CATCH YOU AND I CAN: Revealing Source Voiceprint Against Voice Conversion

Jiangyi Deng[1], Yanjiao Chen[1], Yinan Zhong[1], Qianhao Miao[1], Xueluan Gong[2], Wenyuan Xu[1]

[1]*Zhejiang University,* [2]*Wuhan University*

Abstract

Voice conversion (VC) techniques can be abused by malicious parties to transform their audios to sound like a target speaker, making it hard for a human being or a speaker verification/identification system to trace the source speaker. In this paper, we make the first attempt to restore the source voiceprint from audios synthesized by voice conversion methods with high credit. However, unveiling the features of the source speaker from a converted audio is challenging since the voice conversion operation intends to disentangle the original features and infuse the features of the target speaker. To fulfill our goal, we develop REVELIO, a representation learning model, which learns to effectively extract the voiceprint of the source speaker from converted audio samples. We equip REVELIO with a carefully-designed differential rectification algorithm to eliminate the influence of the target speaker by removing the representation component that is parallel to the voiceprint of the target speaker. We have conducted extensive experiments to evaluate the capability of REVELIO in restoring voiceprint from audios converted by VQVC [54], VQVC+ [63], AGAIN [9], and BNE [34]. The experiments verify that REVELIO is able to rebuild voiceprints that can be traced to the source speaker by speaker verification and identification systems. REVELIO also exhibits robust performance under inter-gender conversion, unseen languages, and telephony networks.

1 Introduction

The nets of Heaven are wide, but nothing escapes its grasp.

Lao Tzu's Tao Te Ching [33]

Voice conversion, a commonly-used speech synthesis technique, enables a person to transform their voice to sound like another person without changing the linguistic content. While the initial incentive for voice conversion could be simply novelty and curiosity, the technological breakthrough of deep learning has made voice conversion available for real-life applications, such as voice dubbing for movies [73], aids for the speech-impaired [60], voice mimicry [65] and disguise [24]. Unfortunately, when falling into the wrong hand, voice conversion may be used to carry out misdeeds. In 2020, the U.S. district court of Columbia tried a case where hackers had heisted $40 million by calling company managers using "deep voice" technology to simulate the voice of the Director [5].

Given the high frequency of phone calls nowadays and the ability to retain phone recordings, it is unsurprising that voice analysis has become a key tool for criminal forensics. However, audio data processed by speech synthesis techniques such as voice conversion may compromise the integrity of voice forensics. In particular, voice conversion distorts the features of the source speaker, making it difficult, if not impossible, to identify the true speaker. Existing works mainly focus on determining whether an audio sample is genuine or fake [6–8, 11, 20, 31, 35, 42, 53, 57, 61, 68, 71], but cannot trace the source of a fake audio sample. A few previous works [30, 62, 74] have attempted to restore transformed voices, but they are restricted to traditional simple frequency-domain voice transformations (e.g., pitch scaling and vocal tract length normalization) [62, 74] or man-made disguises (e.g., mimicking) [30]. As far as we know, there is a lack of success in the case of VC that adopts complex learning models [74].

In this paper, we propose the first effective approach to restore the voiceprint of the source speaker from audio processed by voice conversion techniques. Our developed system, REVELIO[1], aims to re-construct the original personal features of the converted audio such that the extracted voiceprint can match the source speaker via a speaker verification or identification model. In the case of a phone scam, as shown in Figure 1, REVELIO may assist law enforcement officials in investigations.

Nonetheless, recovering an identifiable voiceprint from

---

[1]The Revelio Charm is taught by Professor McGonagall in Transfiguration Classes in Hogwarts. The Revelio Charm reveals the true form of things or makes the invisible visible again [46].
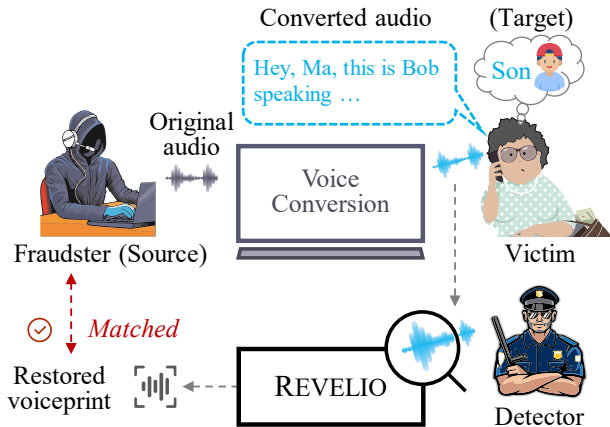
Figure 1: Application scenario of REVELIO. A fraudster uses voice conversion to transform their voice to sound like the target person, e.g., family members or acquaintances of the victim. REVELIO helps restore the voiceprint of the source speaker for investigation purposes.

Figure 2: Illustration of the process of voice conversion. The original audio sample $\mathbf{x}$ is disentangled into the content and the phonetic features. Then the original content and the phonetic features of the target speaker are synthesized into a new speech.

voice-converted (VC) audio with high credit is a challenging task. There are two major difficulties that we need to overcome. (1) The voice conversion operation can transform the audio of a source speaker to different target speakers and the audio of different source speakers to the same target speaker, making it inaccurate to identify the source speaker based on a simple classification model. To deal with this problem, we utilize representation learning to rebuild rather than re-labeling the voiceprint of the source speaker. The key idea is for the representation learning model to learn to generate the embedding of the source speaker voiceprint given the input of a VC audio of the same speaker. To train a well-performed representation learning model, we construct a large dataset consisting of a total of 6,779,000 samples converted by four popular VC techniques with 9,691 source speakers. (2) The features contained in VC audio are a mixture of the source speaker features and the target speaker features, and, in large part, the target speaker features. To dig the hidden source speaker features buried under the target speaker features, we design a novel differential rectification algorithm. The main intuition is to remove the component of the extracted voice feature that is parallel to the feature of the target speaker and only keep the orthogonal component. In this way, the influence of the target speaker is expected to be reduced to the maximum extent. The voiceprint recovered by REVELIO is to be compared with the voiceprint pool collected by the police to determine the identity of the source speaker[2]. Note that REVELIO only recovers the voiceprint of the source speaker of the audio

---

[2]In an open-world, the voiceprint of the source speaker may not be collected, i.e., no match between the recovered voiceprint and the voiceprint in the pool (the identity of the source speaker cannot be determined) but does not affect the fact that REVELIO obtains the voiceprint of the source speaker.
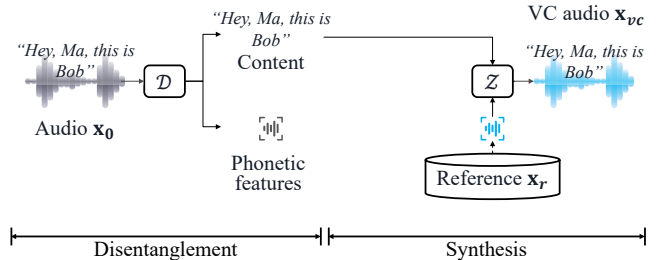
input of the VC, but cannot guarantee that the voiceprint is organic if the input of VC is already manipulated..

We have conducted extensive experiments to evaluate the effectiveness of REVELIO in restoring voiceprint from four popular voice conversion techniques, i.e., VQVC [54], VQVC+ [63], AGAIN [9], and BNE [34]. We demonstrate that the restored voiceprint can be correctly identified as the source speaker by speaker verification or identification systems with more than 95% accuracy (the source speaker is even unseen in the training dataset of REVELIO). An English-trained REVELIO model is shown to be able to recover voiceprint of German-, French- and Spanish-speaking VC audios. We also show by experiments that when the VC audios are contaminated by noises from telephony codecs and 8k/4kHz subsampling (from 16kHz), the restoration capability of REVELIO is only slightly degraded. We study four codecs for public switched telephone network (PSTN) and voice over Internet protocol (VoIP) network in §5.7.

We summarize our main contributions as follows.

- We propose the first effective approach to perform voiceprint restoration from audios processed by voice conversion techniques, which may be added as a tool of voice forensics to help trace and identify the source speaker of an audio.

- We develop a novel and effective representation learning model to extract identifiable voice characteristics of the source speaker by removing the features that are highly related to the target speaker.

- We conduct extensive experiments to verify the effectiveness and robustness of our method, with a generated voice conversion dataset using four voice conversion methods featuring over 9,600 speakers.
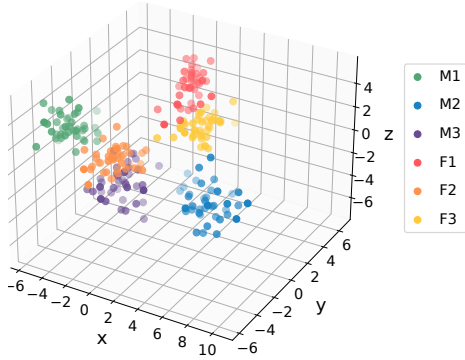
Figure 3: Disentangled content samples of different speakers by BNE [34].

2 Background

2.1 Voice Conversion

Audio data contains two kinds of information, i.e., speech contents and phonetic features. Speech contents refer to linguistic information, i.e., "what are the words spoken" and phonetic features refer to the way the speech contents are conveyed, i.e., "how are the words spoken." Voice conversion (VC) alters the speaker-related phonetic features of an audio sample but maintains the linguistic information and non-speaker-related phonetic features (e.g., emotion). The objective of VC is to make human listeners mis-identify the speaker of a VC-processed audio sample as the target person (e.g., family members or friends), while the original audio sample is actually articulated by a different person (e.g., a fraudster). To achieve this goal, VC techniques attempt to replace the phonetic features of the source speaker with those of the target speaker during processing.

As shown in Fig. 2, given an input audio sample $\mathbf{x}_0$ that features a source speaker, the disentanglement module $\mathcal{D}$ first splits $\mathbf{x}_0$ into two components: one mostly contains the speech content, and the other mostly contains personal phonetic features of the source speaker. Then, the content component is synthesized with a set of reference audios $\mathbb{X}_R = \{\mathbf{x}_r\}$ that features the target speaker in the synthesis module $\mathcal{Z}$. In this way, the converted audio $\mathbf{x}_{vc}$ sounds like the target speaker but preserves the content of $\mathbf{x}_0$.

Based on the availability of parallel reference audio samples, VC techniques can be categorized into parallel VC and non-parallel VC. Parallel VC needs to be trained with the reference sample $\mathbf{x}_r$ that has exactly the same speech content as the original sample $\mathbf{x}_0$ (i.e., the parallel corpus). Non-parallel VC can be trained with reference samples that have different speech contents from the original sample. In comparison, parallel VC is less practical than non-parallel VC due to the need for a parallel corpus for training, but the converted samples of parallel VC generally have higher quality than those of non-parallel VC also due to the benefit of the parallel corpus.

Rationale of voiceprint recovery. The imperfect disentanglement enables the recovery of the voiceprint of the source speaker. As shown in Figure 2, the converted audio is synthesized from the content of the input audio $\mathbf{x}_0$ and the phonetic features of the target speaker. As the disentanglement module cannot perfectly separate content and phonetic features, the phonetic features of the source speaker are leaked into the content parts and synthesized into the converted audio. To demonstrate this, we collect audio samples from six volunteers (three males and three females), each speaking "Hello World" for 50 times. We utilize the disentanglement modules of four VCs to obtain the content of these samples, and perform a dimensional reduction via Linear Discriminant Analysis (LDA) [18]. We present the results in Figure 3 and Figure 10 in the Appendix of the extended version [12]. In addition, we train a linear Support Vector Machine (SVM) to classify the disentangled content samples and achieve an accuracy of 100.0%, 95.7%, 88.0%, 100.0% for the four VCs respectively, which indicates that the content parts of different speakers are almost linearly separable. The visualization and the classification results verify that VC audios still carry information of the source speaker.

2.2 Speaker Verification & Identification

There are two mainstream systems for examining speaker identity: speaker verification and speaker identification. The former tries to determine whether the speaker of an audio sample is the enrolled speaker or not, and the latter attempts to identify who the speaker (of an audio sample) is out of a set of enrolled speakers. To fulfill the above goals, the essential building block of both speaker verification and speaker identification systems is the voiceprint extraction method.

Voiceprint extraction. Given an audio sample $\mathbf{x} \in \mathbb{R}^{1 \times t}$ of an arbitrary length of t, a voiceprint extractor $\mathcal{V} : \mathbb{R}^{1 \times t} \rightarrow \mathbb{R}^{1 \times v}, \forall t \in (0, +\infty)$ maps $\mathbf{x}$ into a fixed-length voiceprint vector, representing the speaker characteristics of the audio. There are mainly two ways to build the voiceprint extractor, i.e., statistical models and Deep Neural Network (DNN)-based models. Gaussian mixture model (GMM) is a traditional statistical model to extract ivector voiceprints. ivector-PLDA [43] is a popular speaker verification/identification implementation that matches ivector voiceprints via probabilistic linear discriminant analysis (PLDA). X-vector [49], Deep Speaker [32], ECAPA-TDNN [14] are DNN-based voiceprint extractors, which outperform GMM as DNNs are more effective in extracting feature representations from large-scale voice datasets. Among them, ECAPA-TDNN is the state-of-the-art voiceprint extractor based on time delay neural network (TDNN). ECAPA-TDNN outputs a 192-dimensional voiceprint vector, i.e., $v = 192$. The extracted voiceprints of legitimate users during enrolment are stored as the *speaker model*, which are later compared with the voiceprint of the input audio for speaker verification or identification.

Speaker verification. In a speaker verification system, there is a single enrolled user with the voiceprint $\mathbf{v}$. The speaker verification system aims to verify whether the speaker of an input audio is the enrolled user (e.g., the fraudster) or not. Given an input audio $\mathbf{x}$, the similarity score between the extracted voiceprint and the enrolled voiceprint is computed as $S(\mathcal{V}(\mathbf{x}), \mathbf{v})$. If the score is higher than the threshold ϕ, $\mathbf{x}$ is considered to be matched with the enrolled speaker.

$$O(\mathbf{x}) = \begin{cases} \text{matched}, & \text{if } S(\mathcal{V}(\mathbf{x}), \mathbf{v}) \geq \phi, \\ \text{not matched}, & \text{otherwise}, \end{cases}$$

where $O(\mathbf{x})$ is the output of the verification system.

Speaker identification. In a speaker identification system, there is a set of enrolled users with the voiceprints $\{\mathbf{v}_i\}_{i \in I}$. Speaker identification aims to determine which enrolled user the speaker of the input audio is most likely to be. In particular, the voiceprint of the input audio is compared with each speaker model to calculate the similarity scores $S(\mathcal{V}(\mathbf{x}), \mathbf{v}_i), i \in I$. The output index of the enrolled users is the one with the highest score.

$$O(\mathbf{x}) = \arg\max_{i \in I} S(\mathcal{V}(\mathbf{x}), \mathbf{v}_i),$$

Without voice conversion, the source speaker of an audio sample should be correctly verified or identified by a speaker verification or identification system that has the speaker's voiceprint enrolled. Nonetheless, after voice conversion, the phonetic features of the source speaker are compromised, thus the VC-processed audio sample will evade the detection of speaker verification or identification systems.

This inspires us to seek a way to recover the voiceprint of the source speaker, which is helpful in forensic investigations. We will demonstrate the effectiveness of REVELIO in restoring voiceprints for both speaker verification and speaker identification tasks of different forensic investigation scenarios in §5.

2.3 System Model

Without loss of generality, we refer to the party who uses voice conversion to convert their voiceprint to a target speaker as the *Dodger* and the party who wants to restore the original voiceprint as the *Detector*.

2.3.1 Dodger Model

We assume that the dodger has a corpus of reference audios $\mathbb{X}_R = \{\mathbf{x}_r\}$ of the target speaker whose real voiceprint is $\mathbf{v}_r$, based on which the dodger uses voice conversion techniques to convert their original audio sample $\mathbf{x}_0$ into $\mathbf{x}_{vc}$ (referred to as the VC audio) that sounds like the target speaker. The voiceprint of the dodger is denoted as $\mathbf{v}_d$. Note that we assume that the original audio is uttered by the dodger. The case where the original audio is generated by text-to-speech (TTS)

techniques is out of our scope since TTS leaves no source voiceprint to be recovered.

2.3.2 Detector Model

We assume that the detector obtains the VC audio $\mathbf{x}_{vc}$ and is aware that $\mathbf{x}_{vc}$ is processed by voice conversion techniques but may not know the particular VC method being used. The detector knows the target speaker by listening to the VC audio and may obtain an audio sample of the target speaker (referred to as the evidence audio $\mathbf{x}_e$) during investigations. The evidence audio $\mathbf{x}_e$ can be non-parallel with the VC audio $\mathbf{x}_{vc}$. Having collected the VC audio $\mathbf{x}_{vc}$ and the evidence audio $\mathbf{x}_e$, the detector attempts to recover the voiceprint of the source speaker of $\mathbf{x}_{vc}$ that can be verified by a speaker verification model (e.g., the detector has a single suspect) or be identified by a speaker identification model (e.g, the detector has a group of suspects).

3 Problem Formulation & Materialization

In this section, we first formulate the voiceprint restoration as a representation learning problem. Then we materialize the representation learning problem in three different scenarios.

3.1 Representation Learning Problem Formulation

Given the raw audio $\mathbf{x}_0 = [x_1, \cdots, x_t] \in \mathbb{R}^{1 \times t}$ of length t, the VC audio $\mathbf{x}_{vc}$ is generated via voice conversion techniques to mimic the target speaker. We have $\mathbf{x}_{vc} = \psi(\mathbf{x}_0, \mathbb{X}_R)$, where $\psi(\cdot, \cdot)$ denotes the conversion function of the VC technique and $\mathbb{X}_R = \{\mathbf{x}_r\}$ contains reference samples of the target speaker. As a result of voice conversion, the similarity score between $\mathcal{V}(\mathbf{x}_{vc})$ and $\mathbf{v}_r$ greatly increases, i.e., $S(\mathcal{V}(\mathbf{x}_{vc}), \mathbf{v}_r) \gg S(\mathcal{V}(\mathbf{x}_0), \mathbf{v}_r)$, and the similarity score between $\mathcal{V}(\mathbf{x}_{vc})$ and $\mathbf{v}_d$ drops to a great extent, i.e., $S(\mathcal{V}(\mathbf{x}_{vc}), \mathbf{v}_d) \ll S(\mathcal{V}(\mathbf{x}_0), \mathbf{v}_d)$, making it difficult (if not impossible) for the detector to unveil the identity of the dodger.

The objective of REVELIO is to restore $\mathbf{v}_d$ from $\mathbf{x}_{vc}$ with high probability, which can be formulated as a representation learning problem $\mathcal{E} : \mathbb{R}^{1 \times t} \to \mathbb{R}^{1 \times v}, \forall t \in (0, +\infty)$. The representation learning model $\mathcal{E}$ aims at minimizing the distance between the output vector (i.e., the reconstructed voiceprint) and the voiceprint of the dodger.

$$\max \; S\big(\mathcal{E}(\mathbf{x}_{vc}, *|\theta), \mathbf{v}_d\big), \tag{1}$$

where θ is the parameters of the representation learning model $\mathcal{E}$ and $*$ denotes any extra information available to the detector, e.g., an evidence audio $\mathbf{x}_e$ from the target speaker.

To train the representation learning model $\mathcal{E}$, we resort to the final objective of identifying the dodger. More specifically, we integrate $\mathcal{E}$ into a classification model that aims to classify the reconstructed voiceprint (i.e., the output of $\mathcal{E}$) into
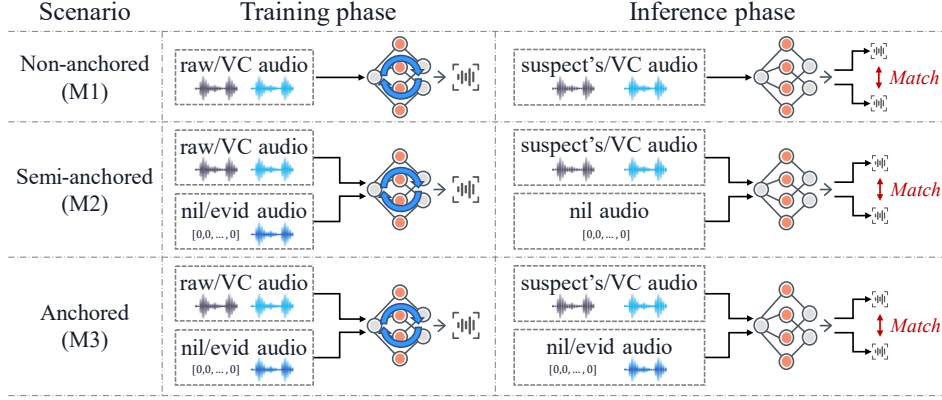
Figure 4: Three materializations of REVELIO. In the non-anchored case M1, the evidence audios of the target speaker are unavailable. In the semi-anchored case M2, the evidence audios of the target speaker are available during the training phase, but unavailable during the inference phase. In the anchored case M3, evidence audios of the target speaker are available during the training and the inference phases. In the semi-anchored and the anchored cases, the nil audio $\mathbf{0}$ is used as a placeholder when the evidence audio is unavailable.

the label that represents the dodger. In other words, $\mathcal{E}$ performs feature extraction for the classification task, followed by which a fully-connected layer is attached to produce the final classification result. We denote the classification model as $\mathcal{C} = [\mathcal{E} \oplus \mathcal{F}]$, where $\mathcal{F}$ is the final layer.

The loss function of the classification model is defined as the cross entropy loss w.r.t the model parameter θ.

$$\mathcal{L}(\theta) \triangleq -\sum_i y_i \log(p_i), \tag{2}$$

where $p_i = \mathcal{C}(\mathbf{x}_{vc}, *|\theta)$ is the output confidence score of the classification model regarding the i-th class. $y_i = 1$ if i is the real dodger and $y_i = 0$ otherwise. $\mathcal{F}$ usually adopts the softmax activation function, i.e.,

$$p_i = \frac{\exp(\mathcal{E}_i(\mathbf{x}_{vc}, *))}{\sum_j \exp(\mathcal{E}_j(\mathbf{x}_{vc}, *))}, \tag{3}$$

where $\mathcal{E}_i(\mathbf{x}_{vc}, *)$ is the i-th output of $\mathcal{E}$. Note that softmax does not explicitly drive the voiceprint towards the dodger voiceprint $\mathbf{v}_d$ and away from the voiceprints of other suspects. We will enhance the discriminative ability of softmax in our design in §4.5.

After training the classification model $\mathcal{C}$, the resulting submodel $\mathcal{E}$ can be obtained as our voiceprint restoration model.

3.2 Materialization Scenarios

According to the availability of the evidence audio $\mathbf{x}_e$ during the training phase and the inference phase, we materialize REVELIO in three cases, as shown in Figure 4.

- **Non-anchored case**. In a non-anchored case, the evidence audio $\mathbf{x}_e$ that features the target speaker of the

VC audio $\mathbf{x}_{vc}$ is unavailable, both during the training phase and the inference phase. $\mathcal{E}$ has a single input in this case. During the training phase, $\mathcal{E}$ learns to output the ground-truth voiceprint $\mathbf{v}_d$ given a training VC audio $\mathbf{x}_{vc}^{train}$ or a training raw audio $\mathbf{x}_0^{train}$ (both originates from the dodger).

- **Semi-anchored case**. In a semi-anchored case, the evidence audio $\mathbf{x}_e$ that features the target speaker of the VC audio $\mathbf{x}_{vc}$ is available during the training phase but unavailable during the inference phase. $\mathcal{E}$ has dual inputs in this case. During the training phase, $\mathcal{E}$ learns to output the ground-truth voiceprint $\mathbf{v}_d$ given a training VC-evidence audio pair $[\mathbf{x}_{vc}^{train}, \mathbf{x}_e^{train}]$ or a training raw-nil audio pair $[\mathbf{x}_0^{train}, \mathbf{0}]$.

- **Anchored case**. The anchored case M3 is the same as the semi-anchored case M2 except that during the inference phase,

The three above cases can be applied to different scenarios. For the non-anchored case, the detector only needs a training dataset of VC audios and their ground-truth labels of the dodger identity during training and a piece of VC audio during inference. In the semi-anchored case, the detector also needs to obtain the corresponding evidence audio for each VC audio during training. In the anchored case, the detector further should have the evidence audio for the VC audio during inference. Thus, in the anchored case we assume that the detector knows the input audio is processed by VC (possibly via existing VC detection techniques), and obtains an evidence audio from the target for inference. But we do not make this assumption in the non-anchored or the semi-anchored cases. Note that given a non-VC audio samples, REVELIO is still able to recover the correct voiceprint. In the semi-anchored

Figure 5 diagram (left column):

VC audio: **x**   evid audio: $\mathbf{x}_e$

Feature extraction ①

Filter Bank
$f \times T_M$ $f \times T_N$
Conv1D + ReLU + BN (k=5, d=1)
$c \times T_i$
SE-Res2Block (k=3, d=2)
$c \times T_i$
SE-Res2Block (k=3, d=3)
$c \times T_i$
SE-Res2Block (k=3, d=4)
$3c \times T_i$

Differential rectification ②

M $3c \times T_M$ $3c \times T_N$ **N**
Residual-Orthogonalization Block
$3c \times T_M$

Dimension normalization ③

Conv1D + ReLU (k=1, d=1)
$3c \times T_M$
Attentive Stat Pooling + BN
$6c \times 1$
FC + BN

Voiceprint enhancement ④
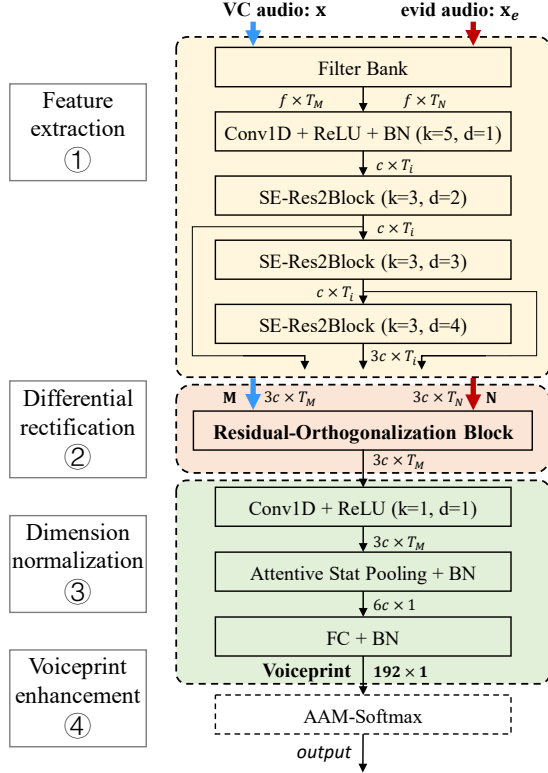
Voiceprint 192×1
AAM-Softmax
output

Figure 5: REVELIO model architecture. The feature extraction block ① extracts crude voiceprint from the VC audio. The differential rectification block ② eliminates the influence of the target speaker based on the evidence audio. The dimension normalization block ③ reshapes the voiceprint into a fixed-length vector. The voiceprint enhancement block ④ further alters the voiceprint to be more discriminative from the voiceprints of other suspects.
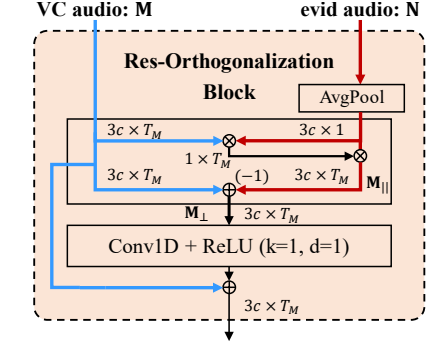
and the anchored cases, the nil audio **0** is used as a placeholder when the evidence audio is unavailable. Note that, in all three cases, the dodger to be identified or the target speaker does not have to be in the training set of REVELIO, i.e., REVELIO can handle unseen dodger and unseen target speaker.
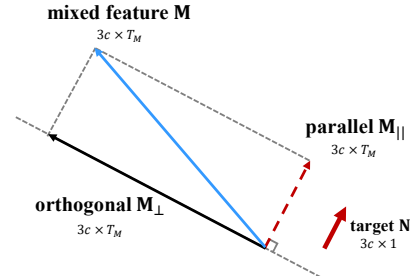
4 REVELIO: Construction Details

4.1 Model Architecture Overview

As shown in Figure 5, the voiceprint restoration model $\mathcal{E}$ consists of four blocks.

- *Feature extraction* ①. The feature extraction block extracts a crude voiceprint representation from the input VC audio.
- *Differential rectification* ②. The differential rectification block cleanses the influence of the target speaker's voiceprint by rectifying the direction of the voiceprint to be orthogonal to that of the target speaker.

Figure 6 diagrams (right column):

VC audio: **M**   evid audio: **N**

Res-Orthogonalization Block
AvgPool
$3c \times T_M$ $3c \times 1$
$1 \times T_M$
$3c \times T_M$ (-1) $3c \times T_M$ $\mathbf{M}_\parallel$
$\mathbf{M}_\perp$ $3c \times T_M$
Conv1D + ReLU (k=1, d=1)
$3c \times T_M$

(a) Residual Orthogonalization Block architecture

mixed feature **M**
$3c \times T_M$

parallel $\mathbf{M}_\parallel$
$3c \times T_M$

orthogonal $\mathbf{M}_\perp$
$3c \times T_M$

target **N**
$3c \times 1$

(b) Extracting Invariant Features by Orthogonalization

Figure 6: The design of the Residual Orthogonalization Block. The block decomposes the voiceprint into two components that are parallel and orthogonal to the voiceprint of the target speaker. By subtracting the parallel component, the influence of the target speaker voiceprint is mitigated.

- *Dimension normalization* ③. The dimension normalization block normalizes the varied-length input into a fixed-length output.
- *Voiceprint enhancement* ④. The voiceprint enhancement block further reduces the distance between the generated voiceprint and the ground-truth voiceprint and increases the distance between the generated voiceprint and other voiceprints, thus enhancing the discriminative capability of speaker verification/identification systems.

Note that differential rectification can only be performed with the assistance of evidence audios. Therefore, in the non-anchored case, $\mathcal{E}$ is formed by the feature extraction, dimension normalization, and voiceprint enhancement blocks only.

4.2 Feature Extraction

We leverage the state-of-the-art time delay neural network (TDNN)-based model for feature extraction [14].

As shown in Figure 5, a filter bank layer first transforms time-domain audios into "FBank" features, which are Mel-frequency-domain features of dimension $f \times T$, where f is the number of Mel filters used and T is the number of frames produced. FBank feature is widely used as input feature for newly

speaker identification systems [14, 32, 49] because it carries more information than MFCC, and is more aligned with human auditory than LPCC [26]. Then, a TDNN layer and three squeeze-excitation SE-Res2Blocks are used to extract temporal features of audios with an expanding receptive field along the time axis [19, 70]. A TDNN block consists of a TDNN (Conv1D) layer, followed by a rectified linear unit (ReLU) activation layer and a batch normalization (BN) layer. An SE-Res2Block consists of a TDNN layer, a Res2-Dilated-TDNN layer, another TDNN layer, and a squeeze-excitation layer. TDNNs are used to extract time-domain features, and squeeze-excitation layers model the global channel interdependencies. The convolutional kernel size and the spacing between kernel elements in the dilated convolution are denoted as k and d in Figure 5. Finally, the outputs of three SE-Res2Blocks are concatenated. The feature extraction block $\mathcal{E}_1$ converts the input audio into a feature map with $3c$ channels, where c is a hyperparameter. The larger the c is, the more features are extracted, but the more complex the model becomes.

In the non-anchored case, the result of the feature extraction block $\mathcal{E}_1(\mathbf{x}_{vc})$ is fed directly to the dimension normalization block. In the semi-anchored and the anchored cases, the feature extraction block outputs both $\mathbf{M} = \mathcal{E}_1(\mathbf{x}_{vc})$ and $\mathbf{N} = \mathcal{E}_1(\mathbf{x}_e)$, which are fed into the differential rectification block. Note that the dimensions of $\mathbf{M}$ and $\mathbf{N}$ are $3c \times T_M$ and $3c \times T_N$ respectively. In the semi-anchored and the anchored cases, if the first input is the raw audio of suspects, the feature extraction block will output $\mathbf{M} = \mathcal{E}_1(\mathbf{x}_0)$ and $\mathbf{N} = \mathcal{E}_1(\mathbf{0})$.

4.3 Differential Rectification

The differential rectification block is the key to restoring the voiceprint of the dodger with high quality by utilizing the evidence audio of the target speaker.

As mentioned in §1, due to the voice conversion process, the VC audio of the dodger may be close to the raw audio of the target speaker but faraway from the raw audio of the dodger in the feature space. Therefore, following the traditional feature extraction method (as ①) may not be able to generate discriminative features that are sufficient for the speaker verification/identification model to zero in on the dodger.

To tackle this problem, we propose to distill the *invariant* and intrinsic features of the dodger by removing the influence of the target speaker guided by the evidence audio. To realize this goal, we develop a novel Res-Orthogonalization Block (ROB) as shown in Figure 6(a), which decomposes the feature vector of the VC audio $\mathbf{M}$ into component $\mathbf{M}_{\parallel}$ that is parallel to the feature vector of the evidence audio $\mathbf{N}$ and $\mathbf{M}_{\perp}$ that is perpendicular to $\mathbf{N}$. We present a visualization of this process in Figure 6(b). Our intuition is that removing the parallel component $\mathbf{M}_{\parallel}$ from $\mathbf{M}$ may suppress the influence of the target speaker induced by VC. Note that only M2 and M3 have ROB, since M1 has not been designed to handle an evidence audio.

We first calculate the average direction of $\mathbf{N}$ along the time axis as

$$\bar{\mathbf{N}} = \frac{1}{T_N} \sum_{i=1}^{T_N} \mathbf{N}_i, \qquad (4)$$

where $\mathbf{N}_i$ is the i-th frame of $\mathbf{N}$.

Then we obtain the unit direction of $\mathbf{N}$ as

$$\bar{\mathbf{n}} = \frac{\bar{\mathbf{N}}}{|\bar{\mathbf{N}}| + \varepsilon} \qquad (5)$$

where the unit vector $\bar{\mathbf{n}}$ has a length of $3c$. To avoid NaN values in practice, a small constant (e.g. $\varepsilon = $1e-6) is added to the denominator.

We decompose the feature vector $\mathbf{M}$ into $\mathbf{M}_{\parallel}$ and $\mathbf{M}_{\perp}$ according to $\bar{\mathbf{N}}$ as

$$\begin{aligned}\mathbf{M}_{\parallel} &= \bar{\mathbf{n}}(\bar{\mathbf{n}}^{\mathrm{T}}\mathbf{M}), \\ \mathbf{M}_{\perp} &= \mathbf{M} - \mathbf{M}_{\parallel} = \mathbf{M} - \bar{\mathbf{n}}(\bar{\mathbf{n}}^{\mathrm{T}}\mathbf{M}) = (\mathbf{I} - \bar{\mathbf{n}}\bar{\mathbf{n}}^{\mathrm{T}})\mathbf{M},\end{aligned} \qquad (6)$$

where $\mathbf{I}$ is an identity matrix of size $3c \times 3c$.

Finally, we add TDNN-processed $\mathbf{M}_{\perp}$ to the original feature vector $\mathbf{M}$ in a residual learning fashion [21] as

$$\mathcal{E}_2(\mathbf{M}, \mathbf{N}) = \mathrm{TDNN}(\mathbf{M}_{\perp}) + \mathbf{M}, \qquad (7)$$

where $\mathcal{E}_2(\mathbf{M}, \mathbf{N})$ has the same size as $\mathbf{M}$. We show the effectiveness of our proposed differential rectification block in §5.2.

4.4 Dimension Normalization

The output of the differential rectification block $\mathcal{E}_2(\mathbf{M}, \mathbf{N})$ has a dimension of $3c \times T_M$, which depends on the length of the input VC audio t_m. To obtain a fixed-length voiceprint, the dimension normalization block $\mathcal{E}_3$ applies channel- and context-dependent statistics pooling to produce an integrated representation over all channels and frames. The output of the dimension normalization block will be treated as the reconstructed voiceprint of the dodger.

4.5 Voiceprint Enhancement

As mentioned in §3.1, training with normal softmax cannot yield contrasting embeddings that help speaker verification/identification models to differentiate the dodger and other suspects. Inspired by ArcFace [13] and ECAPA-TDNN [14], we utilize additive angular margin (AAM) loss instead of softmax loss as the output layer to explicitly force the recovered voiceprint to have high similarity with the dodger voiceprint and low similarity with the voiceprints of other suspects.

Through our experiments, we find that the entire voiceprint restoration model $\mathcal{E}$ is hard to converge during training. To boost the training efficiency, we utilize a pretrained ECAPA-TDNN model to initialize the feature extraction block $\mathcal{E}_1$.

This solution enables practical training of the model with limited computing resources.

In addition, to increase the robustness of REVELIO, we introduce data augmentation techniques in the training phase, including modifying speed, adding noises, randomly dropping frames. We evaluate the robustness of REVELIO in §5.7.

5 Evaluation

5.1 Setup

5.1.1 Prototype

We have implemented a prototype of REVELIO on the Pytorch [41] and SpeechBrain [45] platforms and trained the model according to Equation (2) using four NVIDIA 3090 GPUs. In the training phase, both the input VC audios and the evidence audios are re-sampled to 16 kHz and randomly trimmed to 6 seconds due to the memory limitation of GPUs, i.e., the length of the audio samples is $t_m = t_n = 16,000 \times 6$. For the model, we set the default configuration as $c = 1,024$. We detail the parameters of the blocks of the model in Table 10 in the Appendix.

In the training phase, we initialize the feature extraction block with the parameters of a pretrained ECAPA-TDNN [45] and use an Adam [29] optimizer to update the parameters of the model for 4∼8 epochs (until a satisfying validation loss is achieved or the loss can no longer decrease), with a learning rate of 1e-3, a weight decay (L_2 penalty) rate of 2e-6, and a batch size of 24.

5.1.2 Datasets and VC methods

To train and evaluate REVELIO, we generate a large-scale voice conversion dataset with four voice conversion methods, consisting of audios from 9,691 speakers. We implement four non-parallel one-shot voice conversion methods, i.e.,

- **VQVC** [64] is a classic disentanglement-based voice conversion method, which is based on vector quantization to disentangle speech contents and phonetic features. We utilize an unofficial implementation available on GitHub[3], and train the VC model on *train-clean-360* of Librispeech [40].

- **VQVC+** [63] combines VQVC and U-Net to produce high-quality audios. We utilize an official implementation available on GitHub[4], and train the VC model on *train-clean-360* of Librispeech [40].

- **AGAIN** [9] is an improvement of the VC method AdaIN [10] with an activation guidance method. We

---

[3]https://github.com/Jackson-Kang/VQVC-Pytorch. Note that there is no official implementation of VQVC.
[4]https://github.com/ericwudayi/SkipVQVC

Table 1: Performance comparison of REVELIO and baseline.

Method	Metrics[†]	B1	B2	B3	M1	M2	M3
VQVC	EER (↓)	31.39%	38.62%	47.95%	5.57%	4.40%	**4.08%**
	Top-1 ACC (↑)	6.86%	2.89%	3.53%	91.35%	93.11%	**93.21%**
	Top-5 ACC (↑)	25.16%	14.10%	14.10%	97.95%	**98.81%**	98.62%
	Top-10 ACC (↑)	42.89%	32.05%	26.60%	99.07%	**99.87%**	99.55%
VQVC+	EER (↓)	27.66%	41.47%	48.97%	3.95%	3.66%	**3.35%**
	Top-1 ACC (↑)	11.03%	5.45%	3.85%	95.03%	94.65%	**95.93%**
	Top-5 ACC (↑)	33.81%	19.87%	17.31%	99.23%	98.88%	**99.36%**
	Top-10 ACC (↑)	54.01%	33.65%	34.30%	99.65%	99.55%	**99.68%**
AGAIN	EER (↓)	22.55%	40.42%	42.79%	2.57%	1.94%	**1.87%**
	Top-1 ACC (↑)	21.80%	7.05%	4.17%	97.66%	98.59%	**99.07%**
	Top-5 ACC (↑)	51.03%	28.85%	16.67%	99.74%	99.90%	**99.94%**
	Top-10 ACC (↑)	68.88%	44.55%	31.73%	99.97%	99.97%	**100.0%**
BNE	EER (↓)	31.28%	45.06%	42.12%	5.14%	**3.58%**	3.78%
	Top-1 ACC (↑)	0.93%	2.89%	1.60%	92.66%	95.10%	**96.12%**
	Top-5 ACC (↑)	16.64%	15.71%	14.42%	99.52%	99.78%	**99.97%**
	Top-10 ACC (↑)	33.33%	27.89%	26.28%	99.97%	99.97%	**100.0%**

[†] ↑ indicates that the value is the higher, the better, and ↓ indicates that the value is the lower, the better. The best results are highlighted in bold.

utilize the official implementation[5], and train the model on *train-clean-360* of Librispeech.

- **BNE** (**BNE-Seq2seqMoL**) [34] is a state-of-the-art disentanglement-based voice conversion method that has better subjective hearing quality than previous VC methods (and also shown to be the most difficult for voiceprint restoration in our experiments). We utilize both the official implementation and the official pretrained model[6].

To prepare the training datasets for REVELIO, we apply VQVC, VQVC+, AGAIN, and BNE on three training subsets of LibriSpeech and VoxCeleb1&2 [36, 37] to generate a substantial number of VC audios. We also apply the above four VC methods to the *test-clean* subset of LibriSpeech for testing. Note that there is no overlap between any speakers (source or target) in the training set and the test set. We present the detailed information on dataset generation in Appendix A.

To evaluate the effectiveness of REVELIO in restoring voiceprint from audios of languages different from the language of the training set, we apply BNE on three datasets of other languages, including German, French, and Spanish. The dataset details are summarized in Table 7 in the Appendix.

5.1.3 Evaluation Metrics

Two metrics are used to evaluate the extracted voiceprint of REVELIO under the speaker verification model and the speaker identification model, respectively.

- **Equal Error Rate (EER)** is the rate at which false positive rate (FPR) equals false negative rate (FNR), which is a metric for the speaker verification task. A lower

---

[5]https://github.com/KimythAnly/AGAIN-VC
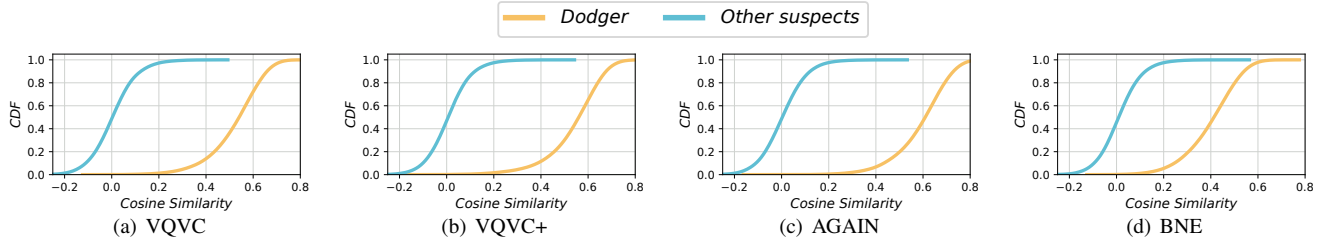[6]https://github.com/liusongxiang/ppg-vc

Figure 7: The distributions of cosine similarity scores between the restored voiceprint by REVELIO and the dodger or other suspects. The voiceprints are recovered by M3.

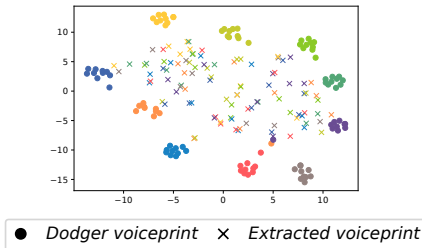

Dodger voiceprint × Extracted voiceprint

Figure 8: Distribution (visualized using t-SNE) of the dodger voiceprint and the extracted voiceprint by the baseline. Colors denote different dodgers. It is shown that the extracted voiceprints by the baseline are faraway from the voiceprint of the real dodger.

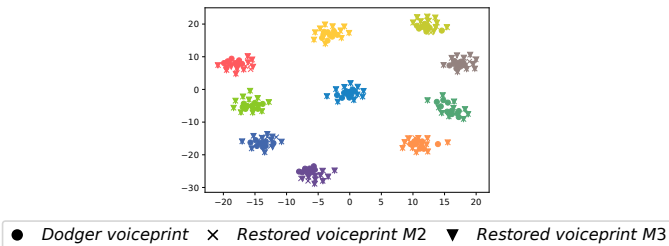● Dodger voiceprint × Restored voiceprint M2 ▼ Restored voiceprint M3

Figure 9: Distribution (visualized using t-SNE) of the dodger voiceprint and the restored voiceprint by REVELIO. Colors denote different dodgers. It is shown that the restored voiceprints by REVELIO are close to the voiceprint of the real dodger.

- EER denotes that the speaker verification can verify the identity of the dodger with higher precision, i.e., a better restoration performance of REVELIO.

- **Top-k Accuracy (Top-k ACC)** is the rate at which the correct label is among the top k labels predicted (ranked by similarity scores) by the speaker identification model. A higher Top-k ACC means that the restored voiceprint can better narrow the scope of the dodger.

5.1.4 Baseline

We utilize the state-of-the-art voiceprint extractor, ECAPA-TDNN, trained on VoxCeleb1&2 as our baseline B1. We also implement two existing works [62, 74] on voice recovery,

Table 2: Generalizability to unseen VCs.

Method	Metrics[†]	Audios converted by			
		VQVC	VQVC+	AGAIN	BNE
VQVC	EER (↓)	4.08%	7.96%	6.24%	36.79%
	Top-1 ACC (↑)	93.21%	77.31%	90.71%	4.26%
	Top-5 ACC (↑)	98.62%	95.06%	99.33%	26.35%
	Top-10 ACC (↑)	99.55%	98.08%	99.78%	45.61%
VQVC+	EER (↓)	6.46%	3.35%	2.67%	35.14%
	Top-1 ACC (↑)	85.64%	95.93%	98.27%	3.11%
	Top-5 ACC (↑)	97.34%	99.36%	99.90%	28.27%
	Top-10 ACC (↑)	99.10%	99.68%	100.0%	47.15%
AGAIN	EER (↓)	8.99%	5.82%	1.87%	31.23%
	Top-1 ACC (↑)	75.77%	86.25%	99.07%	2.40%
	Top-5 ACC (↑)	93.30%	96.70%	99.94%	27.44%
	Top-10 ACC (↑)	97.82%	99.33%	100.0%	46.70%
BNE	EER (↓)	30.35%	19.84%	12.11%	3.78%
	Top-1 ACC (↑)	20.77%	28.21%	50.42%	96.12%
	Top-5 ACC (↑)	52.69%	64.01%	83.75%	99.97%
	Top-10 ACC (↑)	71.19%	79.30%	93.33%	100.0%

[†] ↑ indicates that the value is the higher, the better, and ↓ indicates that the value is the lower, the better.

Table 3: Impact of VC audio length.

Metrics[†]	5s	10s	15s	20s	25s
EER (↓)	5.55%	3.71%	3.39%	3.78%	4.48%
Top-1 ACC (↑)	88.24%	92.31%	94.55%	96.12%	95.51%
Top-5 ACC (↑)	98.69%	99.62%	99.87%	99.97%	99.87%
Top-10 ACC (↑)	99.62%	99.87%	100.0%	100.0%	100.0%

[†] ↑ indicates that the value is the higher, the better, and ↓ indicates that the value is the lower, the better.

referred to as B2 and B3. We will show the baselines cannot recover the voiceprint of the source speaker of the audios processed by voice conversion.

5.2 Overall Effectiveness

In this part, we evaluate the effectiveness of REVELIO in restoring the voiceprint of the dodger. As shown in Table 7, we generate four large-scale training sets, i.e., *VQ-Train*, *V+-Train*, *AG-Train*, *BN-Train* with VQVC, VQVC+, AGAIN, and BNE respectively. As stated in §3.2, we have three materializations of REVELIO for the non-anchored (M1), semi-anchored (M2), and anchored (M3) cases. We train the three

Table 4: Impact of evidence audio length.

Metrics[†]	5s	10s	15s	20s	25s
EER (↓)	3.70%	3.67%	3.68%	3.78%	3.68%
Top-1 ACC (↑)	95.61%	95.67%	95.80%	96.12%	95.83%
Top-5 ACC (↑)	99.68%	99.81%	99.71%	99.97%	99.84%
Top-10 ACC (↑)	99.97%	99.94%	99.97%	100.0%	99.97%

[†] ↑ indicates that the value is the higher, the better, and ↓ indicates that the value is the lower, the better.

Table 5: Over-the-telephony robustness.

Method	Metrics[†]	μ-law	A-law	GSM[‡]	AMR[‡]	8kHz[‡]	4kHz[‡]
VQVC	EER (↓)	5.47%	5.39%	7.12%	9.54%	5.11%	15.78%
	Top-1 ACC (↑)	91.60%	91.76%	80.32%	69.90%	92.89%	49.23%
	Top-5 ACC (↑)	99.58%	99.68%	98.49%	92.76%	99.58%	82.92%
	Top-10 ACC (↑)	99.94%	99.94%	99.97%	97.47%	99.84%	92.37%
VQVC+	EER (↓)	4.60%	4.42%	5.23%	7.68%	4.56%	22.62%
	Top-1 ACC (↑)	91.03%	92.56%	81.28%	77.95%	91.12%	44.90%
	Top-5 ACC (↑)	99.42%	99.33%	98.88%	96.54%	99.71%	78.91%
	Top-10 ACC (↑)	99.94%	99.87%	99.78%	99.07%	99.94%	90.10%
AGAIN	EER (↓)	2.08%	2.15%	2.63%	3.30%	2.11%	5.96%
	Top-1 ACC (↑)	98.33%	97.72%	96.41%	95.61%	97.69%	90.32%
	Top-5 ACC (↑)	100.0%	99.97%	99.78%	99.65%	99.87%	99.90%
	Top-10 ACC (↑)	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
BNE	EER (↓)	4.27%	4.16%	4.87%	6.87%	4.03%	6.47%
	Top-1 ACC (↑)	92.21%	94.46%	90.61%	79.07%	92.89%	76.67%
	Top-5 ACC (↑)	99.52%	99.58%	99.04%	95.90%	99.55%	95.13%
	Top-10 ACC (↑)	99.97%	99.94%	99.81%	98.69%	100.0%	98.08%

[‡] Short for GSM-FR, AMR-NB, 8kHz subsampling, and 4kHz subsampling.
[†] ↑ indicates that the value is the higher, the better, and ↓ indicates that the value is the lower, the better.

Table 6: Performance against the adaptive adversary.

Method / Dataset	VC-specific M3			
	VQVC[2]	VQVC+[2]	AGAIN[2]	BNE[2]
EER (↓)	3.49%	3.55%	2.27%	7.73%
Top-1 ACC (↑)	96.60%	95.64%	99.39%	70.32%
Top-5 ACC (↑)	99.65%	99.78%	100.00%	91.67%
Top-10 ACC (↑)	99.90%	99.97%	100.00%	96.31%
Method / Dataset	Ensembled M3			
	VQVC[2]	VQVC+[2]	AGAIN[2]	BNE[2]
EER (↓)	6.34%	4.95%	2.99%	9.84%
Top-1 ACC (↑)	92.60%	95.99%	99.42%	61.22%
Top-5 ACC (↑)	98.81%	99.55%	100.0%	88.88%
Top-10 ACC (↑)	99.58%	99.94%	100.0%	96.25%

[2] denotes an adaptive adversary who feeds a raw audio to a specific VC twice with two different target speakers.

We visualize the cumulative distribution functions (CDF) of the similarity scores between the restored voiceprint and that of the real dodger and other suspects in Figure 7 (REVELIO) and Figure 11 in the Appendix of the extended version [12] (baseline B1). The scores are calculated from 31,200 pairs of audios from the same speaker (the VC audio whose source speaker is the same as the non-VC audio) and 31,200 pairs of audios from different speakers (the VC audio whose source speaker is different from the non-VC audio). It shows that for REVELIO, the similarity scores between the restored voiceprint and that of the dodger are much higher than those between the restored voiceprint and other suspects, making the two CDFs easily distinguishable with a preset threshold. However, for the baseline voiceprint extractor, the distributions of the real dodger and other suspects are too close to be separated, justifying the effectiveness of REVELIO. From Figure 7, we can see that the distribution gaps of VQVC, VQVC+, and AGAIN are more distinct than those of BNE, meaning that the VC audios produced by BNE are harder to recover. The potential reason is that BNE has the best voice conversion quality, which indicates that BNE removes the phonetic features of the dodger *more thoroughly*, thus the distributions of BNE are less distinguishable than the other three VCs. It indicates that REVELIO may be leveraged as a way to evaluate the performance of voice conversion techniques.

As for the speaker identification task, as shown in Table 1, we can see that REVELIO achieves an average Top-1 ACC of 94.18% (M1), 95.36% (M2), and 96.08% (M3) on the four VCs. REVELIO achieves an average Top-5 ACC of 99.11% (M1), 99.34% (M2), and 99.47% (M3) on the four VCs respectively. We also visualize the original voiceprints of 10 speakers (i.e., the dodgers) and the corresponding extracted voiceprints by the baseline model in Figure 8, and the recovered voiceprints by M2 and M3 in Figure 9 using t-SNE [59]. We can see that the original voiceprint and the recovered voiceprint of the same speaker are tightly clustered. The results imply that the detector can utilize REVELIO to narrow

materializations M1∼M3 on each of these four training sets respectively. Then we test our models and the baseline on the corresponding test sets, i.e., *VQ-Test, V+-Test, AG-Test, BN-Test*. All test audios and evidence audios are trimmed or padded to 20 seconds unless state otherwise. It takes less than 0.1 seconds for REVELIO to recover one voiceprint. Note that the speakers in the test set and the training set have no overlap.

As shown in Table 1 (the best results highlighted), REVELIO achieves an average EER of 4.31% (M1), 3.40% (M2) and 3.27% (M3) on the four VCs. Note that the average EER of the baseline B1 is as high as 28.22%. Compared with B1, the results show that REVELIO can effectively recover voiceprints of the dodger from VC audios, while the plain voiceprint extractor cannot. The reason is that plain voiceprint extractors tend to extract features that are related to the target speakers introduced by VC techniques. In comparison, REVELIO suppresses the influence of the target speaker. As shown in Table 1, B2 and B3 fail to recover voiceprint from VC audios, since they are restricted to traditional frequency-domain voice transformation but not learning-based voice transformation. Note that, in Table 1, we use 5-second test audios and evidence audios for VQVC, VQVC+ and AGAIN to show the differences among three materializations of REVELIO. Further evaluation of audio length is presented in §5.5.

down the suspects and even pinpoint the dodger directly with a high probability. Note that we do not present the results of M1 in Figure 9 because M1 and M2 (or M3) are different models and they map the same speaker into different voiceprints.

The results in Table 1 also verify the effectiveness of our proposed differential rectification block, since M2 and M3 outperform M1 in most cases. In addition, the performances of M3 are better than M2 in most cases, which implies that the model indeed leverages the extra information of the target speaker to improve the voiceprint restoration ability.

5.3 Generalizability to Unseen VCs

In this part, we evaluate the generalizability of REVELIO to VC methods that are unseen in the training phase. We utilize M3 trained on one VC to identify the source speaker of VC audios generated by the other three VCs respectively. We present the results in Table 2. The models trained on VQVC, VQVC+ and AGAIN can generalize well to each other, e.g., the model trained on VQVC achieves an EER of 7.96% (resp. 6.24%) and a Top-1 ACC of 77.31% (resp. 90.71%) on VC audios generated by VQVC+ (resp. AGAIN). A possible reason is that these three VC methods share a similar architecture for disentanglement. Due to significant differences in architecture, these three models can not generalize well to BNE, vice-versa. But the model trained on BNE can still narrow down the dodger on VC audios converted by VQVC, VQVC+ and AGAIN with a Top-10 ACC of 71.19%, 79.30%, and 93.33% respectively.

5.4 Intra- & Inter-gender Performance

In this part, we evaluate the voiceprint restoration performance of M3 on different types of dodger-target pairs, i.e., male-to-male (M→M), female-to-female (F→F), female-to-male (F→M), and male-to-female (M→F). The four test sets are split from *BN-Test*. As shown in Table 11, we can see that the performances of intra- and inter-gender cases are similar, with an average EER of 2.69% (M→M), 3.72% (F→F), 3.90% (F→M), and 3.73% (M→F) respectively. The same is true for the speaker identification performance. Note that we still perform speaker identification among all 40 speakers. The results imply that REVELIO can handle both intra- and inter-gender conversion even if the intra- and inter-gender conversion may induce different intensities of distortions on the dodger voiceprint.

5.5 Audio Length

In this part, we first examine how long the VC audio needs to be to recover the dodger voiceprint. We trim BNE-processed VC audios to 5s, 10s, 15s, 20s, and 25s. As shown in Table 3, we find that a 10-second VC audio is sufficient for voiceprint restoration, achieving an EER of 3.71% and a Top-1 ACC

of 92.31%. This requirement is practical in the real world since a phone scam can hardly be completed in less than 10 seconds. In addition, the results imply that a longer VC audio is beneficial to voiceprint restoration. Note that in Table 3, the performance decreases at 20s and 25s because some of the VC audios in the dataset have a shorter length and are padded to 20s and 25s with zeros, which affects the restoration.

Moreover, we examine how long the evidence audio needs to be to help recover the dodger voiceprint. We trim evidence audios to 5s, 10s, 15s, 20s, 25s and see the performances of voiceprint restoration. As shown in Table 4, we find that a 5-second evidence audio is sufficient for improving the voiceprint restoration performance, yielding a Top-1 ACC of 95.61%. In comparison, with no evidence audio, Table 1 shows that REVELIO achieves a Top-1 ACC of 95.10%. This implies that the evidence audio can help recover the dodger voiceprint.

5.6 Unseen Language

In this part, we evaluate whether REVELIO can generalize to VC audios in other languages that are unseen in the training phase. We convert German, French, and Spanish test sets of multilingual LibriSpeech using BNE and evaluate the performance of M3. As shown in Table 8 in the Appendix, REVELIO achieves an EER of 6.08%, 5.78%, 3.68% and a Top-1 ACC of 80.40%, 64.71%, 92.11% on German, French, and Spanish test sets respectively. The results show that REVELIO trained on English training sets can generalize to German, French, and Spanish audios, especially Spanish audios, on which REVELIO achieves comparable performances as on English audios. This language generalization capability is useful in multilingual scenarios. For example, the VC audio targets a Spanish-speaking victim living in English-speaking countries. In this case, the local detector with a REVELIO model trained with English corpus can also recover the voiceprint of the dodger in the Spanish-speaking VC audio.

5.7 Performance over Telephony

In this part, we evaluate whether REVELIO can recover the voiceprint of the dodger if the VC audio is encoded by telephony codecs, especially those used in a public switched telephone network (PSTN) or a voice over Internet protocol (VoIP) network. We apply four commonly-used codecs and 8k/4kHz subsampling on the VC audios to simulate the coding and transmission telephony channel.

- G.711 (μ-law) and G.711 (A-law) [25] are companding algorithms primarily used in 8-bit PCM digital telecommunication systems to optimize the dynamic range of an analog signal for digitization. μ-law is commonly used in North America and Japan, while A-law is used in Europe.

- GSM Full Rate (GSM-FR) [16] is the first digital speech coding standard in the GSM digital mobile phone system. The speech encoder accepts 13-bit linear PCM at an 8 kHz sampling rate.

- Adaptive Multi-Rate (AMR-NB or GSM-AMR) [1] consists of a multi-rate narrowband speech codec that encodes narrowband (200–3400 Hz) signals. AMR-NB is widely used in GSM and UMTS.

Audios converted by VQVC, VQVC+, AGAIN, and BNE are encoded using these four codecs or subsampled to 8k/4kHz, and we evaluate whether the corresponding M3 models can still perform voiceprint restoration. As shown in Table 5, REVELIO achieves an average EER of 4.11%, 4.03%, 4.96%, 6.85%, 3.95%, and 12.71% on recovering audios distorted by μ-law, A-law, GSM-FR, AMR-NB and 8k/4kHz subsampling. REVELIO almost maintains its original effectiveness against μ-law, A-law and 8kHz subsampling as compared to recovering non-codec/non-subsampled audios. REVELIO still maintains acceptable performances under GSM-FR (3.27%→4.96%) and AMR-NB (3.27%→6.85%), which might result from the data augmentation techniques we use in the training phase. The performance degradation on VQVC and VQVC+ audios subsampled to 4kHz is due to the significant distortion brought by subsampling. Even so, REVELIO can still narrow down the dodger with a Top-5 ACC of 82.92%, 78.91%, 99.90% and 95.13% from subsampled VQVC, VQVC+, AGAIN and BNE audios respectively.

5.8 Model Ensembling

In this part, we evaluate whether one model can learn to restore voiceprints processed by different VCs. We sample one-fifth of the audios from the four VC training sets and the raw training sets respectively and make a multi-VC training dataset. Then we train REVELIO on the multi-VC dataset. As shown in Table 9 in the Appendix, REVELIO achieves an EER of 2.94%, 2.22%, 1.65%, 4.66% and a Top-1 ACC of 99.84%, 99.94%, 99.97% and 90.22% on the test sets processed by VQVC, VQVC+, AGAIN, and BNE respectively. We visualize the distributions of the similarity scores between the restored voiceprint and that of the real dodger and other suspects in Figure 12 in the extended version [12]. We can see that the distributions are similar to those shown in Figure 7, presenting an obvious gap between the real dodger and other suspects. It shows that the ensemble model can learn to recover voiceprint from four different VCs with promising performances. The detector can train an ensemble REVELIO model on a multi-VC dataset generated with all popular VC techniques, thus being able to recover the voiceprint of most VC audios. We also evaluate the over-the-telephony robustness of the ensemble model, as shown in Table 12 in the extended version [12]. We can see that the ensemble model still maintains acceptable performances in most cases.

5.9 Adaptive Adversary

In this part, we consider an adaptive adversary who feeds a raw audio to VC twice with two different target speakers, e.g., Alice first converts the raw audio to sound like Bob and then converts the converted audio to sound like Eve, both by the same VC. In such a case, Alice is the source speaker to be identified and Eve is the final target speaker of the evidence audio. As shown in Table 6, REVELIO still achieves very low EER and very high ACC on the audios that are converted twice by VQVC, VQVC+, and AGAIN. The performance on BNE audios is not as good as others. The possible reason is that BNE's disentanglement module is better at erasing the voice characteristics of the source speaker in the process of two conversions. For BNE, REVELIO can still achieve an EER of 7.73%, a Top-1 ACC of 70.32% and a Top-5 ACC of 91.96%.

6 Related Work

6.1 Voice Conversion (VC)

VC techniques can be categorized into parallel VC and non-parallel VC according to whether they require parallel corpus for training.

Parallel VC. Parallel VC leverages parallel corpus for training the VC model. Abe et al. [2] created code vectors based on hard clustering using vector quantization (VQ) for feature mapping. Several subsequent works [3, 48, 58] tried to reduce the quantization error of VQ by using a fuzzy VQ based on soft clustering and the output is weighed according to the source speaker features. Wu et al. [66] output a linear combination of the exemplars with weights determined by the source speaker features. Stylianou et al. [51] utilized a continuous probabilistic transformation method based on Gaussian mixture models (GMMs). Toda et al. [56] alleviated the over-smoothing effect of the GMM-based method with global variance. Nakashika et al. [38] modeled the source and the target speaker with deep belief nets (DBNs) and converted the audio using a neural network (NN). Sun et al. [52] used a bidirectional long short-term memory-based recurrent neural network (BiLSTM) for conversion.

Non-Parallel VC. To leverage non-parallel corpora, several attempts [4, 15, 50, 72] have been made to adapt parallel VC to use non-parallel training sets, but mismatches in alignment may occur. To avoid explicit alignment, Hsu et al. [22] used a variational auto-encoder (VAE) for speech feature transformation. Later works proposed different models for transformation [17, 23, 27, 28, 39, 47, 55], but fail to handle target speakers that are unseen in the training process. To tackle this shortcoming, advanced disentanglement-based methods have been proposed. Chou et al. [10] disentangled speaker and content features with an average pooling layer and the instance normalization technique (IN), and synthe-

sized them with the adaptive instance normalization technique (AdaIN). Wu et al. [64] proposed VQVC, utilizing VQ and IN for disentanglement. Later, they combined VQ with U-Net to produce high-quality audios [63]. Chen et al. [9] developed AGAIN that uses activation as an information bottleneck to further separate speaker features from content features. Liu et al. [34] extracted content features with an encoder-decoder-based hybrid connectionist-temporal-classification-attention (CTC-attention) phoneme recognizer. Disentanglement-based methods have become a trend of real-time VC because non-parallel and one-shot conversion can be achieved.

6.2 Voice Conversion Detection

Speech deepfake detection is one of the aims of the ASVspoof 2021 Challenge [67] and is the main topic of the 2022 Audio Deep Synthesis Detection (ADD) Challenge [69]. VC is a large part of deepfake techniques. Many efforts have been devoted to VC detection, e.g., spectro-temporal graph attention network (GAT) [53], differentiable architecture [20], a cascade of an embedding extractor [6], continual learning [35], and multi-task learning [71]. Several attempts have been made into dataset generation [68] and data augmentation [7, 8, 11, 57] for model generalization. Nonetheless, existing works only aim at detecting whether the audio is genuine or fake, while we take a step further to restore the voiceprint of the source speaker of a VC audio.

7 Responsible Disclosure

There are both legitimate and illegitimate reasons to use VCs. The rest of this paper discusses illegitimate; legitimate reasons include voice anonymization for privacy concern, voice dubbing for one who is unavailable, aids for the speech-impaired, etc. Any system that can determine the identity of a person behind a VC has the potential to harm those who are using it for legitimate purposes and has the ability to find those who are using it for illegitimate purposes. Having weighed the benefits and harms, we are releasing REVELIO in a limited way, i.e., we provide the code of REVELIO upon request to, e.g., law enforcement or professors at other institutions who are doing related research. This helps with providing the code to those who can use it for legitimate purposes, while reducing the potential harms from releasing it publicly. In addition, we have contacted companies that provide voice conversion services, reported the issues that might be brought by REVELIO, and recommended that they inform users about the privacy concerns.

8 Discussion

In this section, we discuss the limitations of REVELIO.

Adaptive voice conversion. Being aware of REVELIO, a dodger may adapt the voice conversion process to compromise the voiceprint restoration effort of REVELIO, e.g., converting to a target speaker twice with two different VCs. Our preliminary experiments show that the recovery results of VC combinations, e.g., AGAIN(BNE($\mathbf{x}_0$)), are not ideal but still achieve an EER of 17.9% and a Top-1 ACC of 57.9%. The possible reason is that a combination of VCs can be seen as a new kind of VC, to which REVELIO fails to generalize. A possible solution is to build a generative adversarial network (GAN) to simulate the game between voice disguise and voiceprint recovery. We consider this an interesting future direction.

Inverse voice conversion. With the recovered voiceprint of the dodger, it is possible to reconstruct the raw audio before voice conversion. Therefore, we may perform an end-to-end recovery with the converted audio as the input and the recovered raw audio as the output. Nonetheless, existing VCs require a raw audio as input and extract the speaker embeddings internally. It is difficult to align the speaker embedding recovered by REVELIO with the one used in VC as we only have black-box access to the VC model. In the future, we will explore this direction.

VC quality. In §5.2, we have found that the VC quality impacts the performance of voiceprint recovery. A better VC (like BNE) performs a more thorough conversion to the target speaker, thus preserves less information of the source speaker. Nonetheless, there is no existing VC technique that can perfectly disentangle content and phonetic features, which leaves space for voiceprint restoration.

9 Conclusion

In this paper, we present the rationale, design and evaluation of REVELIO, a voiceprint restoration model that attempts to retrieve the voiceprint of the source speaker from audios processed by voice conversion techniques. We have established a representation learning model that can effectively learn the voiceprint embedding of the source speaker given the input VC audios. With the target speaker evidence audio, we have designed a novel differential rectification to further refine the extracted voiceprint. Our experiments have confirmed the effectiveness of REVELIO in rebuilding voiceprint under various voice conversion techniques.

Acknowledgments

References

[1] 3GPP. Mandatory Speech Codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec; Transcoding functions. Technical specification 3G TS 26.090 version 3.0.1, The 3rd Generation Partnership Project, 1999.

[2] Masanobu Abe, Satoshi Nakamura, Kiyohiro Shikano, and Hisao Kuwabara. Voice conversion through vector quantization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1988.

[3] Levent M. Arslan and David Talkin. Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum. In *Conference on Speech Communication and Technology*. ISCA, 1997.

[4] Hadas Benisty, David Malah, and Koby Crammer. Nonparallel voice conversion using joint optimization of alignment by temporal context and spectral distortion. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014.

[5] Thomas Brewster. Fraudsters cloned company director's voice in $35 million bank heist, police find. https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions.

[6] Tianxiang Chen, Elie Khoury, Kedar Phatak, and Ganesh Sivaraman. Pindrop Labs' submission to the ASVspoof 2021 Challenge. In *Automatic Speaker Verification and Spoofing Countermeasures Challenge*. ISCA, 2021.

[7] Tianxiang Chen, Avrosh Kumar, Parav Nagarsheth, Ganesh Sivaraman, and Elie Khoury. Generalization of audio deepfake detection. In *The Speaker and Language Recognition Workshop*. ISCA, 2020.

[8] Xinhui Chen, You Zhang, Ge Zhu, and Zhiyao Duan. UR channel-robust synthetic speech detection system for ASVspoof 2021. In *Automatic Speaker Verification and Spoofing Countermeasures Challenge*. ISCA, 2021.

[9] Yen-Hao Chen, Da-Yi Wu, Tsung-Han Wu, and Hung-yi Lee. AGAIN-VC: A one-shot voice conversion using activation guidance and adaptive instance normalization. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021.

[10] Ju-Chieh Chou and Hung-yi Lee. One-shot voice conversion by separating speaker and content representations with instance normalization. In *Conference of the International Speech Communication Association*. ISCA, 2019.

[11] Rohan Kumar Das. Known-unknown data augmentation strategies for detection of logical access, physical access and speech deepfake attacks: ASVspoof 2021. In *Automatic Speaker Verification and Spoofing Countermeasures Challenge*. ISCA, 2021.

[12] Jiangyi Deng, Yanjiao Chen, Yinan Zhong, Qianhao Miao, Xueluan Gong, and Wenyuan Xu. Catch You and I Can: Revealing source voiceprint against voice conversion. *arXiv preprint arXiv:2302.12434*, 2023.

[13] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[14] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Conference of the International Speech Communication Association*. ISCA, 2020.

[15] Daniel Erro, Asunción Moreno, and Antonio Bonafonte. INCA algorithm for training voice conversion systems from nonparallel corpora. *IEEE Transactions on Speech and Audio Processing*, 18(5):944–953, 2010.

[16] ETSI. Digital cellular telecommunications system (Phase 2+) (GSM); Full rate speech; Transcoding (GSM 06.10 version 8.1.1 Release 1999). Standard ETSI EN 300 961, European Telecommunications Standards Institute, 2000.

[17] Fuming Fang, Junichi Yamagishi, Isao Echizen, and Jaime Lorenzo-Trueba. High-quality nonparallel voice conversion based on cycle-consistent adversarial network. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[18] R. A. FISHER. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[19] Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip H. S. Torr. Res2Net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652–662, 2021.

[20] Wanying Ge, Jose Patino, Massimiliano Todisco, and Nicholas Evans. Raw differentiable architecture search for speech deepfake and spoofing detection. In *Automatic Speaker Verification and Spoofing Countermeasures Challenge*. ISCA, 2021.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[22] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang. Voice conversion from non-parallel corpora using variational auto-encoder. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2016.

[23] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang. Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks. In *Conference of the International Speech Communication Association*. ISCA, 2017.

[24] Chien-yu Huang, Yist Y Lin, Hung-yi Lee, and Lin-shan Lee. Defending your voice: Adversarial attack on voice conversion. In *IEEE Spoken Language Technology Workshop*, 2021.

[25] ITU-T. Pulse code modulation (PCM) of voice frequencies. Recomandation T-REC-G.711-198811, International Telecommunication Union, 1988.

[26] Jaison Joshy and Koj Sambyo. A comparison and contrast of the various feature extraction techniques in speaker recognition. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 9(11):99–108, 2016.

[27] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks. In *IEEE Spoken Language Technology Workshop*, 2018.

[28] Takuhiro Kaneko and Hirokazu Kameoka. CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks. In *European Signal Processing Conference*. IEEE, 2018.

[29] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*. OpenReview.net, 2015.

[30] Sanjeev Kumar, D Kishore, and SV Kiran. Acoustic features extraction of non-electronic disguised voice for speaker identification. *Joint Journal of Novel Carbon Resource Sciences & Green Asia Strategy*, 9(3):853–860, 2022.

[31] Phillip L. De Leon, Michael Pucher, Junichi Yamagishi, Inma Hernáez, and Ibon Saratxaga. Evaluation of speaker verification security and detection of hmm-based synthetic speech. *IEEE Transactions on Speech and Audio Processing*, 20(8):2280–2290, 2012.

[32] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu. Deep Speaker: An end-to-end neural speaker embedding system. *arXiv preprint arXiv:1705.02304*, 2017.

[33] Derek Lin et al. *Tao Te Ching: Annotated & Explained.* SkyLight Paths Publishing, 2006.

[34] Songxiang Liu, Yuewen Cao, Disong Wang, Xixin Wu, Xunying Liu, and Helen Meng. Any-to-many voice conversion with location-relative sequence-to-sequence modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1717–1728, 2021.

[35] Haoxin Ma, Jiangyan Yi, Jianhua Tao, Ye Bai, Zhengkun Tian, and Chenglong Wang. Continual learning for fake audio detection. In *Conference of the International Speech Communication Association*. ISCA, 2021.

[36] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. VoxCeleb: A large-scale speaker identification dataset. In *Conference of the International Speech Communication Association*. ISCA, 2017.

[37] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. VoxCeleb2: Deep speaker recognition. In *Conference of the International Speech Communication Association*. ISCA, 2018.

[38] Toru Nakashika, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. Voice conversion in high-order eigen space using deep belief nets. In *Conference of the International Speech Communication Association*. ISCA, 2013.

[39] Toru Nakashika, Tetsuya Takiguchi, and Yasuhiro Minami. Non-parallel training in voice conversion using an adaptive restricted boltzmann machine. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2032–2045, 2016.

[40] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Conference on Neural Information Processing Systems*. PMLR, 2019.

[42] Patrick Perrot, Guido Aversano, and Gérard Chollet. Voice disguise and automatic detection: Review and perspectives. In *Progress in nonlinear speech processing,*

Workshop on Nonlinear Speech Processing. Springer, 2005.

[43] 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 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[44] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. MLS: A large-scale multilingual dataset for speech research. In *Conference of the International Speech Communication Association*. ISCA, 2020.

[45] 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. *arXiv preprint arXiv:2106.04624*, 2021.

[46] Joanne K Rowling and Gerhard Lauer. *Harry Potter*. Bloomsbury London, 2008.

[47] Yuki Saito, Yusuke Ijima, Kyosuke Nishida, and Shinnosuke Takamichi. Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors. In *IEEE International Conference on Acoustics, Speech and Signal*, 2018.

[48] Kiyohiro Shikano, Satoshi Nakamura, and Masato Abe. Speaker adaptation and voice conversion by codebook mapping. In *IEEE International Symposium on Circuits and Systems*, 1991.

[49] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-Vectors: Robust DNN embeddings for speaker recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[50] Peng Song, Wenming Zheng, and Li Zhao. Non-parallel training for voice conversion based on adaptation method. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[51] Yannis Stylianou, Olivier Cappé, and Eric Moulines. Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing*, 6(2):131–142, 1998.

[52] Lifa Sun, Shiyin Kang, Kun Li, and Helen M. Meng. Voice conversion using deep bidirectional long short-term memory based recurrent neural networks. In *IEEE*

International Conference on Acoustics, Speech and Signal Processing, 2015.

[53] Hemlata Tak, Jee weon Jung, Jose Patino, Madhu Kamble, Massimiliano Todisco, and Nicholas Evans. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. In *Automatic Speaker Verification and Spoofing Countermeasures Challenge*. ISCA, 2021.

[54] Huaizhen Tang, Xulong Zhang, Jianzong Wang, Ning Cheng, and Jing Xiao. AVQVC: One-shot voice conversion by vector quantization with applying contrastive learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

[55] Patrick Lumban Tobing, Yi-Chiao Wu, Tomoki Hayashi, Kazuhiro Kobayashi, and Tomoki Toda. Non-parallel voice conversion with cyclic variational autoencoder. In *Conference of the International Speech Communication Association*. ISCA, 2019.

[56] Tomoki Toda, Alan W. Black, and Keiichi Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Speech and Audio Processing*, 15(8):2222–2235, 2007.

[57] Anton Tomilov, Aleksei Svishchev, Marina Volkova, Artem Chirkovskiy, Alexander Kondratev, and Galina Lavrentyeva. STC antispoofing systems for the ASVspoof 2021 Challenge. In *Automatic Speaker Verification and Spoofing Countermeasures Challenge*. ISCA, 2021.

[58] Oytun Türk and Levent M. Arslan. Robust processing techniques for voice conversion. *Computer Speech & Language*, 20(4):441–467, 2006.

[59] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[60] Christophe Veaux, Junichi Yamagishi, and Simon King. Towards personalised synthesised voices for individuals with vocal disabilities: Voice banking and reconstruction. In *Workshop on Speech and Language Processing for Assistive Technologies*. Association for Computational Linguistics, 2013.

[61] Run Wang, Felix Juefei-Xu, Yihao Huang, Qing Guo, Xiaofei Xie, Lei Ma, and Yang Liu. DeepSonar: Towards effective and robust detection of ai-synthesized fake voices. In *ACM International Conference on Multimedia*, 2020.

[62] Yong Wang, Haojun Wu, and Jiwu Huang. Verification of hidden speaker behind transformation disguised voices. *Digital Signal Processing*, 45:84–95, 2015.

[63] Da-Yi Wu, Yen-Hao Chen, and Hung-yi Lee. VQVC+: one-shot voice conversion by vector quantization and u-net architecture. In *Conference of the International Speech Communication Association*. ISCA, 2020.

[64] Da-Yi Wu and Hung-Yi Lee. One-shot voice conversion by vector quantization. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020.

[65] Zhizheng Wu and Haizhou Li. Voice conversion versus speaker verification: An overview. *APSIPA Transactions on Signal and Information Processing*, 3:17, 2014.

[66] Zhizheng Wu, Tuomas Virtanen, Tomi Kinnunen, Eng Siong Chng, and Haizhou Li. Exemplar-based voice conversion using non-negative spectrogram deconvolution. In *ISCA Tutorial and Research Workshop on Speech Synthesis*, 2013.

[67] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md. Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas W. D. Evans, and Héctor Delgado. ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection. *arXiv preprint arXiv:2109.00537*, 2021.

[68] Jiangyan Yi, Ye Bai, Jianhua Tao, Haoxin Ma, Zhengkun Tian, Chenglong Wang, Tao Wang, and Ruibo Fu. Half-Truth: A partially fake audio detection dataset. In *Conference of the International Speech Communication Association*. ISCA, 2021.

[69] Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, Shan Liang, Shiming Wang, Shuai Zhang, Xinrui Yan, Le Xu, Zhengqi Wen, and Haizhou Li. ADD 2022: The first audio deep synthesis detection challenge. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

[70] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations*. OpenReview.net, 2016.

[71] Lin Zhang, Xin Wang, Erica Cooper, and Junichi Yamagishi. Multi-task learning in utterance-level and segmental-level spoof detection. In *Automatic Speaker Verification and Spoofing Countermeasures Challenge*. ISCA, 2021.

[72] Meng Zhang, Jianhua Tao, Jilei Tian, and Xia Wang. Text-independent voice conversion based on state mapped codebook. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 4605–4608, 2008.

[73] Mingyang Zhang, Xin Wang, Fuming Fang, Haizhou Li, and Junichi Yamagishi. Joint training framework for text-to-speech and voice conversion using multi-source tacotron and wavenet. *arXiv preprint arXiv:1903.12389*, 2019.

[74] Linlin Zheng, Jiakang Li, Meng Sun, Xiongwei Zhang, and Thomas Fang Zheng. When automatic voice disguise meets automatic speaker verification. *IEEE Transactions on Information Forensics and Security*, 16:824–837, 2021.

A Dataset Generation

As shown in Table 7, we have nine clean audio datasets for voice conversion dataset generation, i.e., *train-clean-100*, *train-clean-360*, *train-other-500*, VoxCeleb1, VoxCeleb2, *test-clean*, MLS German, MLS French, and MLS Spanish.

- *train-clean-100* has 251 speakers. For each source speaker in *train-clean-100*, the remaining 250 speakers are used as the target speaker to generate VC samples.

- For each source speaker in *train-clean-360*, *train-other-500*, and VoxCeleb1, we randomly sample another 300 speakers in the same dataset as the target speaker.

- For each source speaker in VoxCeleb2, we randomly sample another 100 speaker as the target speaker.

We generate one VC sample for each of the above source-target pair for training.

- *test-clean* has 40 speakers. For each source speaker, the remaining 39 speakers in the test set are used to generate test samples. We generate 20 samples for each source-target pair (a total of $40 \times 39 \times 20 = 31,200$ samples as shown in Table 7). The speaker identification system is enrolled with all 40 speakers for evaluations.

- MLS German, MLS French, and MLS Spanish have 30, 18, 20 speakers respectively. For each source speaker, all remaining speakers in the dataset are used to generate test samples. We generate 20 samples for each source-target pair.

Table 7: Voice conversion dataset.

Method	Dataset[†]	Alias	Lang.	#Speaker	#Sample
VQVC	*train-clean-100*			251	62,750
	train-clean-360			921	276,300
	train-other-500	VQ-Train	English	1,166	349,800
	VoxCeleb1			1,251	375,300
	VoxCeleb2			5,994	599,400
	test-clean	VQ-Test	English	40	31,200
VQVC+	*train-clean-100*			251	62,750
	train-clean-360			921	276,300
	train-other-500	V+-Train	English	1,166	349,800
	VoxCeleb1			1,251	375,300
	VoxCeleb2			5,994	599,400
	test-clean	V+-Test	English	40	31,200
AGAIN	*train-clean-100*			251	62,750
	train-clean-360			921	276,300
	train-other-500	AG-Train	English	1,166	349,800
	VoxCeleb1			1,251	375,300
	VoxCeleb2			5,994	599,400
	test-clean	AG-Test	English	40	31,200
BNE	*train-clean-100*			251	62,750
	train-clean-360			921	276,300
	train-other-500	BN-Train	English	1,166	349,800
	VoxCeleb1			1,251	375,300
	VoxCeleb2			5,994	599,400
	test-clean	BN-Test	English	40	31,200
BNE	MLS German	BN-GE	German	30	17,400
	MLS French	BN-FR	French	18	6,120
	MLS Spanish	BN-SP	Spanish	20	7,600

[†] *train-clean-100*, *train-clean-360* and *train-other-500* are training sets in LibriSpeech [40]. We use the test sets in German, French, and Spanish of Multilingual LibriSpeech (MLS) [44] for evaluation.

Table 8: Generalization to unseen languages.

Metrics[†]	English	German	French	Spanish
EER ($\downarrow$)	3.78%	6.08%	5.78%	3.68%
Top-1 ACC ($\uparrow$)	96.12%	80.40%	64.71%	92.11%
Top-5 ACC ($\uparrow$)	99.97%	97.30%	97.22%	100.0%
Top-10 ACC ($\uparrow$)	100.0%	99.14%	99.67%	100.0%

[†] There are 40, 30, 18, and 20 speakers in *BN-Test*, *BN-GE*, *BN-FR*, and *BN-SP* respectively. $\uparrow$ indicates that the value is the higher, the better, and $\downarrow$ indicates that the value is the lower, the better.

Table 9: Trained on the multi-VC dataset.

Metrics[†]	VQVC	VQVC+	AGAIN	BNE
EER ($\downarrow$)	2.94%	2.22%	1.65%	4.66%
Top-1 ACC ($\uparrow$)	99.84%	99.94%	99.97%	90.22%
Top-5 ACC ($\uparrow$)	100.0%	100.0%	100.0%	99.55%
Top-10 ACC ($\uparrow$)	100.0%	100.0%	100.0%	100.0%

[†] $\uparrow$ indicates that the value is the higher, the better, and $\downarrow$ indicates that the value is the lower, the better.

Table 10: The detailed implementation of REVELIO.

Module	Block	Output size	#Params
Feature extraction	Filter bank	$(80, T)$	-
	Conv1D+ReLU+BN	$(1024, T)$	412,672
	SE-Res2Block	$(1024, T)$	2,713,344
	SE-Res2Block	$(1024, T)$	2,713,344
	SE-Res2Block	$(1024, T)$	2,713,344
	Block summary	$\mathbf{M}: (3072, T_M)$ $\mathbf{N}: (3072, T_N)$	8,552,704
Differential rectification	Res-Orth. Block	$(3072, T_M)$	9,446,400
Dimension normalization	Conv1D+ReLU+BN	$(3072, T_M)$	9,446,400
	ASP+BN	$(6144, T_M)$	1,588,608
	FC+BN	$(192, 1)$	1,179,840
Voiceprint enhancement	AAM-Softmax	$(9583, 1)$	1,839,936
Model summary	**Input**: $(1, t_m) \times (1, t_n) \rightarrow$ **Output**: $(9583, 1)$ **Total params**: 32,053,888		

(i) $\mathbf{M}$ and $\mathbf{N}$ are defined in §4.2. (ii) We allow the input VC audios and evidence audios to have arbitrary lengths, so t_m and t_n do not have to be the same.

Table 11: Performance of voiceprint restoration on intra- and inter-gender VC.

Method	Metrics[†]	M→M	F→F	F→M	M→F
VQVC	EER ($\downarrow$)	3.07%	4.55%	5.13%	3.29%
	Top-1 ACC ($\uparrow$)	94.87%	92.37%	90.88%	93.38%
	Top-5 ACC ($\uparrow$)	99.08%	97.76%	98.00%	98.75%
	Top-10 ACC ($\uparrow$)	99.61%	98.95%	99.00%	99.50%
VQVC+	EER ($\downarrow$)	3.25%	3.83%	3.93%	3.25%
	Top-1 ACC ($\uparrow$)	95.38%	93.82%	93.63%	95.38%
	Top-5 ACC ($\uparrow$)	99.25%	98.95%	99.13%	99.25%
	Top-10 ACC ($\uparrow$)	99.88%	99.47%	99.75%	99.88%
AGAIN	EER ($\downarrow$)	1.66%	2.07%	2.15%	1.78%
	Top-1 ACC ($\uparrow$)	99.61%	98.42%	99.25%	99.50%
	Top-5 ACC ($\uparrow$)	99.87%	100.0%	99.75%	99.75%
	Top-10 ACC ($\uparrow$)	100.0%	100.0%	100.0%	100.0%
BNE	EER ($\downarrow$)	2.78%	4.43%	4.39%	2.88%
	Top-1 ACC ($\uparrow$)	95.26%	95.53%	94.00%	93.00%
	Top-5 ACC ($\uparrow$)	99.87%	99.74%	99.50%	99.75%
	Top-10 ACC ($\uparrow$)	100.0%	100.0%	100.0%	100.0%

[†] $\uparrow$ indicates that the value is the higher, the better, and $\downarrow$ indicates that the value is the lower, the better.