Voxstructor: Voice Reconstruction from Voiceprint

Panpan Lu[1], Qi Li[2], Hui Zhu[1], Giuliano Sovernigo[2], and Xiaodong Lin[2](✉)

[1] Xidian University, Xi'an, China
{lupanpan@stu.,zhuhui@}xidian.edu.cn
[2] University of Guelph, Guelph, Canada
{qli15,gsovernigo,xlin08}@uoguelph.ca

Abstract. With the rapid development of machine learning technologies, voiceprint has become widely used as a personal identifier in daily life. Because of that, it is essential to determine to what extent a voiceprint derived from machine learning can be inverted to obtain the original speaker characteristic. However, the reconstruction of voiceprint templates is still a challenging issue. It has also not been proven whether the widespread use of voiceprint poses a privacy leakage risk. In this paper, we implement the first comprehensive, holistic, and systematic reconstruction study targeting voiceprint templates. We present Voxstructor, a voiceprint-based voice constructor that can be used for bulk template reconstruction attacks. An attacker can reconstruct a new voice based only on the victim's voiceprint data instead of the voice itself. Specifically, we formalize the voice reconstruction work as an objective optimization problem and merge voice cloning with voiceprint template conversion work. We have conducted extensive experiments on multiple mapping models, loss functions, voiceprint template extraction models, scoring methods, and two types of speaker verification attacks. Thorough experiments show that our attacks are effective, achieving a fairly high success rate which is similar to the results generated by voice cloning methods. The time overhead of Voxstructor is far less than other attacks. Our study not only demonstrates the need for protection of voiceprint templates in speaker recognition systems, but also shows that Voxstructor can be used as a privacy measure tool for voiceprint privacy-preserving schemes.

Keywords: Privacy · Reconstruction · Speaker verification · Voiceprint.

1 Introduction

The application of speaker recognition systems is on the rise, such as in banking, voice assistants, online authentication among numerous others. At the same time, the attacks against speaker recognition systems become correspondingly more common. Several representative attack methods have been proposed separately, such as adversarial noise[1–3], replay attack[4], speech synthesis[5], and others.

All these attacks point out the vulnerability of some modules in the speaker recognition system.

It is well known that a speaker recognition system operates by extracting the voiceprint vector from the speaker's speech, and then comparing it with the stored voiceprint to calculate the similarity. This process can be used to determine the identity of the speaker. Voiceprints are typically compact binary or real-valued feature representations that are extracted from voice samples or voice features to increase the efficiency and accuracy of similarity computation. Over the past couple of decades, a large number of approaches have been proposed for voiceprint[6–8].

In this paper, we focus on template reversibility and reconstruction attacks in speaker recognition systems. In a voiceprint reconstruction attack, if voice can be reconstructed from the target's voiceprint, it can be used to gain access to the target through the target or other user-registered systems, thus threatening the target's interests and safety. Template reconstruction attacks generally assume that templates of target subjects and the corresponding black-box template extractor can be accessed [9]. First, templates of target users can be exposed in hacked databases. Second, the corresponding black-box template extractor can potentially be obtained by purchasing the speaker recognition SDK. To our knowledge, almost all of the speaker recognition vendors store voiceprints without template protection.

There are existing works on face template reconstruction [9], but these cannot be applied to voiceprint reconstruction. Faces are static features and do not change dynamically. Human voice however, changes dynamically with content, emotion, and other factors. Voiceprints are text-independent for increasing recognition accuracy, which makes it more difficult to reconstruct the target's speech. Similar work has been done such as voice synthesis [5, 10], and adversarial voice attack [1–3]. However, such work requires the original speech of the target as input, which is difficult to obtain these speeches in reality.

To address the issues of voiceprint reconstruction, inspired by voice cloning [10, 5], we propose a voice constructor from voiceprint, called Voxstructor "vox", as derived from the etymology of "voice"). First, we use a common voiceprint extractor such as i-vector as a black box to extract voiceprint. Second, we use multiple neural networks to construct a mapping transformation model from voiceprint to speaker embedding in voice clones. Third, the speaker embedding is used to generate speech from existing voice cloning technology. In our study of voiceprint reconstruction attacks, we made no assumptions about subjects used to train the target speaker recognition system. Therefore, we use Kaldi's pretrained voiceprint extractors in our research, and use public datasets to train our attack model. We experiment and analyze our mapping transformation model from multiple loss functions and multiple model structures. We also abstract several attack scenarios, conduct experiments and analysis for these scenarios. In summary, we make the following contributions:

- We conduct a comprehensive study on the reversibility of voiceprint. To our best knowledge, this is the first study on voiceprint reconstruction and voiceprint privacy.
- Voxstructor is developed for reconstructing voice samples from voiceprint. We implement and analyze voiceprint reconstruction under three mapping network structures, three loss functions, three voiceprint extractors and three discrimination thresholds while achieving a very high attack success rate.
- We discuss the multiple implications of our scheme. It not only exposes the reversibility and sensitivity of the voiceprint, but also demonstrates the need for privacy protection. Moreover, it can be used in several aspects such as computer forensics, and privacy-preserving effect metrics.

The remainder of this paper is organized as follows. We review the relevant background information in Section 2. The proposed scheme and the performance evaluation are followed in Section 3 and Section 4 respectively. In Section 5, we review some related works. Finally, we draw our conclusion in Section 6.

2 Background

In this section, we introduce the basic knowledge of the speaker verification system and threat model.

2.1 Speaker verification system

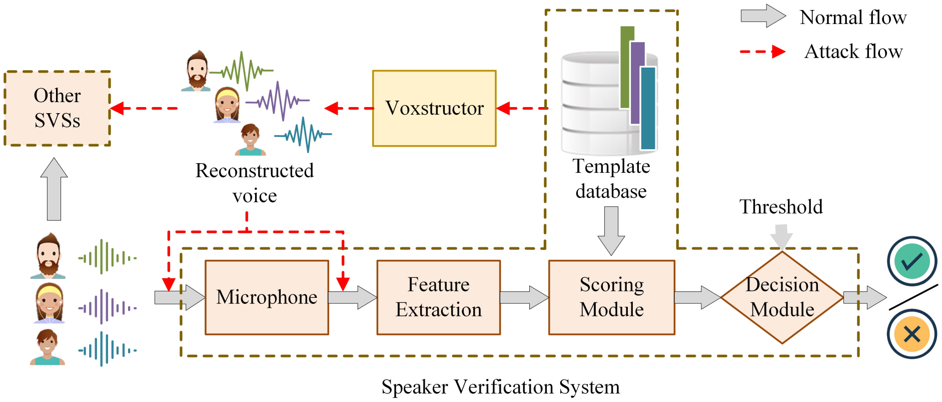

Fig. 1. Overview of the proposed system for reconstructing voices from the corresponding templates.

Speaker recognition is an automatic technology which can recognize the speaker's identity according to the sound characteristics extracted from their speech. The flow of a speaker verification system (SVS) is shown with the dashed box in the normal flow in Figure 1, which mainly includes five modules. **Microphone** is used to collect user's registered voice and verification voice. **Extractor** is used to extract speaker characteristics in voice. **Database** is used to store user's registered voiceprint template. **Scoring module** is used to match speaker feature vector extracted from verification voice with registered voiceprint template stored in database and outputs a similarity score. **Decision module** is used to compare the result of scoring module with threshold and gives the decision result of pass or fail.

At present, there are several popular technologies used in speaker recognition system as follows.

i-vector The method based on the i-vector involves modeling the global difference, and modeling the speaker and channel as a whole [6]. In this way, the restrictions on the training corpus are relaxed, the calculation is simple, and the performance is better. The i-vector contains both the speaker differential information and the channel differential information, so it is necessary to remove the channel interference in the i-vector and use channel compensation technology to eliminate the channel interference. An i-vector can be written as 400 or 600 dimensional vector.

x-vector Snyder et al. [7] defined the x-vector and proposed an extraction model based on a multi-layer delayed neural network, which can transform the input features at the frame level into the feature expression at the sentence level. The embedded vector extracted from the model is called the x-vector, which can be used similarly to the i-vector. The dimension of this vector is 512, and it also contains the channel information. Channel compensation technology is needed to eliminate the interference (PLDA classifier is used in the training process).

Resnet34 model Heo et al. [8] proposed the Resnet34 voiceprint extraction model based on residual networks. In this paper, different loss functions are used to train the model. Comparing the accuracies of the models, GE2E and the original network (AP + softmax) have the highest accuracy.

Scoring methods It consists of three main scoring methods in SVS systems: PLDA, cosine distance, and Euclidean distance. As a channel compensation algorithm of i-vector and x-vector, PLDA is widely used in SVS because its compensation effect is better than other channel compensation algorithms (such as LDA) and the scoring methods are based on calculating log likelihood. Research shows that channel information will cause the size of feature vector to change, while speaker information mainly affects the direction of i-vector feature vector [6], so cosine distance weakens the influence of channel information to a certain extent. Finally, Euclidean distance is widely used as a way to measure the distance between two points (vectors). Here we combine it with the Resnet34 model as an SVS to show the effect of Euclidean distance in the field of voiceprint feature recognition.

2.2 Threat model

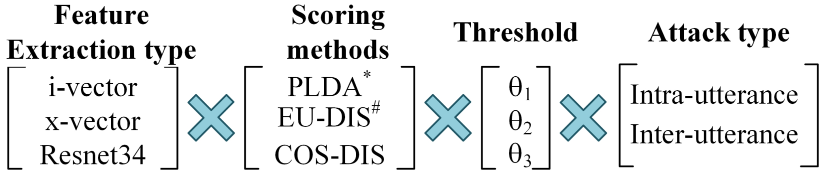

Fig. 2. Attack scenarios, where * means that PLDA is not used for SVS with Resnet34 as the voiceprint extractor, and # means that EU-DIS is only used for SVS with Resnet34 as the voiceprint extractor. $\theta_1, \theta_2, \theta_3$ are based on EER, high user experience degree, and high security degree respectively.

We assume that the adversary has the voiceprint template of the registered speaker and hopes to design a voice sample to defeat the SVS.

The scenario of the template reconstruction attack is shown in Figure 1. The adversary obtains the target's voiceprint template through some means, such as purchasing it illegally or obtaining it through unauthorized access, for example, caused by software vulnerabilities in the SVS such as buffer overflow privilege escalation. Then the adversary uses it to reconstruct the target's voice through voxstructor, and uses the voice to attack the target's speaker verification system. According to whether the voiceprint template obtained by the adversary is from the target SVS, there are two types of attacks in voiceprint reconstruction attack: (1)intra-utterance, the reconstructed voiceprint template comes from the victim registration voiceprint template stored in the target SVS; (2)inter-utterance, the reconstructed voiceprint template comes from the victim's unregistered voiceprint template.

Our proposed voiceprint reconstruction attack is a black-box attack. In other words, the attacker can attack the system without knowing the neural network model in the SVS (such as structure, parameters, and training data set, etc.). The reason is that the proposed voiceprint reconstructor does not need to understand the neural model, and only needs to input the voiceprint template to reconstruct the victim's voice.

In our attack model, in order to fully demonstrate the effect of voiceprint reconstruction attack, we design six SVSs based on three mainstream voiceprint extraction models (i-vector, x-vector, Resnet34) and three popular scoring methods (PLDA, Euclidean distance, cosine distance). According to the system's availability and security requirements, we set three different thresholds based on the system accuracy evaluation target equal error rate (EER), high user experience degree, and high security degree for each SVS. Including two types of attacks, as shown in Figure 2, there are a total of 36 attack scenarios.

3 Voxstructor

Voxstructor is mainly realized by voice cloning technology [5] and voiceprint mapping model. Figure 3 shows the structure diagram of voxstructor. Voiceprint templates are transformed into the speaker embedding vectors by mapping model, and the vector and text content are synthesized into mel spectrograms by the synthesizer, and the reconstructed speech is generated by the vocoder. In this section, we will introduce the details and technologies in the process of voiceprint reconstruction.

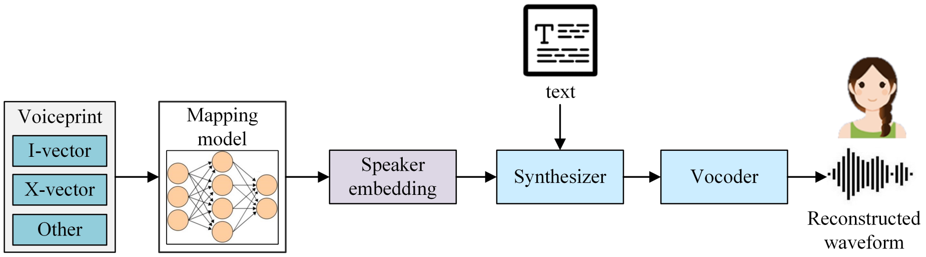

Fig. 3. Voxstructor structure diagram.

3.1 Voice cloning

Google proposes real-time voice cloning (RTVC) technology, which is based on a text-to-speech (TTS) synthesis system that can generate speech audio in different speaker's voices [5]. The technology consists of three main components: (1) Speaker encoder network: a multilayer LTSM network, trained as a speaker verification task, that generates a fixed dimensional embedding vector speaker embedding from only a few seconds of reference speech of the target speaker, which has the characteristics of the speaker; (2) Synthesizer: a Tacotron 2-based [11] inter-sequence synthesis network that generates mel spectrograms from text conditional on the speaker embedding vector generated in the speaker encoder; (3) Vocoder: an autoregressive WaveNet-based [12] vocoder network that converts the generated in the synthesizer mel spectrograms into time domain waveform samples, thus generating the speech audio of the target person. While voice cloning provides our research with a strong base, our voiceprint reconstruction converts different forms of voiceprint templates into speaker embedding vectors, synthesizes mel spectrograms with text through Synthesizer, and finally generates speech files through the vocoder.

3.2 Problem formation

Given a target person's voiceprint template x, the target person's voice v is reconstructed from x, and v must be similar enough to the target person's voice in order to pass the SVS. In other words, the voiceprint x' extracted from v by the voiceprint extractor in the SVS and the registered template x can be scored against the threshold after passing the scoring method. Therefore, we wish the difference between x and x' to be as small as possible. Then the problem is formalized as:

$$\min\ d(x, x'), \tag{1}$$

where d represents the distance between the two vectors.

Our ultimate goal is to reconstruct the voice from the voiceprint, i.e., we wish to construct a voiceprint reconstructor $v = R(x, t)$, where t is the text of the reconstructed voice, and v can pass the SVS. The known voice cloning technique [5] for cloning voice is: $v = g(se, t)$, where se is the origin speaker embedding of speaker's voice. Therefore we can introduce the mapping model $m(.)$ to establish the link between x and se: $se' = m(x)$, then the voice reconstruction machine can be expressed as:

$$v = R(x, t) = g(m(x), t). \tag{2}$$

Because we build a link between voiceprint and speaker embedding, the problem Eq.(1) can be transformed to Eq.(3):

$$\min\ d(se, se'). \tag{3}$$

In summary, our goal is to find mapping model $m(.)$, which can achieve min $d(se, se')$.

3.3 Mapping model

The mapping models are three-layer fully connected structures and convolutional neural network, which have simple structures, easy implementation, and few parameters. The normalization of the voiceprint is needed before the input model. According to three different types of voiceprint characteristics, we design three different mapping models. (1) I2E: this model maps i-vector to speaker embedding; (2) X2E: this model maps x-vector to speaker embedding; (3) R2E: this model maps the voice eigenvector extracted by Resnet34 to speaker embedding. The mapping model we proposed is not limited to these three types.

Let $D(\cdot, \cdot)$ denote the reconstruction loss function between different speaker embedding, v denote a training voice sample from the public datasets, $f(\cdot)$ denote voiceprint extractor function, $g(\cdot)$ denote voice cloning function, t denote the text in the reconstructed voice, θ denote the parameters of mapping model, e denote the speaker embedding extractor function. According to Eq.(1) and Eq.(3), the objective function for training mapping models can be formulated as

$$\arg\min_{\theta} \mathcal{L}(v,\theta) = \arg\min_{\theta} \frac{1}{N}\sum_{i}^{N} \mathcal{D}(f(v_i), f(g(m_\theta(f(v_i)), t))) \tag{4}$$

$$\approx \arg\min_{\theta} \frac{1}{N}\sum_{i}^{N} \mathcal{D}(m_\theta(f(v_i)), e(v_i)), \tag{5}$$

where N denotes the number of voice samples. After training, we can get the reconstructed voice sample v' using the target's voiceprint x: $v' = g(m_\theta(f(v_i)), t)$.

We use three different loss functions (MSELoss, L1Loss and SmoothL1Loss) as $D(\cdot,\cdot)$ in Eq.(4) to train our model, so as to show the influence of different distance measurement methods on the voiceprint mapping model. For each mapping model, we use the above loss functions to train the mapping models, which contains nine models.

4 Experiment Evaluation

We evaluate the reconstruction attack capability of Voxstructor[3] in a SVS from the following four aspects: effectiveness, efficiency, human perceived similarity and privacy-reserving methods effect metric. Then, we introduce the datasets, experimental design and the above four aspects. For convenience and clarity, we list the abbreviations used in the experiment in Table 1.

4.1 Dataset and design

Dataset The databases we used are voxceleb1 [13] and librispeech [14]. Voxceleb1 contains about 100000 voice samples of 1251 celebrities from YouTube Videos. The data is reasonably gender balanced (55% male). For the speaker verification system, the data set can be divided into a development set and a test set, and there is no overlap between them.

Librispeech is the most authoritative mainstream open-source dataset to measure speech recognition technology. It is an audiobook data set containing text and voice. The data comes from the audio recordings of reading materials from the Librivox project, and is carefully subdivided and consistent. Each recording is split into segments of 10 seconds and linked to its corresponding section of the accompanying text.

Design In order to better and more comprehensively evaluate the voice reconstruction attack capability of Voxstructor, we target three speaker verification systems: i-vector, x-vector and deep residual network ResNset34 model. These are test in the most popular open-source platform-kaldi [15]. In addition, we use

[3] https://github.com/voxstructor/voxstructor

Table 1. The list of notations used in the experiments.

Notations	Descriptions
Average-utterance	Voxstructor based on the mean voiceprint
CONV	MSE & convolutional network
Inter-utterance	Voxstructor based on the Unregistered voiceprint
Intra-utterance	Voxstructor based on the registered voiceprint
IV-COS	i-vector & cosine distance as score
IV-PLDA	i-vector as voiceprint & PLDA as score
L1L	L1 loss & fully connected network
L1L-text2	L1L & use second text to generate
L1LNO	L1L & without normalization
MSE	MSE loss & fully connected network
Rand-vector	Randomly generated voiceprint vectors
Rand-wav	4874 randomly generated voices
RN-EU	Resnet & euclidean distance as score
RN-COS	Resnet & cosine distance as score
Smooth	Smooth loss & fully connected network
RTVC	Voice generated by voice cloning tool
RTVC-text2	RTVC & use second text to generate
XV-PLDA	x-vector as voiceprint & PLDA as score
XV-COS	x-vector & cosine distance as score

the more popular PLDA, cosine distance, and Euclidean distance as the scoring methods for speaker verification systems. The i-vector extractor and the corresponding PLDA use the pre-trained model from the open source tool Kaldi [4]. The x-vector model and its PLDA also use the pre-trained Kaldi model [5]. The Resnet34 model uses the pre-trained model from Joon et al. [16]. The performance of these six systems can be seen in Table 2.

To be able to fully represent the Voxstrutor voice reconstruction capability, we use all the voices in the test set (40 users, 4874 voices) to register the speaker verification system and reconstruct the voices according to the corresponding voiceprint. We also consider the case that some current speaker verification systems register users with a voiceprint template that averages the vectors extracted from multiple voices of the user. Therefore, we also use the average voiceprint of the four voices of each user in the test set to register the system and reconstruct the voice based on this voiceprint template. In addition, we selected 100 speakers from the tran-clean-100 in LibriSpeech with clear and noiseless speech to test the reconstruction ability of Voxstructor.

Evaluation Criteria We use false rejection rate (FRR), false acceptance rate (FAR), and equal error rate (EER) to express the performance of the SVS. We use the attack pass rate to evaluate the ability of the voice-reconstruction attack, i.e., the percentage of voices reconstructed from the voiceprint that pass the

[4] https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v1

[5] https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v2

Table 2. Performance of the six baseline SVSs(%).

SVS	Threshold	EER	FRR	FAR
IV-PLDA	-1.000	5.342	5.371	5.313
	-6.000	——	1.193	17.487
	3.000	——	14.846	1.278
IV-COS	0.060	13.831	13.807	13.855
	-0.020	——	1.283	67.869
	0.128	——	41.559	1.012
XV-PLDA	-3.000	3.134	3.303	2.964
	-8.000	——	1.288	7.635
	1.000	——	7.269	1.103
XV-COS	0.670	9.722	9.369	10.074
	0.530	——	1.082	32.847
	0.770	——	33.330	1.145
RN-EU	80.000	5.199	5.795	4.602
	87.000	——	1.129	18.722
	75.000	——	13.971	1.145
RN-COS	0.350	1.880	1.023	2.736
	0.390	——	2.322	1.198

speaker verification system. To evaluate the efficiency of the voice reconstruction attack, the execution time of the reconstructed voice is used as a metric.

4.2 Effectiveness

Target model In order to evaluate the effectiveness of the voiceprint reconstruction attack, we designed six speaker verification systems (IV-PLDA, IV-COS, XV-PLDA, XV-COS, RN-EU, RN-COS) using a combination of the three most popular voiceprint extraction methods and three scoring methods, PLDA, Euclidean distance, and cosine distance. For simplicity, IV is used to denote i-vector, XV to denote x-vector, RN to denote Resnet34, and PLDA, EU, COS to denote the three scoring methods of PLDA, Euclidean distance, and cosine distance, respectively.

Here, we set three thresholds for each system. The first threshold is determined based on the EER, which is a relatively good compromise between availability and security of the verification system. The second will be determined based on the FRR value of 1.0% , i.e., the verification system focuses on availability. The third will be determined based on the FAR value of 1.0% , i.e., the verification system focuses on security. It is worth noting that similar to other biometric verification techniques, the input voice (or extracted voiceprint from the user input) and the stored voiceprint usually do not match perfectly. As a result, a matching score threshold must be set for SVS to verify the identity of user. In the RN-COS system, the FRR can be taken as 1.023% when the EER is 1.88%, so the second threshold is not set.

Voiceprint reconstruction results To evaluate the voice reconstructed based on different voiceprint forms, we test the pass rate of the voice reconstructed

Table 3. The pass rates of Voxstructor, RTVC and random guessing under intra-utterance type(%).

Model	SVS	Threshold	Voxstructor	RTVC	Rand_vector	Rand-wav
I2E	**IV-PLDA**	-1.000	77.754	84.773	1.950	0.636
		-6.000	97.106	98.071	10.979	6.731
		3.000	41.822	54.853	0.082	0.041
	IV-COS	0.060	41.137	52.626	11.264	15.224
		-0.020	88.162	91.711	66.886	66.639
		0.128	7.817	13.993	0.431	1.847
X2E	**XV-PLDA**	-3.000	72.522	79.889	3.099	2.586
		-8.000	90.971	93.803	16.745	16.068
		1.000	48.964	60.456	0.349	0.369
	XV-COS	0.670	89.208	86.869	0.000	0.041
		0.530	99.713	99.036	0.000	0.759
		0.770	45.466	46.492	0.000	0.000
R2E	**RN-EU**	80.000	64.854	67.357	0.533	12.946
		87.000	88.531	90.008	5.929	52.216
		75.000	40.870	44.276	0.021	1.785
	RN-COS	0.350	63.192	70.086	0.000	1.149
		0.390	45.568	53.037	0.000	0.041

with different types of voiceprints in the corresponding SVS. In addition, we designed two sets of comparison experiments, one is to test the pass rate of the original voice of registered users directly synthesized by the real-time voice cloning tool (RTVC) [5] in six speaker verification systems. The other is to verify the pass rate of random guesses, we generated two forms of data sets randomly based on the test set of voxceleb1, one randomly generated voice vector set (rand_vector) and the other randomly generated voice set (rand_wav). The pass rates of Voxstructor, RTVC and random guessing under intra-utterance case are shown in Table 3, where the loss function of the Voxstructor is the L1Loss function. We verify the impact caused by the voiceprint after reconstruction on the SVS based on 4874 voices in the voxceleb test set. The results show that our attack scheme is fully effective and can achieve similar results to RTVC, far exceeding the two random guesses.

As shown in Table 4, the pass rates of voice reconstructed based on average voiceprint are higher than common intra-uttenance case. This result indicates that the mean voiceprint-based reconstruction attack is much more effective against the SVS based on their models than the single speech-based voiceprint reconstruction attack. That is, the mean voiceprint of multiple voices of a user is better at characterizing the user's voice than the voiceprint of a single voice.

In the inter-utterance case, the pass rate of the reconstructed speech from our voiceprint reconstruction scheme is about 20% lower in SVSs than in the intra-utterance case. This shows that even for different voices of the same person, there are still relatively large differences and it is still not a good way to model a person's speech characteristics. The result is a good illustration of the limitations and drawbacks of short speech registration in speaker verification systems. However, this pass rate is still fatal to SVSs, and once the voiceprint of

Table 4. The pass rates of different datasets, average-utterance, intra-utterance and inter-utterance(%).

Model	SVS	Threshold	librispeech	voxceleb	Average	Intra	Inter
I2E	**IV-PLDA**	-1.000	96.000	77.754	85.000	77.754	48.133
		-6.000	100.000	97.106	97.500	97.106	83.772
		3.000	84.000	41.822	47.500	41.822	18.366
	IV-COS	0.060	81.000	41.137	52.500	41.137	28.706
		-0.020	95.000	88.162	90.000	88.162	80.764
		0.128	43.000	7.817	20.000	7.817	4.369
X2E	**XV-PLDA**	-3.000	94.000	72.522	72.500	72.522	44.944
		-8.000	100.000	90.971	95.000	90.971	71.989
		1.000	84.000	48.964	42.500	48.964	23.544
	XV-COS	0.670	100.000	89.208	100.000	89.208	71.729
		0.530	100.000	99.713	100.000	99.713	95.758
		0.770	66.000	45.466	85.000	45.466	23.293
R2E	**RN-EU**	80.000	82.000	64.854	87.500	64.854	49.968
		87.000	92.000	88.531	100.000	88.531	80.064
		75.000	66.000	40.870	80.000	40.870	28.112
	RN-COS	0.350	93.000	63.192	92.500	63.192	40.027
		0.390	82.000	45.568	80.000	45.568	24.624

a registered user of all SVSs is leaked, then SVS using the same model as that SVS will also be threatened.

The pass rates of two datasets in the intra-utterance case are shown in Table 4, where the loss function of the model is the L1Loss function. As can be seen from the table, among these six SVSs, the pass rate of the speech reconstructed by Voxstructor based on the voiceprint in librispeech speech can reach more than 90% in the SVS with the threshold value of voxceleb, which is about 20% higher than the pass rate tested with voxceleb data. This is mainly due to the fact that the voxceleb speech is mainly from YouTube videos, which contains additional background noise, while the librispeech speech is clean speech from audiobook readings.

Effect of loss functions In order to test the effect of different loss functions on the mapping models, we designed three loss functions, L1Loss, MSELoss, and SmoothL1Loss, to train our proposed three models, I2E, X2E, R2E, and test their pass rates. The pass rates of the mapping models trained with different loss functions are shown in Table 5. The results show that SmoothL1Loss has the highest accuracy for i-vector voiceprints and MSELoss has the highest accuracy for x-vector and Resnet voiceprints.

Model structure For the mapping models in the proposed scheme, we also design a set of comparison experiments using three different structures of mapping models for the acoustic vectors. These are namely the fully connected, convolutional, and unnormalized.

Table 5. The pass rates of different loss function under intra-utterance type(%).

Model	**SVS**	**Threshold**	**L1L**	**MSE**	**Smooth**
I2E	**IV-PLDA**	-1.000	77.754	76.790	80.176
		-6.000	97.106	96.696	97.496
		3.000	41.822	41.740	43.854
	IV-COS	0.060	41.137	43.906	45.507
		-0.020	88.162	89.516	89.577
		0.128	7.817	9.643	10.423
X2E	**XV-PLDA**	-3.000	72.522	72.604	71.681
		-8.000	90.971	91.525	90.806
		1.000	48.964	49.292	48.245
	XV-COS	0.670	89.208	88.346	89.229
		0.530	99.713	99.815	99.733
		0.770	45.466	43.188	44.337
R2E	**RN-EU**	80.000	64.854	67.111	64.198
		87.000	88.531	91.075	88.941
		75.000	40.870	42.388	39.352
	RN-COS	0.350	63.192	65.429	64.362
		0.390	45.568	47.476	46.984

The pass rates of the fully connected mapping model, the fully connected mapping model without normalization of the voiceprint, and the convolution-based mapping model in the speaker verification system are shown in Table 6. As we can see, The pass rate for the unnormalized fully-connected mapping model of the voiceprint is very low, about 70% lower than that of the normalized fully-connected model. It was observed that in the reconstruction for i-vector and Resnet voiceprint, the pass rates of the fully connected structure is almost the same as that of the convolutional structure. And the fully connected structure outperforms the convolutional structure in the reconstruction for the x-vector. This indicates that the correlation between the components of the voiceprint template is small and there is no local receptive field.

Due to the limited space, we only show the pass rates in the intra-utterance attack type here, and the pass rate for the inter-utterance type can be found in the Appendix.

Text independence Finally, to characterize the text-independent speaker verification system, two different texts were used to synthesize two sets of speech. Text 1 is: "This is being said in my own voice. The computer has learned to do an impression of me." Text 2 is: "The prince loves his roses, but felt disappointed by something the rose said. As doubt grows, he decides to explore other planet." We synthesize two sets of speech based on the two texts after mapping the voiceprints using a mapping model trained with the L1Loss loss function. In addition, to compare and demonstrate the effect of our reconstruction scheme to reconstruct two different contents of speech in a text-independent system, we take the speech from the voxceleb1 test set directly through the RTVC tool to synthesize two sets of speech with different texts.

Table 6. The pass rates of different mapping models under intra-utterance type(%).

Model	SVS	Threshold	L1L	L1LNO	CONV
I2E	**IV-PLDA**	-1.000	77.754	8.537	77.940
		-6.000	97.106	21.547	97.189
		3.000	41.822	1.642	42.110
	IV-COS	0.060	41.137	13.500	44.112
		-0.020	88.162	66.229	89.516
		0.128	7.817	0.985	10.176
X2E	**XV-PLDA**	-3.000	72.522	1.847	52.473
		-8.000	90.971	6.587	79.356
		1.000	48.964	0.451	29.079
	XV-COS	0.670	89.208	21.892	78.539
		0.530	99.713	29.237	99.056
		0.770	45.466	7.571	28.272
R2E	**RN-EU**	80.000	64.854	2.400	61.202
		87.000	88.531	11.899	86.725
		75.000	40.870	0.472	36.048
	RN-COS	0.350	63.192	1.805	57.858
		0.390	45.568	0.677	41.157

The pass rates of the voice reconstructed using two different English texts with the mapped vectors through the speaker verification system is shown in Table 7 and Table 16 (see Appendix). The results indicate that our attack can still achieve a high pass rate even when generating a voice with different text content than the original registered voice. The pass rate for both texts is essentially the same. This shows that Voxstructor is fully applicable to diverse attacks and can generate commands with sensitive semantics to further threaten the security of smart voice assistants, smart homes, and other environments.

4.3 Efficiency

We test the time of synthesizing sound in our voiceprint reconstruction scheme on a Windows PC with NVIDIA p5000 GPU. The test time is shown in table 8. Our voiceprint reconstruction attack can reconstruct the user's voice from the voiceprint vector without accessing the verification system. However, the FAKEBOB proposed in the paper [1] not only needs a segment of speech as the original speech of the target speech, but also needs to visit the verification system many times to make the speech conversion successful. Therefore, the time consumed by our proposed voiceprint reconstruction attack is much less than that of FAKEBOB attack, which is 80 times faster than that of FAKEBOB attack.

4.4 Manual listening experiment

We randomly select 10 people from librispeech and pick 2 sentences each at random. We extract three voiceprints for each sentence and reconstruct them using Voxstructor to get 3*2*10=60 new voices. At the same time, we use RTVC

Table 7. Pass rate of text-independent under Intra-utterance type(%).

Model	SVS	Threshold	L1L	L1L-TEXT2	RTVC	RTVC-TEXT2
I2E	**IV-PLDA**	-1.000	77.754	74.635	84.773	82.865
		-6.000	97.106	95.834	98.071	97.353
		3.000	41.822	39.113	54.853	52.719
	IV-COS	0.060	41.137	39.598	52.626	51.252
		-0.020	88.162	88.059	91.711	92.983
		0.128	7.817	6.483	13.993	12.659
X2E	**XV-PLDA**	-3.000	72.522	67.063	79.889	76.093
		-8.000	90.971	86.723	93.803	91.340
		1.000	48.964	45.392	60.456	56.105
	XV-COS	0.670	89.208	92.286	86.869	89.146
		0.530	99.713	99.733	99.036	99.015
		0.770	45.466	53.160	46.492	48.133
R2E	**RN-EU**	80.000	64.854	88.141	67.357	86.438
		87.000	88.531	98.420	90.008	97.435
		75.000	40.870	69.696	44.276	67.152
	RN-COS	0.350	63.192	67.009	70.086	69.655
		0.390	45.568	49.097	53.037	53.775

Table 8. The time consumed by voiceprint reconstruction attack and FAKEBOB attack (seconds).

	I2E	X2E	R2E	FAKEBOB
Time (seconds)	23.781	27.254	25.340	2014

to generate the same 2*10=20 strips. These are combined, and we invite 10 volunteers to perform a manual listening test to evaluate the similarity with the original speech. We ask the testers to score the speech on a scale of 0-5, where a score of 0 indicates that it is completely unlike the original speech and a score of 5 indicates that it is identical. The results of the manual scoring are shown in Table 9.

Table 9. The manual listening scores of voxstructor and voice cloning[5].

	I2E	X2E	R2E	Average of three	RTVC
Scores	4.12	3.86	3.94	3.97	4.25

From Table 9, we can see that the average score of Voxstructor reconstructed out is 3.97, and the average score of RTVC is 4.25. The results show that the effect of our reconstructed speech using voiceprints is very close to the effect of RTVC using voice directly. Furthermore, both resemble the original voice so much that humans cannot distinguish whether it is the generated voice or not.

4.5 Privacy-preserving methods metric

Due to the sensitive nature of voiceprint biometrics, many privacy-preserving speaker recognition schemes have been developed in recent years. Thus, it is important to evaluate the effectiveness of these privacy protection mechanisms. In order to show the effectiveness of Voxstructor on the metric of voiceprint-based privacy-preserving schemes, we designed the following experiments.

Setup: This experiment is also conducted mainly under the intra-utterance case. The test data are obtained from 20 different speakers' voices in librispeech's tran-clean-100. We test the pass rates of reconstructed voice from the three kinds of protected voiceprints by Voxstructor in their SVSs. In order to exclude the influence of the voice text in the metric of the privacy protection scheme of the voiceprint, we reconstruct the voice content as well as the content of the original voice text. We metric for four current privacy-preserving methods for voiceprint.

- MR: multiplying the voiceprint value by a random number for protection purpose.
- ARV: Adding the voiceprint by a random vector for protection purpose.
- MOM: multiplying the voiceprint by an orthogonal matrix for protection purpose.
- MMV: multiplying the voiceprint by an orthogonal matrix followed by a random vector for protection purposes, where the elements in the random vector and the random orthogonal matrix are generated by normal distribution, and we designed the mean value to be 0 and the scalar vertebral difference to be 0, 0.5, 1, 2, 3, 4, 5, to verify its pass rate in the SVS, respectively.

Results: For the MR approach, the protection method of multiplying the voiceprint by a random number does not achieve the effect of protecting the voiceprint because the voiceprint will be normalized when the voiceprint mapping model of Voxstructor is passed.

For the ARV, MOM, and MMV approaches, our test data are shown in Table 10, Table 11, and Table 12 respectively. When the variance is small, the Voxstructor pass rate is very high. When the variance is large, the pass rate of the reconstructed speech is low. The privacy of the voice template is fully protected at this time. Therefore, we can conclude that the pass rate of Voxstructor is inversely correlated with the degree of privacy protection. In conclusion, Voxstructor can be used as a tool for evaluating privacy-preserving approaches for speaker verification systems.

5 Related work

At present, there are many studies on the security of intelligent voice system. In this part, we discuss the attacks on an intelligent voice system and compare them with Voxstructor.

Li et al. [2] introduce an imperceptible disturbance into the original speech signal to defeat the SVS. From the perspective of voiceprint template, they

Table 10. Pass rate of Voxstructor for ARV privacy-preserving schemes(%).

SVS	Threshold	std-0	std-0.5	std-1	std-2	std-3	std-4	std-5
IV-PLDA	-3.000	95.000	95.000	95.000	60.000	25.000	40.000	25.000
	-8.000	100.000	100.000	100.000	75.000	60.000	50.000	45.000
	1.000	95.000	80.000	70.000	40.000	20.000	20.000	10.000
IV-COS	0.060	100.000	95.000	100.000	80.000	75.000	65.000	60.000
	-0.020	100.000	100.000	100.000	100.000	100.000	95.000	95.000
	0.128	80.000	75.000	65.000	35.000	30.000	30.000	20.000
XV-PLDA	-3.000	100.000	100.000	100.000	100.000	95.000	95.000	85.000
	-8.000	100.000	100.000	100.000	100.000	100.000	100.000	100.000
	1.000	95.000	95.000	100.000	100.000	85.000	85.000	70.000
XV-COS	0.670	100.000	100.000	100.000	100.000	100.000	90.000	90.000
	0.530	100.000	100.000	100.000	100.000	100.000	100.000	100.000
	0.770	65.000	75.000	80.000	65.000	55.000	50.000	35.000
RN-EU	80.000	95.000	100.000	100.000	100.000	100.000	95.000	90.000
	87.000	100.000	100.000	100.000	100.000	100.000	100.000	100.000
	75.000	95.000	95.000	95.000	95.000	90.000	85.000	80.000
RN-COS	0.350	95.000	100.000	95.000	90.000	90.000	75.000	55.000
	0.390	80.000	90.000	90.000	80.000	80.000	65.000	45.000

Table 11. Pass rate of Voxstructor for MOM privacy-preserving schemes(%).

SVS	Threshold	std-0	std-0.5	std-1	std-2	std-3	std-4	std-5
IV-PLDA	-3.000	95.000	10.000	25.000	20.000	10.000	30.000	10.000
	-8.000	100.000	45.000	40.000	45.000	45.000	45.000	40.000
	1.000	95.000	0.000	10.000	10.000	5.000	20.000	0.000
IV-COS	0.060	100.000	60.000	75.000	85.000	70.000	70.000	55.000
	-0.020	100.000	85.000	90.000	95.000	95.000	95.000	90.000
	0.128	80.000	20.000	30.000	20.000	25.000	25.000	5.000
XV-PLDA	-3.000	100.000	20.000	15.000	0.000	15.000	15.000	5.000
	-8.000	100.000	25.000	35.000	20.000	25.000	20.000	20.000
	1.000	95.000	15.000	0.000	0.000	10.000	10.000	0.000
XV-COS	0.670	100.000	10.000	10.000	10.000	10.000	10.000	20.000
	0.530	100.000	40.000	50.000	40.000	45.000	40.000	35.000
	0.770	65.000	5.000	5.000	0.000	0.000	0.000	0.000
RN-EU	80.000	95.000	60.000	40.000	50.000	35.000	45.000	60.000
	87.000	100.000	90.000	85.000	85.000	80.000	95.000	95.000
	75.000	95.000	45.000	15.000	25.000	10.000	35.000	45.000
RN-COS	0.350	95.000	25.000	5.000	10.000	10.000	15.000	5.000
	0.390	80.000	10.000	5.000	5.000	5.000	10.000	5.000

generate the sample voice for spoofing SVSs by leveraging the Genetic algorithm, the fitness function in which is mainly designed according to the similarity score between the target's voiceprint and the voiceprint extracted from the sample speech. Comparatively, Voxstructor does not need multiple iterations to reconstruct voice, so its efficiency is high. Additionally, Voxstructor can realize the black-box attack without knowing the voiceprint extraction model. The voiceprint mimicry attack [3], realized the gray box or black box attack, but

Table 12. Pass rate of Voxstructor for MMV privacy-preserving schemes(%).

SVS	Threshold	std-0	std-0.5	std-1	std-2	std-3	std-4	std-5
IV-PLDA	-3.000	95.000	15.000	40.000	20.000	20.000	52.000	20.000
	-8.000	100.000	35.000	65.000	45.000	30.000	30.000	35.000
	1.000	95.000	0.000	20.000	10.000	10.000	5.000	0.000
IV-COS	0.060	100.000	55.000	60.000	55.000	50.000	70.000	70.000
	-0.020	100.000	100.000	95.000	85.000	95.000	90.000	90.000
	0.128	80.000	30.000	25.000	30.000	15.000	30.000	30.000
XV-PLDA	-3.000	100.000	10.000	20.000	10.000	15.000	20.000	20.000
	-8.000	100.000	35.000	40.000	40.000	25.000	45.000	35.000
	1.000	95.000	5.000	10.000	0.000	5.000	5.000	15.000
XV-COS	0.670	100.000	5.000	15.000	20.000	15.000	15.000	30.000
	0.530	100.000	30.000	55.000	50.000	50.000	40.000	55.000
	0.770	65.000	5.000	0.000	5.000	0.000	0.000	0.000
RN-EU	80.000	95.000	60.000	45.000	65.000	70.000	50.000	55.000
	87.000	100.000	95.000	85.000	90.000	85.000	85.000	85.000
	75.000	95.000	20.000	30.000	40.000	35.000	20.000	25.000
RN-COS	0.350	95.000	10.000	5.000	35.000	25.000	15.000	5.000
	0.390	80.000	10.000	0.000	10.000	15.000	5.000	0.000

it is in essence an adversarial voice attack. That is to say, only through most iterations can the sample speech contain the target's voiceprint template. The FAKEBOB proposed by Chen et al. [1] attacks speaker recognition systems (e.g. SV, OSI, CSI). However, FAKEBOB also needs to access the system several times, while Voxstructor does not need access to one. In summary, compared with adversarial voice attacks, our voiceprint reconstruction attack can quickly reconstruct the target's voice from the voiceprint template without additional voice samples.

The spoofing attack is to mimic the target's voice to trick the SVS. There are four main kinds of attacks. The first and second attacks are mimicking and replaying. The attacker creates a speech sample by mimicking or pre-recording the speech sample of a given target speaker, which are the simplest ways to cheat the speaker verification system. However, playback technology can not meet the requirements of text-dependent SVS when producing specific utterances, and mimic is quite hard to find in reality. Our voiceprint reconstruction attack can meet this kind of attack scenario. The third one is voice synthesis [17]. The attacker uses text to speech (TTS) synthesis system to synthesize the target's audio. However, the training of this synthesis model requires the target's speech set of at least tens of minutes. Voxstructor does not need to obtain any speech set of the target, it only needs the target's voiceprint. The fourth attack is voice conversion [18]. It is to modify the voice of one speaker (source) to make it sound like the voice of another speaker (target) without changing the language content. However, this kind of attack also needs the target's voice to train the transfer function. In addition, our attack can reconstruct the voice of most targets in a short time, which other spoofing attacks cannot achieve.

6 Conclusion

In this paper, we conducted the first comprehensive and systematic research on voiceprint reconstruction, by proposing a novel, efficient voiceprint re-constructor, called Voxstructor. At the same time, our voiceprint reconstruction attack was verified under 36 attack scenarios. This paper not only reveals the high sensitivity of voiceprint template through a large number of experiments, but also has the following significance:

- Voxstructor can carry out high simulation and batch spoofing attack on speaker recognition system. And it automatically completes the attack end-to-end without human participation.
- Voxstructor can be used to measure the effect of voiceprint privacy protection method; it can also be used to measure privacy in voiceprint. For noise-added privacy-preserving schemes, Voxstructor can also reconstruct the voice sample very well and achieve a high pass rate.
- Voxstructor can be used in computer forensics. This technology can also restore the voice of the suspect from the voiceprint of the suspect to provide evidence or clues for the police.

Acknowledgments

This work was supported by National Natural Science Foundation of China (61972304 and 61932015), National Natural Science Foundation of Shaanxi Province (2019ZDLGY12-02), and Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. G. Chen, S. Chen, L. Fan, X. Du, Z. Zhao, F. Song, and Y. Liu. Who is real bob? adversarial attacks on speaker recognition systems. In *2021 2021 IEEE Symposium on Security and Privacy (SP)*, pages 55–72, Los Alamitos, CA, USA, may 2021. IEEE Computer Society.
2. Qi Li, Hui Zhu, Ziling Zhang, Rongxing Lu, Fengwei Wang, and Hui Li. Spoofing attacks on speaker verification systems based generated voice using genetic algorithm. In *ICC 2019-2019 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2019.
3. Lei Zhang, Yan Meng, Jiahao Yu, Chong Xiang, Brandon Falk, and Haojin Zhu. Voiceprint mimicry attack towards speaker verification system in smart home. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pages 377–386. IEEE, 2020.
4. Lian Huang and Chi-Man Pun. Audio replay spoof attack detection by joint segment-based linear filter bank feature extraction and attention-enhanced densenet-bilstm network. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:1813–1825, 2020.

5. Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez-Moreno, and Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 4485–4495, 2018.
6. Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Trans. Speech Audio Process.*, 19(4):788–798, 2011.
7. David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333. IEEE, 2018.
8. Hee Soo Heo, Bong-Jin Lee, Jaesung Huh, and Joon Son Chung. Clova baseline system for the voxceleb speaker recognition challenge 2020. *arXiv preprint arXiv:2009.14153*, 2020.
9. Guangcan Mai, Kai Cao, Pong C. Yuen, and Anil K. Jain. On the reconstruction of face images from deep face templates. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(5):1188–1202, 2019.
10. Ji-won Seong, Wookey Lee, and Suan Lee. Multilingual speech synthesis for voice cloning. In Herwig Unger, Jinho Kim, U Kang, Chakchai So-In, Junping Du, Walid Saad, Young-guk Ha, et al., editors, *IEEE International Conference on Big Data and Smart Computing, BigComp 2021, Jeju Island, South Korea, January 17-20, 2021*, pages 313–316. IEEE, 2021.
11. Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE, 2018.
12. Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.
13. Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.
14. Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.
15. Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December 2011. IEEE Catalog No.: CFP11SRW-USB.
16. Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han. In defence of metric learning for speaker recognition. In *Interspeech*, 2020.
17. Phillip L De Leon, Michael Pucher, Junichi Yamagishi, Inma Hernaez, and Ibon Saratxaga. Evaluation of speaker verification security and detection of hmm-based synthetic speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(8):2280–2290, 2012.

18. Dibya Mukhopadhyay, Maliheh Shirvanian, and Nitesh Saxena. All your voices are belong to us: Stealing voices to fool humans and machines. In *European Symposium on Research in Computer Security*, pages 599–621. Springer, 2015.

Appendix

Experimental results about Inter-utterance case

The pass rates of Voxstructor, RTVC and random guessing under inter-utterance case are shown in Table 13. The pass rate of Voxstructor is close to that of RTVC with speech as direct input and significantly higher than that of the two random guessing schemes. These results illustrate that Voxstructor is still valid under inter-utterance case.

Table 13. Pass rate of Voxstructor, RTVC and random guessing under inter-utterance case.

Model	**SVS**	**Threshold**	**L1L**	**RTVC**	**Rand_vector**	**Rand-wav**
I2E	**IV-PLDA**	-1.000	48.133	57.194	1.782	0.743
		-6.000	83.772	88.186	11.039	7.480
		3.000	18.366	26.122	0.228	0.058
	IV-COS	0.060	28.706	33.622	11.659	15.747
		-0.020	80.764	83.648	65.292	66.840
		0.128	4.369	6.389	0.562	1.644
X2E	**XV-PLDA**	-3.000	44.944	49.326	3.188	2.822
		-8.000	71.989	74.143	16.625	15.501
		1.000	23.544	28.997	0.403	0.292
	XV-COS	0.670	71.729	69.077	0.000	0.037
		0.530	95.758	94.677	0.000	0.732
		0.770	23.293	23.844	0.000	0.000
R2E	**RN-EU**	80.000	49.968	51.723	0.514	12.349
		87.000	80.064	80.938	5.790	51.909
		75.000	28.112	28.818	0.016	1.713
	RN-COS	0.350	40.027	46.288	0.000	1.002
		0.390	24.624	29.915	0.000	0.095

The pass rates of three loss functions under inter-utterance case are shown in Table 14. The pass rates of three mapping models under inter-utterance case are shown in Table 15. The pass rates of text-independent reconstructed voice under inter-utterance case are shown in Table 16. These results show that the discussion we made in the main text for the intra-utterance case is also applicable in the inter-utterance case.

By comparing the results in both cases, the pass rate in the intra-utterance case is higher than that in the inter-utterance case. This once again shows that there are still relatively large differences even for different voices of the same person, and short voice sample is still not a good source to model a person's speech characteristics in SVSs.

Table 14. Pass rates of three loss function under inter-utterance case(%)

Model	SVS	Threshold	L1L	MSE	Smooth
I2E	IV-PLDA	-1.000	48.133	46.456	48.027
		-6.000	83.772	81.459	83.756
		3.000	18.366	17.528	17.294
	IV-COS	0.060	28.706	28.955	29.194
		-0.020	80.764	81.437	80.488
		0.128	4.369	4.343	4.894
X2E	XV-PLDA	-3.000	44.944	44.462	43.464
		-8.000	71.989	71.162	71.013
		1.000	23.544	23.698	22.446
	XV-COS	0.670	71.729	69.852	70.053
		0.530	95.758	95.970	96.124
		0.770	23.293	20.827	21.193
R2E	RN-EU	80.000	49.968	50.917	48.059
		87.000	80.064	81.400	79.539
		75.000	28.112	27.736	25.361
	RN-COS	0.350	40.027	40.308	40.758
		0.390	24.624	25.027	25.180

Table 15. Pass rates of three mapping models under inter-utterance case(%)

Model	SVS	Threshold	L1L	L1LNO	CONV
I2E	IV-PLDA	-1.000	48.133	8.313	46.822
		-6.000	83.772	21.332	84.371
		3.000	18.366	1.650	16.218
	IV-COS	0.060	28.706	13.547	29.517
		-0.020	80.764	64.862	80.323
		0.128	4.369	1.071	4.592
X2E	XV-PLDA	-3.000	44.944	1.873	32.446
		-8.000	71.989	6.626	61.347
		1.000	23.544	0.408	14.292
	XV-COS	0.670	71.729	21.797	62.306
		0.530	95.758	28.823	95.180
		0.770	23.293	8.187	14.825
R2E	RN-EU	80.000	49.968	2.126	47.879
		87.000	80.064	11.893	78.298
		75.000	28.112	0.477	25.764
	RN-COS	0.350	40.027	1.803	38.234
		0.390	24.624	0.530	23.712

Table 16. Pass rates of text-independent reconstructed voice under inter-utterance case

Model	SVS	Threshold	L1L	L1L-TEXT2	TRVC	TRVC-TEXT2
I2E	IV-PLDA	-1.000	48.133	45.332	57.194	54.578
		-6.000	83.772	80.085	88.186	86.462
		3.000	18.366	16.881	26.122	23.947
	IV-COS	0.060	28.706	26.808	33.622	33.473
		-0.020	80.764	80.398	83.648	84.093
		0.128	4.369	3.197	6.389	5.483
X2E	XV-PLDA	-3.000	44.944	40.801	49.326	45.003
		-8.000	71.989	65.767	74.143	69.220
		1.000	23.544	21.814	28.997	25.973
	XV-COS	0.670	71.729	77.200	69.077	71.898
		0.530	95.758	96.739	94.677	95.599
		0.770	23.293	28.181	23.844	24.173
R2E	RN-EU	80.000	49.968	78.287	51.723	76.485
		87.000	80.064	95.594	80.938	94.639
		75.000	28.112	53.218	28.818	52.195
	RN-COS	0.350	40.027	42.815	46.288	46.760
		0.390	24.624	26.007	29.915	30.636