural source-filter (NSF) [16] waveform model, denoted as B1. Although the system can effectively anonymize the speech data, it is complicated to build and deploy. Moreover, the ASR AM of this system is language-dependent and requires language-specific resources. The system cannot be directly applied to another language. The second baseline, which is denoted as B2, is based on conventional signal processing techniques and does not require training data. However, it has been demonstrated that B2 cannot protect the speaker information as well as the neural-network-based systems [4, 5].

Recent success in self-supervised learning (SSL) has shown remarkable performance for speech synthesis tasks [17–19], One of the new speech resynthesis works [18] learned discrete speech units from the unsupervised clustering of speech representations using a large unlabeled speech corpus. The idea of discretized speech units is to disentangle speech representations, separating content from speaker information. In addition, a number of researchers [20, 21] have found that incorrect quantization predictions from discretization will cause mispronunciation problems.

In this paper, we propose a new SSL-based language-independent neural speaker anonymization method. We first introduce soft content representations [21] instead of discretized ones to mitigate ambiguous mispronunciations in the anonymized speech. Next, as VPC 2020 has shown that x-vectors are important to encode speaker identity information for the speaker anonymization task, we update the most frequently used d-vector [18, 22] or TDNN x-vector speaker encoders [4, 5, 13] to the state-of-the-art ECAPA-TDNN [23]. More importantly, the current VPC 2020 primary baseline B1 requires large amounts of language-specific resources to train a language-dependent ASR AM, while the SSL-based soft content encoder of the proposed method learns universal representations by training with unlabeled audio data, which improves the portability to a new language. Extensive experiments were conducted on the VPC 2020 datasets in English and AISHELL-3 [24] datasets in Mandarin to demonstrate the effectiveness of our proposed SSL-based language-independent speaker anonymization method.

In the remainder of this paper, Section 2 overviews the speaker anonymization baseline systems of VPC 2020. Section 3 describes our proposed SSL-based language-independent speaker anonymization method, Section 4 evaluates the system, and Section 5 concludes the paper.

[1]English and Mandarin Audio samples and source code are available at https://github.com/nii-yamagishilab/SSL-SAS

10.21437/Odyssey.2022-39

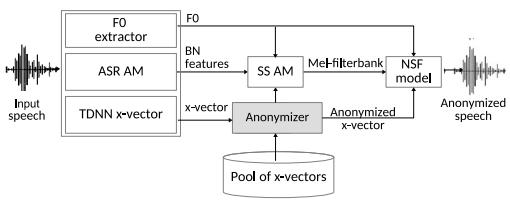

Figure 1: Architecture of the VPC 2020 baseline B1 system.

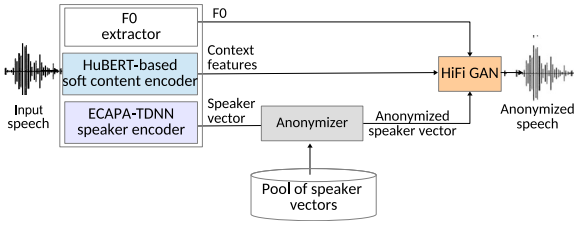

Figure 2: Architecture of the proposed system. The main differences between the B1 system and proposed system are highlighted in color.

2. Speaker Anonymization Baselines

2.1. Baseline using x-vector and Neural Waveform Models

The VPC 2020 primary baseline aims to decompose speech into speaker identity, fundamental frequency, and linguistic information, which is illustrated in Figure 1 and is composed of three main procedures:

1) Original F0, bottleneck (BN), and x-vector extraction. There are two pre-trained and fixed encoders, including a TDNN-F ASR AM [14, 15] trained on the *LibriSpeech-train-clean-100* and *LibriSpeech-train-other-500* datasets [25] to extract 256-dimensional BN features as content representation, and a TDNN x-vector model trained on the *VoxCeleb-1 & 2* [26, 27] datasets to extract 512-dimensional x-vector as the speaker identity representation.

2) x-vector anonymization. A pseudo x-vector is obtained by searching the 200 furthest same-gender x-vectors from an external x-vector pool (*LibriTTS-train-other-500* [28]), according to probabilistic linear discriminant analysis (PLDA) [13] distances. Then averaging 100 randomly-selected ones [29].

3) Anonymized speech synthesis. An SS AM first generates a Mel-filterbank from the anonymized x-vector, original F0, and BN features. The generated Mel-filterbank, F0, and anonymized x-vector are fed into an NSF [16] model to synthesize the anonymized waveform. Both SS AM and NSF models are trained with *LibriTTS-train-clean-100* [28].

2.2. Baseline using McAdams Coefficient

In short-time speech analysis, the power spectrum of a short-term segment of speech is fit with an all-pole model using linear predictive coding analysis. The formant frequencies are determined by the angles of the corresponding complex poles. It is known that the McAdams coefficient can be used to alter the pole angles [30, 31]. Thus, in the second baseline of VPC 2020, the speaker anonymization process was achieved by varying the pole angles to shift the formants using the McAdams coefficient.

3. Proposed Method using SSL Models

In this section, we describe the proposed SSL-based language-independent speaker anonymization system as illustrated in Figure 2 and a more detailed structure is shown in Figure 3. We can see the differences clearly between B1 and the proposed SSL-based system from Figure 2, which consists of two pre-trained and fixed encoders: a HuBERT-based soft content encoder E_c, ECAPA-TDNN speaker encoder E_{spk}, one F0 extractor and one decoder HiFi-GAN neural vocoder [32]. These approaches have been successfully applied to the different tasks. We are the first to reassemble them for the purpose of the speaker anonymization task.

The HuBERT-based soft content encoder [21] is obtained by fine-tuning a pretrained HuBERT Base model [33], learning finer grained continuous context representations from discrete units. The ECAPA-TDNN speaker encoder provides an utterance-level discriminative speaker-related representation. We follow the same x-vector anonymization scheme of the VPC baseline B1 system [4, 5] to anonymize a speaker vector of each source input utterance. For the F0 extractor, the YAAPT algorithm [34] is used to extract F0 from the input signal. Then the content features, F0 and anonymized speaker vector are fed into the HiFi-GAN neural vocoder after upsampling and concatenating to generate the anonymized speech.

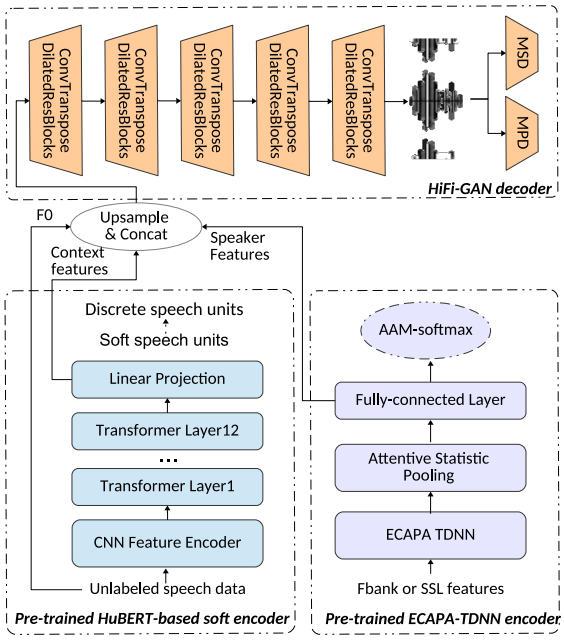

Figure 3: Language-independent speaker anonymization system using SSL models showing expanded detail of the combined structure.

3.1. HuBERT-based Soft Content Encoder

Many self-supervised speech models have been proposed in previous studies. It has been observed that different layers of SSL-based models contain different information like speaker identity, content and semantics [35–38]. In particular, higher and middle layers of SSL-based models tend to capture richer linguistic information. Replacing the typical ASR bottleneck features with SSL-based content features has the advantage of eliminating the need for ASR resources such as phone labels, phonesets, lexicons, and language models.

This study mainly focuses on HuBERT [33]. It was trained on the partially masked frames using pseudo labels, which can be obtained and refined by iterative clustering processes. There

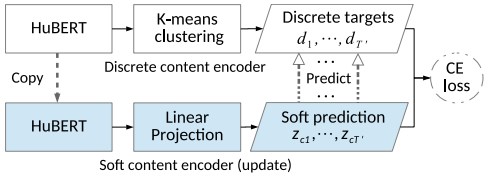

Figure 4: Outline of HuBERT-based soft content encoder.

are two common ways to use a pre-trained HuBERT model in speech synthesis. One is directly using the continuous output features of the HuBERT model [38]. However, it has been proven that the continuous representations contain both context and speaker information, and are thus not suitable for speech disentanglement. Another work [18] applied a k-means algorithm over HuBERT continuous representations to obtain discrete speech units, getting rid of speaker identity attributes. While inaccurate discrete features are known to lead to incorrect pronunciation, and we confirm the issue in our experiments later as well.

To compromise a trade-off between continuous representations and discrete speech units, we implement the HuBERT-based soft content encoder that is similar to [21] to capture soft content features by predicting a distribution over discrete speech units. These soft content features are expected to represent more accurate content information while suppressing speaker information effectively. Figure 4 presents the training process of the HuBERT-based soft content encoder. Given a sequence of an input recording $\mathbf{x} = (x_1, \cdots, x_T)$, where $x_t \in \mathcal{R}$, the output of the soft content E_c is the sequence of soft prediction $\mathbf{z}_c = (\mathbf{z}_{c1}, \cdots, \mathbf{z}_{cT'})$, where $T'/T = 1/320$ is determined by the convolutional neural network (CNN) stride of HuBERT. During the training stage, given an input utterance, the pre-trained HuBERT with a k-means clustering model is firstly introduced as the fixed model to extract discrete speech units $(d_1, \cdots, d_{T'})$ as targets. Then, the k-means clustering strategy of the pre-trained HuBERT is replaced with a linear projection layer that transforms the HuBERT output to a sequence of soft prediction $\mathbf{z}_c = (\mathbf{z}_{c1}, \cdots, \mathbf{z}_{cT'})$. The parameterized distribution of each soft unit over the dictionary of discrete speech units can be defined in Equation 1:

$$p(d_{t'} = i \mid \mathbf{z}_{ct'}) = \frac{\exp(\text{sim}(\mathbf{z}_{ct'}, \mathbf{w}_i)/\tau)}{\sum_{k=1}^{K} \exp(\text{sim}(\mathbf{z}_{ct'}, \mathbf{w}_k)/\tau)}, \quad (1)$$

where i denotes the cluster index of the i^{th} discrete target, $\mathbf{w}_i$ is a corresponding trainable embedding vector, $\text{sim}(\cdot, \cdot)$ measures the cosine similarity, and τ is a temperature parameter set as 0.1 by the experimental selection. The final step is to update the soft content (including HuBERT) by minimizing the cross-entropy (CE) loss function, which measures the distance between the distribution and discrete targets $(d_1, \cdots, d_{T'})$.

3.2. ECAPA-TDNN Speaker Encoder

Current neural network ASV systems mainly consist of input feature extraction, a deep frame-level feature extractor, pooling layer, and loss function. For input feature extraction, the most commonly used features are Mel-filterbank (FBank) and Mel-frequency cepstrum coefficients (MFCC). Recent works have shown the representations derived from different layers of SSL-based models also contain different levels of discriminative speaker-related information [37]. Specifically, lower layers of SSL-based models have more speaker information than higher layers. For the deep feature extractor, TDNN-based [13,23] and CNN-based [39,40] systems have been successfully used in the field of ASV.

In this paper, we employ the state-of-the-art ECAPA-TDNN [23] as speaker encoder E_s to encode speaker identity information, $\mathbf{z}_{spk} = E_s(\mathbf{x})$, $\mathbf{z}_{spk} \in \mathbb{R}^{192}$. ECAPA-TDNN is an improved version of the original TDNN x-vector. The deep frame-level feature extractor of ECAPA-TDNN includes squeeze-excitation blocks, skip connections, and multi-scale/multi-layer feature aggregation to explore discriminative speaker features and a channel-dependent attentive statistics pooling layer to capture fixed-dimensional representation from variable-length features. In addition to using the commonly used FBank as input features, we also explore the effect of SSL-based input features for the ECAPA-TDNN in the proposed speaker anonymization framework.

3.3. F0 Extractor

We use the YAAPT algorithm [34] to extract F0 from the input signal $\mathbf{x}$. Its F0 sequence is denoted as $\mathbf{z}_f = (\mathbf{z}_{f1}, \cdots, \mathbf{z}_{fT''})$. The ratio of $T''/T = 1/160$ in our experiment.

3.4. HiFi-GAN Neural Vocoder

The frame-wise context $\mathbf{z}_c$ and F0 $\mathbf{z}_f$ sequences are up-sampled and concatenated. The segmental-level speaker embedding $\mathbf{z}_{spk}$ is then integrated to each frame in the up-sampled sequence, namely the encoded representation $\mathbf{z} = (\mathbf{z}_c, \mathbf{z}_f, \mathbf{z}_{spk})$, which is then passed to the neural vocoder HiFi-GAN [32] to generate the speech waveforms.

Unlike the two-stage pipeline strategy in VPC 2020 baseline B1, there is no SS AM before the neural vocoder. The reason for excluding SS AM without generating a Mel-filterbank further is that the content features derived from the HuBERT-based soft context encoder contain rich linguistic information, compared with BN features extracted by ASR AM. In addition, HiFi-GAN is known as a non autoregressive model that has been successfully applied to speech synthesis, achieving both high computational efficiency and sample quality.

HiFi-GAN consists of a generator and two discriminators. The generator has five groups of ResBlock where multiple transposed convolutions upsample low-frequency encoded representation $\mathbf{z}$ to the original audio size, followed by a stack of dilated residual connections to increase the receptive field. During training, the generated audio sample $\hat{\mathbf{x}}$, is passed to two discriminators. The multi-period discriminator (MPD) captures the periodic patterns of audio through five-period sub-discriminators operating on equally spaced samples between the original and generated waveforms. The multi-scale discriminator (MSD) has the advantage of exploring long-range and consecutive interactions of audio by multi-scale average pooling operations. The configuration of HiFi-GAN is the same as [18]. The overall training losses of HiFi-GAN involve a generator loss $\mathcal{L}_G$ and a discriminator loss $\mathcal{L}_D$:

$$\mathcal{L}_G = \sum_{k=1}^{K} \left[\mathcal{L}_{Adv}(G; D_k) + \lambda_{fm} \mathcal{L}_{FM}(G; D_k) \right] + \lambda_{mel} \mathcal{L}_{Mel}(G) \quad (2)$$

$$\mathcal{L}_D = \sum_{k=1}^{K} \mathcal{L}_{Adv}(D_k; G) \quad (3)$$

where D_k denotes the k-th sub-discriminator in the MPD and MSD. In the experiment, we set $\lambda_{fm} = 2$ and $\lambda_{mel} = 45$ to

balance the adversarial losses, the feature matching loss $\mathcal{L}_{FM}$, and mel-spectrogram loss $\mathcal{L}_{Mel}$:

$$\mathcal{L}_{FM}(G; D_k) = \mathbb{E}_{(\mathbf{x},\hat{\mathbf{x}})}\left[\sum_{i=1}^{L} \frac{|D_k^i(\mathbf{x}) - D_k^i(G(\hat{\mathbf{x}}))|_1}{N_i}\right] \quad (4)$$

$$\mathcal{L}_{Mel}(G) = \mathbb{E}_{(\mathbf{x},\hat{\mathbf{x}})}\left[|\phi(\mathbf{x}) - \phi(G(\hat{\mathbf{x}}))|_1\right] \quad (5)$$

where L denotes the number of layers in the discriminator. D_k^i is the i-th layer feature map of the k-th sub-discriminator, and N_i indicates the number of units in the i-th layer. ϕ is a spectral operator computing a mel-spectrogram from a given waveform.

4. Evaluation

In this section, we first test our ECAPA-TDNN-based speaker encoders separately from the anonymization task, using the well explored Voxceleb dataset [26, 27] to show that they have reasonable performance. Then we followed the VPC 2020 evaluation plan [4, 5] to demonstrate that the proposed anonymization system can achieve comparable performance to VPC 2020 baselines on English data. Finally, we conducted anonymization experiments on Mandarin data to show that the proposed system without any explicit language-dependent component can be applied to a different language.

4.1. ASV Experiments

As described in Section 3.2, we built two ECAPA-TDNN-based speaker encoders that can be plugged into our anonymization system. Both of them followed the recipe in the original ECAPA literature [23], but their input features are different:

- F-ECAPA: 80-dimensional FBank with 25ms window size and 10ms frame shift.
- S-ECAPA: 768-dimensional weighted average of the output features from all the hidden layers of a pre-trained HuBERT Base model released by Fairseq toolkit[2]. The parameters of HuBERT Base were updated when training S-ECAPA.

F-ECAPA used the same network typology as the original ECAPA with 512 channels in the convolution frame layers [23]. The training loss was the additive angular margin (AAM) loss [41, 42] with a margin of 0.2. S-ECAPA also used the same typology except for the first layer due to the different input feature dimensions. Both models used the 192-dimensional output from the last linear projection layer as the speaker embedding.

We trained both models using the Voxceleb2 dev set [27], which contains over 1 million utterances from 5,994 speakers. Each audio sample was cropped to segments with a maximum duration of 3s during training. Extra room impulse response data[3] was used for training data augmentation, while voice activity detection (VAD) was not used. We then solely evaluated the trained models on the Voxceleb1 test set [26] for the ASV task, which contains 4,872 utterances from 40 speakers. The evaluation metrics include the equal error rate (EER) and the minimum of the decision cost function (minDCF) calculated using $C_{FA} = C_{Miss} = 1$, and $P_{target} = 0.01$, as in [23]. This ASV experiment was conducted using the SpeechBrain toolkit [43].

[2] https://github.com/pytorch/fairseq/

[3] https://www.openslr.org/28/

Table 1: ECAPA-TDNN-based speaker encoder included in the proposed system was solely evaluated for the ASV task. The speaker encoders were trained on the Voxceleb2 dev set and evaluated on the Voxceleb1 test set. Lower EER is better for the ASV task.

system	**EER** [%]	**minDCF**
TDNN x-vector [42]	2.23	-
ECAPA [23]	1.01	0.127
F-ECAPA	1.10	0.135
S-ECAPA	0.87	0.123

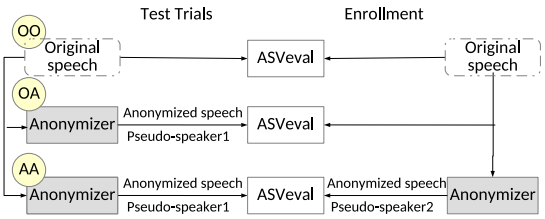

Figure 5: Objective speaker verifiability evaluation scenarios.

Table 1 compares the EER and minDCF results from our speaker encoders with those reported in the literature on TDNN x-vector and ECAPA. Our F-ECAPA performed similarly to the original ECAPA and outperformed the TDNN x-vector by a large margin. Compared with F-ECAPA, the S-ECAPA using the input features extracted from the HuBERT model achieved better results. Since both F-ECAPA and S-ECAPA outperformed the TDNN xvector, which was used in the VPC 2020 primary baseline, we decided to use the two new speaker encoders in the following speaker anonymization experiments.

4.2. Speaker Anonymization Experimental Setup

4.2.1. Evaluation Plan

We followed the VPC 2020 evaluation plan [4,5] for our speaker anonymization experiments. The only difference is that the content encoders of the proposed systems used SSL models pretrained on an external database not included in the VPC 2020 protocol (i.e., *LibriSpeech-train-960*). Details on the content encoders are explained in the next section..

The evaluation plan treated the speaker anonymization task as a game between users and attackers. It assumes that the attackers have access to an ASV model (which is referred to as ASV_{eval}) and a few enrollment trials for each speaker. The attackers then use these resources to identify the speaker identity in anonymized test trials. Specifically, the evaluation plan defines three attack scenarios, and the enrollment and test trials in each scenario are either original (O) or anonymized (A). The three scenarios are illustrated in Figure 5 and listed as follows:

- *Unprotected (OO)*: no anonymization is applied, and attackers verify the original test trials against the original enrollment trials;
- *Ignorant attacker (OA)*: users anonymize their trial data while attackers are unaware and simply use original enrollment data;
- *Lazy-informed (AA)*: users anonymize their trial data, and attackers partially know which anonymizer was used. Attackers use the same anonymizer to anonymize the enrollment data, hoping that it can be better linked with the anonymized trial data from the same target

Table 2: EER (%) on English data (development set of VPC 2020) evaluated by ASV_{eval}^{eng}. The EER results of the OO condition for Libri-female, Libri-male, VCTK-diff-female, and VCTK-diff-male are 8.66%, 1.24%, 2.86%, and 1.43%, respectively. Higher EER indicates better privacy.

Development set	Libri-female		Libri-male		VCTK-diff-female		VCTK-diff-male	
Anony. Type	OA	AA	OA	AA	OA	AA	OA	AA
VPC2020 B1 [4]	50.14	36.79	57.76	34.16	49.97	26.11	53.95	30.92
S-ECAPA + HuBERT	27.84	21.59	7.76	13.66	28.52	11.68	23.28	16.18
S-ECAPA + HuBERT-km	49.43	37.38	48.45	39.29	51.15	26.78	49.68	30.47
S-ECAPA + HuBERT-soft	48.15	35.65	43.32	35.71	51.09	27.40	48.24	30.52
F-ECAPA + HuBERT-soft	47.44	29.55	46.72	34.78	52.11	24.09	48.04	29.48

Table 3: EER (%) on English data (test set of VPC 2020) evaluated by ASV_{eval}^{eng}. The EER results of the OO condition for Libri-female, Libri-male, VCTK-diff-female, and VCTK-diff-male are 7.66%, 1.11%, 4.88%, and 2.06%, respectively. Higher EER indicates better privacy.

Test set	Libri-female		Libri-male		VCTK-diff-female		VCTK-diff-male	
Anony. Type	OA	AA	OA	AA	OA	AA	OA	AA
VPC2020 B1 [4]	47.26	32.12	52.12	36.75	48.05	31.74	53.85	30.94
S-ECAPA + HuBERT	20.99	14.42	12.92	14.52	26.03	16.72	17.57	16.30
S-ECAPA + HuBERT-km	46.17	36.13	44.99	38.98	48.51	33.64	52.81	32.61
S-ECAPA + HuBERT-soft	41.42	31.02	39.87	36.97	48.10	31.22	47.59	35.94
F-ECAPA + HuBERT-soft	41.24	29.74	42.54	33.18	50.31	29.32	48.11	30.71

speaker. However, attackers do not know the detailed parameters and anonymized the enrollment data using different pseudo-speakers.

The ASV_{eval} EER in *Unprotected (OO)* serves as the reference where no anonymization is applied. The EERs from the other two scenarios measure how well an anonymization system protects the speaker information in the test trials when facing attackers with varied amounts of prior knowledge on the anonymization system. Ideally, the EERs should be as high as 50% in both OA and AA scenarios.

The evaluation plan also uses the word error rate (WER) as a utility metric to measure how well the speech content is preserved after anonymization. The WER was computed using an ASR model (ASR_{eval}). An ideal anonymization system should not increase the WER of test trials after anonymization. Note that both ASR_{eval} and ASV_{eval} are independent from the anonymization system. For the experiments on different languages, we use different ASR_{eval} and ASV_{eval} to evaluate the anonymization systems more accurately. Details are explained in the following sections.

Another utility metric is the naturalness of anonymized audios. We used a recently proposed mean opinion score (MOS) prediction network [44] to estimate perceived naturalness rather than conducting time-consuming listening tests.

4.2.2. System Configurations

We built two versions of the proposed anonymization system. While both used the HuBERT-soft content encoder, they use either F-ECAPA or S-ECAPA as the speaker encoder. For reference, we built another two anonymization systems using S-ECAPA and slightly different HuBERT-based content encoders[4].

The first reference system used a pre-trained HuBERT Base model from the Fairseq toolkit. Given an input waveform, this HuBERT model extracts a sequence of 768-dimensional vectors from the output of the sixth Transformer layer as the content representations [18]. The second reference system used the HuBERT-km Base model released in the same toolkit. This model was based on HuBERT, but additional k-means clustering was applied to learn 200 clusters on *LibriSpeech-train-clean-100*. During inference, each continuous-valued content representation was quantized, and the corresponding 200-dimensional cluster centroid was used as the content representation.

Finally, the HuBERT-soft in the proposed systems used HuBERT as the backbone and the index of the quantized content representation from HuBERT-km as the target. It was fine-tuned on *LibriSpeech-train-clean-100* using the strategy in Section 3.1. Each content vector from HuBERT-soft has 200 dimensions.

All the anonymization systems used the same YAAPT F0 extractor. Given the F0, content, and speaker vectors from the corresponding encoders, a HiFi-GAN was trained on *LibriTTS-train-clean-100* for each anonymization system. Noted we use cosine distance to generate the pseudo speaker unlike the B1 system using PLDA scores.

[4]We also tested wav2vec 2.0 as the content encoder and observed similar results to those using HuBERT. We thus only present the results of HuBERT-based systems in this paper.

Table 4: WER (%) on English data evaluated by ASR_{eval}^{eng}. Lower WER indicates better utility.

	Libri.		VCTK	
Anonymization system	Dev.	Test	Dev.	Test
Ground Truth [4]	3.83	4.15	10.79	12.82
VPC 2020 B1 [4]	6.39	6.73	15.38	15.23
S-ECAPA + HuBERT	4.23	4.47	12.12	13.89
S-ECAPA + HuBERT-km	7.84	7.80	19.21	21.74
S-ECAPA + HuBERT-soft	4.47	4.70	12.88	14.57
F-ECAPA + HuBERT-soft	4.50	4.69	12.96	14.86

4.3. Speaker Anonymization Experiments in English

We evaluated the anonymization systems on the official development and test sets of VPC 2020. These two sets contain English utterances of several female and male speakers from the *LibriSpeech* and VCTK [45] corpora. To compute the WER and EER, we used the language-matched ASR_{eval} and ASV_{eval} systems provided by VPC 2020 [4, 5] and denoted them as $ASR_{\text{eval}}^{\text{eng}}$ and $ASV_{\text{eval}}^{\text{eng}}$, respectively. They were trained on the *LibriSpeech-train-clean-360* English dataset.

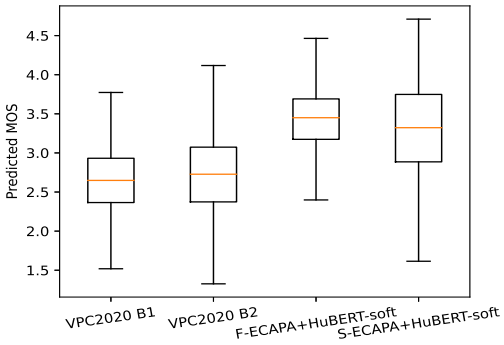

Figure 6: Box-plots on predicted naturalness scores of anonymized speech from experimental systems.

Tables 2 and 3 show the EER results of various speaker anonymization systems, measured by ASV_{eval}^{eng} on the VPC 2020 development and test sets, respectively. Table 4 shows their WER results measured by ASR_{eval}^{eng}. VPC 2020 baseline B2 was not included in the objective evaluation as its objective results are worse than B1 [5]. Here is the point-by-point summary:

Proposed system vs. B1: From the results, the first observation is that compared with VPC 2020 B1, our proposed systems, "F-ECAPA + HuBERT-soft," and "S-ECAPA + HuBERT-soft" (the last two rows), which do not require any language-specific modules, have achieved comparable EER values on both OA and AA conditions and have lower WERs than those of B1.

Soft content encoder: The WER and EER results of "S-ECAPA + HuBERT" show that although the continuous representations extracted from "S-ECAPA + HuBERT" achieved low WERs, ASV EERs are far from 50%. As expected, the continuous representations without k-means clustering do not have proper disentanglement of content and speaker, and this prevents the generation of properly speaker-anonymized speech. Next, for "S-ECAPA + HuBERT-km," we can see that the k-means clustering process managed to suppress speaker information and increased the ASV EERs of speaker-anonymized speech significantly whereas it also increased the ASR WERs, too. As hypothesized, the hard clustering process seems to cause pronunciation issues. Finally, we can see that the proposed "S-ECAPA + HuBERT-soft" system strikes a good balance and speaker-anonymized speech has a low ASR WER and high ASV EER. **S-ECAPA vs. F-ECAPA:** Although not the main focus, it is also meaningful to compare acoustic features used for our speaker encoder. Interestingly, "S-ECAPA + HuBERT-soft" always has higher EER values than "F-ECAPA + HuBERT-soft" on all AA conditions (where attackers have more priors).

To further analyze the effectiveness of our proposed models, we evaluate the naturalness of speaker-anonymized speech based on the MOS prediction network. Box-plots of the predicted MOS scores are shown in Figure 6. From the figure, we can observe a similar trend to the subjective results of VPC 2020 in which B2 had higher naturalness scores than B1 [5]. Moreover we can see that the proposed "S-ECAPA + HuBERT-soft" and "F-ECAPA + HuBERT-soft" are expected to have higher naturalness than the B1 and B2 systems.

Overall, the proposed simpler "ECAPA + HuBERT-soft" anonymization systems without any language-specific models can protect the speaker information almost as well as the VPC 2020 primary baseline, and provide reliable linguistic information among all the anonymization systems.

Table 5: EER (%) on Mandarin data evaluated by ASV_{eval}^{mand}. Higher EER indicates better privacy.

EER(%)	OO	OR	OA	AA
F-ECAPA-HuBERT-soft	2.04	9.13	37.58	22.98
S-ECAPA-HuBERT-soft	2.04	13.45	40.81	23.26

Table 6: CER (%) on Mandarin data evaluated by ASR_{eval}^{mand}. Lower CER indicates better utility.

CER(%)	Syn. type	Test set
Ground Truth	-	10.36
F-ECAPA + HuBERT-soft	resyn.	14.81
F-ECAPA + HuBERT-soft	anony.	18.86
S-ECAPA + HuBERT-soft	resyn.	14.95
S-ECAPA + HuBERT-soft	anony.	21.18

4.4. Speaker Anonymization Experiments in Mandarin

We directly used the proposed same anonymization systems from Section 4.3 on a Mandarin speaker anonymization task without training or fine tuning them on Mandarin data. We evaluated the systems in the *Unprotected (OO)*, *Ignorant attacker (OA)*, *Lazy-informed (AA)*, and a new scenario called *OR*. In this OR scenario, the enrollment is original and the test trial is resynthesized by the anonymization system using the original speaker vector [46]. An ideal system thus should not change speaker information or content of the test trial in the OR scenario, and the WER and EER should be as low as those from the OO scenario.

The evaluation was conducted on a test set sampled from a multi-speaker Mandarin speech corpus called AISHELL-3 [24]. The test set contains 4,267 utterances from 44 speakers. We split them into test trial (88 utterances) and enrollment (4179 utterances) subsets, which are used to produce 10,120 enrollment-test trials for the ASV evaluation. The ASV evaluation model ASV_{eval}^{mand} was a F-ECAPA trained on Mandarin datasets called *CN-Celeb-1 & 2* [47, 48]. ASV_{eval}^{mand} was configured in the same way as Section 4.1 described. The ASR evaluation model ASR_{eval}^{mand} was a publicly available ASR Transformer [43] trained on the 150-hour Mandarin ASR dataset AISHELL-1 [49]. Note that the WER was replaced with character error rate (CER) in Mandarin ASR.

Tables 5 and 6 list the EERs and CERs, respectively. We first observe that both proposed systems increased the ASV EER to around 40% in the OA scenario and 23% in the AA scenario, which suggests that speaker information is protected to an acceptable degree. Meanwhile, the CERs on the anonymized trials were increased to about 20%, suggesting degradation on the speech content. Interestingly, the CERs on anonymized trials are higher than those on resynthesized trials for both proposed systems. The increased CER is likely to be caused by the pseudo x-vectors obtained from the English x-vector pool, while used to generate Mandarin speech. The previous study [50, 51] has proven that in addition to speaker-related information, x-vectors also encode speaker rate, background conditions and lexical content information. Fixing the data mismatch is one potential way to improve our proposed anonymization systems. Finally, experiments on Mandarin an English share a similar trend that compared with "F-ECAPA + HuBERT-soft," "S-ECAPA + HuBERT-soft" tends to sacrifice some linguistic content to protect the speaker information, obtaining a higher EER on OR, OA, and AA conditions but a worse CER on both resynthesis and anonymized trials.

5. Conclusion

This paper proposed an SSL-based language-independent speaker anonymization method, which consists of a HuBERT-based soft content encoder, ECAPA-TDNN speaker encoder, F0 encoder and HiFi-GAN decoder. Experiments on English VPC 2020 datasets and Mandarin AISHELL-3 datasets demonstrated that our proposed speaker anonymization method without any explicit language-specific resources, can be adopted to other languages successfully. Future work will focus on building a more robust speaker anonymization method so that it can better process input trials recorded in an acoustic condition with unseen conditions.

Acknowledgment This study is supported by JST CREST Grants (JPMJCR18A6 and JPMJCR20D3), MEXT KAKENHI Grants (21K17775, 21H04906, 21K11951, 18H04112), and the VoicePersonal project (ANR-18-JSTS-0001).

6. References

[1] Ville Vestman, Tomi Kinnunen, Rosa González Hautamäki, and Md Sahidullah, "Voice mimicry attacks assisted by automatic speaker verification," *Computer Speech & Language*, vol. 59, pp. 36–54, 2020.

[2] Qin Jin, Arthur R Toth, Alan W Black, and Tanja Schultz, "Is voice transformation a threat to speaker identification?," in *Proc. ICASSP*, 2008, pp. 4845–4848.

[3] Fuming Fang, Xin Wang, Junichi Yamagishi, Isao Echizen, Massimiliano Todisco, Nicholas Evans, and Jean-Francois Bonastre, "Speaker anonymization using x-vector and neural waveform models," *Proc. 10th ISCA Speech Synthesis Workshop*, pp. 155–160, 9 2019.

[4] N. Tomashenko, Brij Mohan Lal Srivastava, Xin Wang, Emmanuel Vincent, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Jose Patino, Jean-François Bonastre, Paul-Gauthier Noé, and Massimiliano Todisco, "Introducing the VoicePrivacy Initiative," in *Proc. Interspeech*, 2020, pp. 1693–1697.

[5] Natalia Tomashenko, Xin Wang, Emmanuel Vincent, Jose Patino, Brij Mohan Lal Srivastava, Paul-Gauthier Noé, Andreas Nautsch, Nicholas Evans, Junichi Yamagishi, Benjamin O'Brien, et al., "The VoicePrivacy 2020 challenge: Results and findings," *Computer Speech & Language*, 2022.

[6] Kei Hashimoto, Junichi Yamagishi, and Isao Echizen, "Privacy-preserving sound to degrade automatic speaker verification performance," in *Proc. ICASSP*, 2016, pp. 5500–5504.

[7] Jianwei Qian, Haohua Du, Jiahui Hou, Linlin Chen, Taeho Jung, Xiangyang Li, Yu Wang, and Yanbo Deng, "Voicemask: Anonymize and sanitize voice input on mobile devices," *ArXiv*, vol. abs/1711.11460, 2017.

[8] Qin Jin, Arthur R Toth, Tanja Schultz, and Alan W Black, "Voice convergin: Speaker de-identification by voice transformation," in *Proc. ICASSP*, 2009, pp. 3909–3912.

[9] Jianwei Qian, Haohua Du, Jiahui Hou, Linlin Chen, Taeho Jung, and Xiang-Yang Li, "Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity," in *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, 2018, pp. 82–94.

[10] Chien-yu Huang, Yist Y Lin, Hung-yi Lee, and Lin-shan Lee, "Defending your voice: Adversarial attack on voice conversion," in *Proc. Spoken Language Technology Workshop (SLT)*, 2021, pp. 552–559.

[11] Carmen Magarinos, Paula Lopez-Otero, Laura Docio-Fernandez, Eduardo Rodriguez-Banga, Daniel Erro, and Carmen Garcia-Mateo, "Reversible speaker de-identification using pre-trained transformation functions," *Computer Speech & Language*, vol. 46, pp. 36–52, 2017.

[12] Brij Mohan Lal Srivastava, Aurélien Bellet, Marc Tommasi, and Emmanuel Vincent, "Privacy-preserving adversarial representation learning in ASR: Reality or illusion?," in *Proc. Interspeech*, 2019, pp. 3700–3704.

[13] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. ICASSP*. IEEE, 2018, pp. 5329–5333.

[14] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks.," in *Proc. Interspeech*, 2018, pp. 3743–3747.

[15] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. Interspeech*, 2015, pp. 3214–3218.

[16] Xin Wang, Shinji Takaki, and Junichi Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis," in *Proc. ICASSP*, 2019, pp. 5916–5920.

[17] Kushal Lakhotia, Evgeny Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Adelrahman Mohamed, et al., "Generative spoken language modeling from raw audio," *arXiv preprint arXiv:2102.01192*, 2021.

[18] Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux, "Speech resynthesis from discrete disentangled self-supervised representations," *arXiv preprint arXiv:2104.00355*, 2021.

[19] Wen-Chin Huang, Yi-Chiao Wu, and Tomoki Hayashi, "Any-to-one sequence-to-sequence voice conversion using self-supervised discrete speech representations," in *Proc. ICASSP*, 2021, pp. 5944–5948.

[20] Hyeong-Seok Choi, Juheon Lee, Wansoo Kim, Jie Lee, Hoon Heo, and Kyogu Lee, "Neural analysis and synthesis: Reconstructing speech from self-supervised representations," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[21] Benjamin van Niekerk, Marc-André Carbonneau, Julian Zaïdi, Mathew Baas, Hugo Seuté, and Herman Kamper, "A comparison of discrete and soft speech units for improved voice conversion," *arXiv preprint arXiv:2111.02392*, 2021.

[22] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer, "End-to-end text-dependent speaker verification," in *Proc. ICASSP*, 2016, pp. 5115–5119.

[23] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based

Speaker Verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[24] Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li, "AISHELL-3: A Multi-Speaker Mandarin TTS Corpus," in *Proc. Interspeech*, 2021, pp. 2756–2760.

[25] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[26] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.

[27] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[28] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[29] Brij Mohan Lal Srivastava, Natalia A. Tomashenko, Xin Wang, Emmanuel Vincent, Junichi Yamagishi, Mohamed Maouche, Aurélien Bellet, and Marc Tommasi, "Design choices for x-vector based speaker anonymization," in *Proc. Interspeech*, 2020, pp. 1713–1717.

[30] Stephen Edward McAdams, *Spectral fusion, spectral parsing and the formation of auditory images*, Stanford university, 1984.

[31] Jose Patino, Natalia Tomashenko, Massimiliano Todisco, Andreas Nautsch, and Nicholas Evans, "Speaker Anonymisation Using the McAdams Coefficient," in *Proc. Interspeech*, 2021, pp. 1099–1103.

[32] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," 2020.

[33] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[34] Kavita Kasi and Stephen A Zahorian, "Yet another algorithm for pitch tracking," in *Proc. ICASSP*, 2002, vol. 1, pp. I–361.

[35] Ankita Pasad, Ju-Chieh Chou, and Karen Livescu, "Layer-wise Analysis of a Self-supervised Speech Representation Model," in *Proc. ASRU*, 2021, pp. 914–921.

[36] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Proc. NIPS*, 2020, vol. 33, pp. 12449–12460.

[37] Zhengyang Chen, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, Yanmin Qian, and Michael Zeng, "Large-scale self-supervised speech representation learning for automatic speaker verification," 2022.

[38] Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198.

[39] Weicheng Cai, Jinkun Chen, and Ming Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," in *Proc. Odyssey*, 2018.

[40] Xiaoxiao Miao, Ian McLoughlin, Wenchao Wang, and Pengyuan Zhang, "D-MONA: A dilated mixed-order non-local attention network for speaker and language recognition," *Neural Networks*, vol. 139, pp. 201–211, 2021.

[41] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. CVPR*, 2019, pp. 4690–4699.

[42] Xu Xiang, Shuai Wang, Houjun Huang, Yanmin Qian, and Kai Yu, "Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition," in *Proc. APSIPA ASC*. IEEE, 2019, pp. 1652–1656.

[43] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[44] Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi, "Generalization ability of MOS prediction networks," in *Proc. ICASSP (accepted)*, 2022.

[45] Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.

[46] Pierre Champion, Denis Jouvet, and Anthony Larcher, "Speaker information modification in the VoicePrivacy 2020 toolchain," Research report, INRIA Nancy, équipe Multispeech ; LIUM - Laboratoire d'Informatique de l'Université du Mans, nov 2020.

[47] Lantian Li, Ruiqi Liu, Jiawen Kang, Yue Fan, Hao Cui, Yunqi Cai, Ravichander Vipperla, Thomas Fang Zheng, and Dong Wang, "CN-Celeb: multi-genre speaker recognition," *Speech Communication*, 2022.

[48] Yue Fan, JW Kang, LT Li, KC Li, HL Chen, ST Cheng, PY Zhang, ZY Zhou, YQ Cai, and Dong Wang, "CN-Celeb: a challenging Chinese speaker recognition dataset," in *Proc. ICASSP*. IEEE, 2020, pp. 7604–7608.

[49] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng, "Aishell-1: An open-source Mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.

[50] Desh Raj, David Snyder, Daniel Povey, and Sanjeev Khudanpur, "Probing the information encoded in x-vectors," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 726–733.

[51] Jennifer Williams and Simon King, "Disentangling style factors from speaker representations.," pp. 3945–3949, 2019.