IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 2, 2020, accepted October 21, 2020, date of publication October 27, 2020, date of current version November 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3034253

One-Shot Voice Conversion Algorithm Based on Representations Separation

CHUNHUI DENG[1], YING CHEN[2], AND HUIFANG DENG[2]
[1]School of Computer Engineering, Guangzhou College, South China University of Technology, Guangzhou 510641, China
[2]School of Computer Science and Engineering, South China University of Technology, Guangzhou 510641, China

Corresponding author: Huifang Deng (hdengpp@qq.com)

This work was supported in part by the Department of Education of Guangdong Province through Special Innovation Program (Natural Science), under Grant 2015KTSCX183, and in part by the South China University of Technology through ''Development Fund'' under Grant x2js-F8150310.

ABSTRACT Voice Conversion (VC) is a method of converting the source speaker's speech into the target speaker's speech without changing the source speaker's speech content. The current VC methods have the following problems: (1) they are only applicable to a limited number of speakers, not to any speakers, as a result, the application scenarios are greatly restricted; (2) the representation (feature) separation(RS) effect of the current mainstream technology is not ideal on the source speaker speech and the target speaker speech; and (3) the voice conversion quality of most models is unsatisfactory, and hence needs to be improved. Therefore, in this paper, we constructed a one-shot VC model of Representation Separation, called RS-VC model, implemented by the encoder-decoder structure. The encoder is composed of a content encoder and a speaker encoder. The content encoder separates the content information of the source speaker voice and generates a content representation. The speaker encoder separates the target speaker information of the target speaker voice and generates a speaker representation. The decoder synthesizes the content representation and the speaker representation to generate the converted voice. In this paper, we obtained the optimized speaker verification model SVIGEN2E (Speaker Verification with Instance Normalization using Generalized End-to-End loss) by improving the speaker verification (SV) model. The model SVIGEN2E is used as the speaker encoder. This speaker encoder needs to be trained in advance prior to RS-VC model training, and the pre-trained model of SVINGE2E directly extracts speaker representation of the target speaker's voice, and is used for training and testing RS-VC model. A progressive training method is proposed then for training RS-VC model. Experiments show that the progressive training method can effectively improve the quality of the converted voice. Compared with the basic speaker verification model, both SVINGE2E and RS-VC deliver the impressive improvements in EER (Equal Error Rate).

INDEX TERMS Voice conversion, content representation, speaker representation, representation (feature) separation, speaker verification, one-shot, speaker encoder, content encoder, progressive training method.

I. INTRODUCTION

Although Voice Conversion (VC) is a research branch of speech synthesis, the research history is also very long. It is a method of converting source speaker's speech into target speaker's speech without changing the source speaker's speech content. With the development of technology, voice conversion technology has also undergone some changes. Early VC methods are mainly involved in designing speech feature extraction methods, extracting personal features in speech, and then constructing a representation mapping model to train features through parallel voice data sets, one can get the trained feature mapping parameters, and complete the construction of the VC model and the collection of parameters. When performing VC, two steps are followed: first extract the target speech features and source speech features and then use the feature conversion model to perform VC to obtain the converted speech. There are mainly two types of VCs: the channel spectrum based and the prosodic conversion based. The VC based on channel spectrum is mainly divided into four categories: (1) Codebook mapping-based methods [1]–[4], (2) Gaussian mixing model methods [5]–[9], (3) Hidden Markov model-based methods [10], (4) Neural network-based conversion methods [11], [12]. At present,

The associate editor coordinating the review of this manuscript and approving it for publication was Ioannis Schizas.

the main VC methods are based on the neural network due to rapid development of deep learning and neural network methods. In 2016, Oord *et al.* proposed a neural network-based vocoder model WaveNet [11]. That is a vocoder. It plays an important role in the field of voice synthesis and VC, and can synthesize voice features into original voice with better sound quality. Later, Liu *et al.* [13] explored WaveNet vocoder with limited training data for VC In 2017 Hsu *et al.* [14] proposed a Variational Autoencoding Wasserstein Generative Adversarial Network(VAW-GAN) non-parallel VC framework; Kameoka *et al.* [15] proposed a non-parallel data VC method and Kaneko *et al.* [16] explored parallel-data-free voice conversion using cycle-consistent adversarial networks. It is called cycle-consistent generative adversarial network (Cycle-GAN), does not need to align the data and the model, and can alleviate the excessive smoothness of the generated results to a certain extent. In 2018, Chou *et al.* [17] proposed an adversarial learning framework, in which the Cycle-GAN is used to separate voice speaker's representation in the voice signal from the voice content, and use of the model trained by this framework can achieve VC between multiple speakers. In the same year, Kameoka *et al.* [15] proposed a variant of generative adversarial network named Star-GAN to perform non-parallel many-to-many VC, and experiments show that this variant has higher voice similarity and sound quality than the general Variational Autoencoding Generative Adversarial Network (VAE-GAN). In 2019, Qian *et al.* [18] proposed anew non-parallel many-to-many VC method. This method uses an automatic encoder and decoder and realizes the transfer of distributed matching patterns by training the self-reconstruction loss. Specifically, this method uses the automatic encoder to get the voice content information and uses the decoder to synthesize the content information with the target person information to generate new voice. The VC methods proposed in [14], [16] and [17] build the voice conversion models by improvements to the GAN, but these methods can only conduct the conversion between limited speakers, and the quality of the converted voice needs to be improved. Reference [15] mainly uses an Encoder-Decoder frame to realize the VC of any speaker, but the effect of the encoder representation (feature) separation (RS) in this variant is not obvious and not ideal. Aiming at the above inadequacies, in this paper, we built a one-shot VC model based on representation separation.

At present, deep learning technology as a research hotspot has also made some progress in the field of voice conversion. Therefore, the current technical research of voice conversion is mainly based on deep learning neural networks. The current VC methods have the following problems: (1) they are only applicable to a limited number of speakers, not to any speakers, as a result, the application scenarios are greatly restricted; (2) the feature separation effect of the current mainstream technology is not satisfactory in representation separation (RS) on the source speaker speech and the target speaker speech; (3) the voice conversion quality of most models is still unsatisfactory, and hence needs to be improved. In response to the above problems, in this paper, a new VC model is built based on the Encoder-Decoder structure of representation (feature) separation and deep learning and neural network. The Encoder consists of two parts: The Speaker Encoder and the Content Encoder. The Speaker Encoder separates the target speaker information from the target speaker speech to generate Speaker Representations or Speaker Features; the content encoder separates the content information from the source speaker speech to generate Content Representations or Speaker Features, and the Decoder synthesizes speaker representations and the content representation to generate the target speaker's speech with the source speaker's speech content and target speaker information. In this paper, the proposed VC model only needs to input any source speaker speech and target speaker speech to achieve the voice conversion between any two speakers, also known as one-shot VC. In order to improve the representation separation effect and extract more representative speaker representations, this paper optimizes the basic speaker verification (SV) model and obtains the optimized SV model called SVINGE2E(Speaker Verification with Instance Normalization using Generalized End-to-End loss), which achieved the highest improvement of 41.72% in EER over the basic speaker verification model. Using trained SVIGEN2E as the speaker encoder in the VC model, this speaker encoder can effectively extract the speaker's timbre information. In the same time, in order to generate a content representations without source speaker information, upon constructing the content encoder, we use the bidirectional LSTM(Long-Short Term Memory) as an information filter to filter out content information and the content loss function to optimize the content encoder, so that the content encoder can effectively remove the speaker information and extract the content information in the source speaker's speech. In order to improve the quality of the generated speech, a progressive training method is proposed to train the RS-VC model. In the first step, the reconstruction loss function is used as the model loss function to train the model's capability to reconstruct the speech Mel spectrum. In the second step, the reconstruction loss function and the content loss function are used as the model loss function. The model optimizes the content encoder while reconstructing the speech Mel spectrum. Experiments show that the progressive training method produces better speech quality. Through the above improvements, this paper constructs and implements an arbitrary speaker VC algorithm based on feature separation. The experimental results verify the effectiveness of the algorithm in this paper, and the conversion effect reaches a good level.

II. SVINGE2E MODEL AND RS-VC MODEL

The one-shot voice conversion model built based on representations separation in this paper is implemented by the Encoder-Decoder structure. The encoder is composed of a Speaker Encoder and a Content Encoder. By optimizing the SV model, we obtained the optimized SV model named

SVINGE2E (Speaker Verification with Instance Normalization using Generalized End-to-End loss). We use SVINGE2E as the Speaker Encoder and construct the Content Encoder and Decoder in VC model. A progressive training method is designed for training RS-VC model.

A. SVINGE2E MODEL

In order to get SVINGE2E model, in this paper, we optimize the structure of the basic SV model [19] which is a three-layer structure with LSTM by adding a convolutional neural network (CNN), a fully connected neural network (FCNN) and an instance normalization (IN) to the middle layer (part) ($Y_1 \rightarrow Y_2$) of the three layers as shown in Fig.1. The loss function of the SVINGE2E model is the GE2E (Generalized End-to-End) loss function which is the same as the one for the basic SV model [19]. This improved verification model can more effectively extract features and facilitate the model convergence. The full structure of SVINGE2E model is shown in Fig. 1 and still maintains the three-layer LSTM structure but with extra CNN, FCNN and IN added to the middle layer.

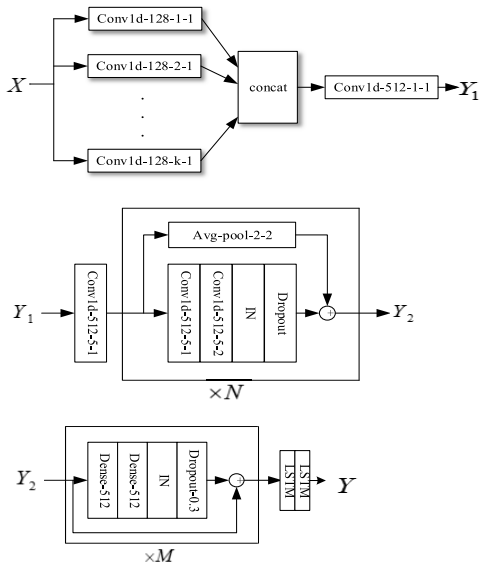


FIGURE 1. The full structure of SVINGE2E model obtained by improving the basic SV model – a three-layer LSTM structure.

In Fig. 1, X is a batch of speech utterances. IN represents instance normalization. These utterances X go through the model SVINGE2E to obtain the output Y of the last layer of LSTM. Take the vector y_T of the last step of the output Y as the representation of the each speech utterance in X. L2 normalization is used to obtain the feature vector e of X, which is used to calculate the model loss.

In this paper, the calculation method of the GE2E loss function of the model is as follows.

Suppose that each batch consists of N speakers, and each speaker has M utterances in training process. Then the feature representation (i.e., representation vector) e_{ji} of utterance i of speaker j obtained after each utterance passes through the model SVIGENE can be expressed as in (1)

$$e_{ji} = \frac{f\left(X_{ji}, W\right)}{\left\|f\left(X_{ji}, W\right)\right\|_2} \tag{1}$$

where X_{ji} is utterance i of speaker j in the batch; f represents the model SVINGE2E through which the utterance data pass; W represents the model SVINGE2E parameters; and $\|\ldots\|_2$ denotes L2 normalization. The vector obtained after the utterance passes through the model needs to be L2 normalized to get e_{ji} which is representation vector of utterance i of speaker j. We expect that the representation (feature) vector of each utterance is close to the representation vector of other utterance of the speaker, but different from the representation vector of the other speaker's utterance. During the model training process, we try to make the representation vectors of the same speaker's utterances as close as possible with each other. The representation vectors of different speakers are separated. The calculation formula of the representation vector of each speaker is given in (2)

$$c_j^{(-i)} = \frac{1}{M-1}\sum_{\substack{m=1 \\ m\neq i}}^{M} e_{jm} \tag{2}$$

$c_j^{(-i)}$ is the representation vector of speaker j that excludes utterance i, For a given representation vector and all speaker representation vectors in this batch, the similarity $S_{ji,k}$ calculation formula between the representation vectors is shown in (3):

$$S_{ji,k} = \begin{cases} w \cdot \cos\left(e_{ji}, c_j^{(-i)}\right) + b, & if\ (k == j) \\ w \cdot \cos\left(e_{ji}, c_k\right) + b, & otherwise \end{cases} \tag{3}$$

where w and b denote weight parameters. c_j is representation vector of the speaker j that includes all utterances of the speaker j. $S_{ji,k}$ is the similarity between the k-th speaker's representation vector and the representation vector of utterance i of speaker j, or in short, it represents the similarity between speaker k and utterance i of speaker j. The cosine distance is used in the formula to measure the similarity between representation vectors. In the model training process, when calculating the similarity between the utterance and the speaker, there are two cases according to whether the utterance belongs to the speaker: when $k = j$ that represents the utterance belongs to the speaker, the utterance's representation vector and the speaker's representation vector are positively correlated, then $c_j^{(-i)}$ is used to calculate the similarity and their similarity should be increased at that time; when $k \neq j$ that represents the utterance does not belong to the speaker, the correlated is negative, then c_j is used to calculate the similarity and, their similarity should be reduced at this time. The formula of the GE2E loss function are as shown in (4) and (5).

$$L\left(e_{ji}\right) = S_{ji,j} - \log\left(\sum_{k=1}^{N}\exp\left(S_{ji,k}\right)\right) \tag{4}$$

$$L_G(X, W) = L_G(S) = \sum_{j,i} L\left(e_{ji}\right) \tag{5}$$

where $L\left(e_{ji}\right)$ is the loss of each utterance, $L_G(X, W)$ is the loss of a batch of utterances.

B. RS-VC MODEL

This paper builds a one-shot voice conversion model based on representation separation (RS-VC). It is an Encoder-Decoder structure model. Encoder consists of a Content Encoder and a Speaker Encoder. Decoder compose of a decoder and a post-network. The Content Encoder encodes content information in the source speaker speech and remove speaker information to get Content Representation. The SE encodes speaker information in the target speaker speech to get Speaker Representation. The decoder synthesizes SR and content representation to generate a new synthesized speech, and the post-network supplements and improves the new synthesized speech. The trained SVING2E that acts as speaker encoder encodes the target speaker's speech and generates a vector containing the speaker's features. This vector is called the (target) speaker representation. The structure of the voice conversion model in this paper is shown in Figure 2.

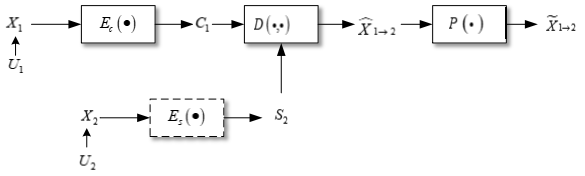


FIGURE 2. The whole architecture of the RS-VC.

In Fig. 2, X_1 is speech(utterance) of the source speaker U_1; X_2 is speech of the target speaker U_2; E_s is the Speaker Encoder in RS-VC, and also is the model SVINGE2E, which is used to extract the speaker representation S_2 of voice X_2, the training is done before RS-VC model is trained; E_C is the Content Encoder in RS-VC, which encodes the content information of the source speaker's speech X_1 to generate content representations C_1; $D(\cdot,\cdot)$ is a decoder, which synthesizes speaker representations S_2 and the content representation C_1 to generate a new Mel spectrum $\widehat{X}_{1\to2}$. $P(\cdot)$ is the post-network, which supplements $\widehat{X}_{1\to2}$ to generate a more perfect $\widetilde{X}_{1\to2}$.This is the voice conversion process from speaker U_1 to speaker U_2.

Suppose that U represents the speaker and Z represents the content in voice, X represents the utterance of the content Z spoken by U. In the RS-VC model, there are two cases to deal with while the speaker encoder extracting the speaker representation:

(1) For different utterances of a same speaker, the speaker representations (features) extracted from his utterances are the same. For example: if $U_1 = U_2$, then $E_s(X_1) = E_s(X_2)$.

(2) For the speech utterances from different speakers, the speaker representations are different from speaker to speaker. For example: if $U_1 \neq U_2$, then $E_s(X_1) \neq E_s(X_2)$.

During the training of RS-VC model, two loss functions are used to optimize the model, namely the content loss function and the reconstruction loss function. The content loss function mainly optimizes the content encoder, and the reconstruction loss function mainly helps decoder synthesize a new speech Mel spectrum. During the RS-VC model training, we use the same source speaker and target speaker, the process of training the RS-VC model is the process of using the target speaker as the source speaker to reconstruct the source speaker's speech. In this model voice conversion process, first the content encoder learns to remove the source speaker information from source speaker speech, retains the content information, and generates a content representation, and then the decoder uses speaker representations and content representations to reconstruct the source speaker speech. In the test and conversion of RS-VC model, when the source speaker and the target speaker are different, we can generate the target speaker's speech with the source speaker's speech content. In order to improve the model training effect, this paper proposes a progressive training method to train the RS-VC model. The training method is divided into two steps: spectrum reconstruction and content encoder optimization.

1) SPECTRUM RECONSTRUCTION

The main purpose of this step is to enable the decoder to correctly complete the reconstruction of the Mel spectrum by using the content representation and speaker representation. Therefore, at this step, we only use reconstruction loss function to train RS-VC model. By randomly selecting two utterances X_1 and X_1' of the same speaker, then the training process of the RS-VC model is shown in (6)-(9).

$$C_1 = E_c(X_1) \tag{6}$$
$$S_1 = E_s\left(X_1'\right) \tag{7}$$
$$\widehat{X}_{1\to1} = D(C_1, S_1) \tag{8}$$
$$\widetilde{X}_{1\to1} = P\left(\widehat{X}_{1\to1}\right) \tag{9}$$

where X_1 is speech of source speaker U_1; X_1' is a speech of target speaker U_1; C_1 is the content representation extracted from X_1 by the Content Encoder; S_1 is the speaker representation extracted from X_1' by Speaker Encoder; and $\widehat{X}_{1\to1}$ is the speech of the content representation C_1 and the speaker representation S_1 generated by the decoder. It is a conversion process from speaker U_1 to speaker U_1, i.e., the speech generated by the conversion of the speech of the U_1 to the speech of the U_1, also known as the reconstruction of the spectrum of the speech of the U_1. That is self-reconstruction of the source speaker's speech. $\widehat{X}_{1\to1}$ goes through post network and then the Mel spectrum $\widetilde{X}_{1\to1}$ is generated, so the training process of the RS-VC model is actually the process of Mel spectrum self-reconstruction. Therefore, the generated speech and the source speaker speech should be the same. Therefore, the loss function is used to measure the distance between the generated speech and the source speaker speech. Minimizing the distance between them is the process of reconstructing the speech. The reconstruction loss function of the RS-VC model

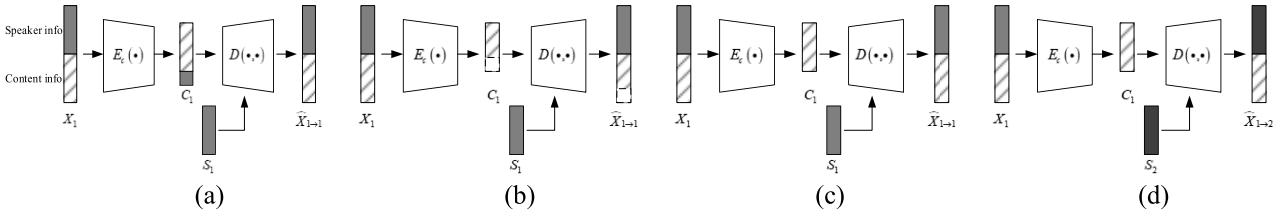


FIGURE 3. Representation (Feature) separation voice conversion (RS-VC) example diagrams: (a) indicates that the filter size is too large; (b) the filter size is too small; (c) the filter size is just right; and (d) represents the feature separation state in voice conversion test after RS-VC model is trained with the same filter size as used in (c) and it can be seen that the good voice conversion results are obtained.

is L_{recon}, as shown in (10)

$$L_{recon} = E\left[\left\|\widehat{X}_{1\to1} - X_1\right\|_1\right] \tag{10}$$

The calculation method of the loss function in this paper uses the average absolute loss. $\widetilde{X}_{1\to1}$ is the complement and perfection to $\|\ldots\|_1$ by the post network, so the post-network is a network structure that further reconstructs high-quality voice. It should be pointed out that during the experimental training, the source speaker and the target speaker are the same, so it's a self to self (i.e., 1→1) mapping, and during the testing, the source speaker is generally different from the target speaker. The loss function calculated using the source speaker is called the initial reconstruction loss function, as shown in (11)

$$L_{recon0} = E\left[\left\|\widetilde{X}_{1\to1} - X_1\right\|_1\right] \tag{11}$$

where $\|\ldots\|_1$ represents mean absolute loss. The calculation methods of the loss function L_{recon0} and loss function L_{recon} are the same, so when training in this step, the loss function L of the model is as follows (12)

$$L = L_{recon} + L_{recon0} \tag{12}$$

2) OPTIMIZATION OF CONTENT ENCODER

In this training step, the model optimizes the Content Encoder on the basis of the reconstruction spectrum. By using content filters and content loss functions, the Content Encoder removes the speaker information and extract content information. This paper uses the content loss function to optimize the Content Encoder. The loss function $L_{content}$ is shown in the (13)

$$L_{content} = E\left[\left\|E_c\left(\widehat{X}_{1\to1}\right) - C_1\right\|_1\right] \tag{13}$$

The speech content before and after the voice conversion is unchanged, so the distance between content representations extracted by the content encoder from the two speech is also close. The content loss function uses the average absolute to measure the distance between the speech contents. In this step, the loss function L of the RS-VC model is given by (14)

$$L = L_{\text{recon}} + \mu L_{content} + L_{recon0} \tag{14}$$

where μ represents the weight of content loss function in the model loss function. This paper uses the content loss function to optimize the Content Encoder, and at the same time set the content filter in the Content Encoder to control the content encoder to filter information, so the conversion principle of the RS-VC model in this paper is shown in the Fig. 3.

In this paper, the size of the content filter in the content encoder is set to control the amount of information that passes through the content filter. In Fig. 3, the grayed bar indicates speaker information, and the striped bar indicates content information. (a), (b), and (c) in the figure 3 indicate voice conversion example diagrams that the filter size is too large, too small and just appropriate in the training of RS-VC model respectively. The following filter-sizes of 16, 8, 4, 2, and 1 are tried respectively by experiments, and the experimental results show that size of 2 is more appropriate as used in (c). (d) shows a schematic diagram of voice conversion during conversion. (a) shows that the filter is too large when training the RS-VC model, not only the source content information is passed through, but also the target speaker information is passed through, so the generated content C_1 contains speaker information. (b) shows that the filter is too small during the training of the voice conversion model, which prevents the speaker information from passing through, and also prevents portion of the content information passing through. Therefore, the generated content indicates that C_1 contains only part of the content information, and the spectrum cannot be well reconstructed. (c) indicates that when the filter size is appropriate, the content information can completely pass through while completely preventing the speaker information from passing through, and the spectrum can be well reconstructed. (d) represents the feature separation state in voice conversion test after RS-VC model is trained with the same filter size as used in (c) and it can be seen that the good voice conversion results are obtained.

III. EXPERIMENTS

A. EXPERIMENT DATA

SVINGE2E: In this experiment LibriSpeech [20], VoxCeleb1 [21], and VCTK are used as training and testing data sets. The LibriSpeech data set is divided into a training set, a testing set, and a development set. The numbers of speakers in these datasets are 2,338, 73, and 73 respectively. The VoxCeleb1 dataset is divided into a training set and a testing set, and the total number of speakers is 1,251. The VCTK data set includes 109 speakers. In the SVINING2E experiment several numbers of speakers from the above three

TABLE 1. EER of SVINGE2E evaluated on test set under different training numbers.

Training set	Training number	Testing set	EER (%)		SVINGE2E Improvement compared with 3L-LSTM (%)
			3L-LSTM	SVINGE2E	
			A	B	$(A–B)/A$
VCTK	98	VCTK	10.46	9.58	8.41
		LibriSpeech	29.19	17.01	41.72
LibriSpeech	1,200	VCTK	6.26	5.68	9.26

TABLE 2. EER of SVINGE2E evaluated on test set under different training numbers.

Model	Training set	Trainee	EER(%)		
			VCTK	LibriSpeech	VoxCeleb1
SVINGE2E (Ours)	VCTK	98	9.58	17.01	37.73
	LibriSpeech	1,200	5.68	5.11	30.30
	LibriSpeech	2,300	4.39	4.77	28.70
	LibriSpeech+VoxCeleb1	2,800	3.66	3.97	24.67
	LibriSpeech+VoxCeleb1	3,500	3.15	3.92	19.81

data sets are selected randomly for training and testing. The Mel spectrum of the original speech is extracted and used as the input data of model SVINGE2. SVINGE2E is trained with a batch size of 64 × 10, using the ADAM optimizer.

RS-VC: In this paper, the RS-VC experiment uses the VCTK data set as the training set and the test set. The data set has a total of 109 speakers, from which 10 speakers are randomly selected as the testing set. The 10 speakers are p225, p226, p227, p228, p229, p230, p231, p232, p233, p234. In the experiment, Wavenet [11] is used as the vocoder to generate the original speech from the Mel spectrum. In our implementation, the frame rate of the Mel spectrum is 62.5Hz and the sampling rate of speech waveform is 16 kHz. RS-VC is trained with a batch size of 24, using the ADAM optimizer. The weight μ in Eq. (14) is set to 100. We use MOS(Mean Opinion Score) and ABX Tests as the evaluation criteria of the RS-VC model experiment. Here A represents the source speaker's speech, B represents the target speaker's speech, and X is the converted speech to determine whether X is more similar to A or B. MOS is divided into naturalness of MOS and similarity of MOS, and their scores range from 1 to 5. The larger the score, the higher target similarity or naturalness. ABX Test determines whether the converted speech is more similar to the original speech or to the target speech.

B. SVINGE2E EXPERIMENT

The model of SVINGE2E in this paper is improved on the basis of the simple basic three-layer LSTM model [19] (3L-LSTM). The model is evaluated on VCTK dataset and LibriSpeech dataset. Tree-layer LSTM model experiment data come from paper of [22].When using the dataset VCTK as the training set of the model, the data set is divided into a training set and a testing set according to the ratio of 9:1. Specifically, there are 98 speakers in the training set and 11 speakers in the testing set, when the VCTK training set is used to train the model. The model is evaluated on VCTK testing set and LibriSpeech testing set. When LibriSpeech dataset is used as the model training set, 1,200 speakers are randomly selected from the training set of the LibriSpeech. After training the model using this LibriSpeech training set, the model is evaluated on VCTK testing dataset and LibriSpeech data set. The model test uses Equal Error Rate (EER) as the evaluation metric which is the main evaluation metric for speaker verification model. In general, there are three metrics used to evaluate the model of SVINGE2E: False Acceptance Rate (FAR), which is defined as $FAR = \frac{FP}{FP+TP}$; False Reject Rate (FRR), which is defined as $FRR = \frac{TN}{TN+FN}$; and Equal Error Rate (EER), which is defined as the value when FAR = FRR. Here *FP* represents that a speech actually does not belong to a speaker but is judged by the model to belong to that speaker; *TP* represents that a speech actually belongs to a speaker and is judged by the model to belong to that speaker; *TN* represents that a speech actually belongs to a speaker but is judged by the model as not belonging to the speaker, while *FN* represents that a speech actually does not belong to the speaker and is judged by the model as not belonging to the speaker. The test results are shown in the Table 1. In this paper, the model SVINING2E uses GE2E loss as model loss function, so the training effect of the model based on this loss function is related to the number of speakers in the training set. Experiments are conducted with different number of trainees, and the obtained results are shown in Table 2. From Table 1, it can be seen that when SVINING2E is trained on the VCTK training set, the EER tested on the VCTK is 9.58, and compared with 10.46 of the 3L-LSTM model, decreased by 0.88, which means an improvement of 8.41% by the model of SVINING2E; The EER tested on the LibriSpeech testing set is 17.01, and compared with 29.19 of the basic model 3L-LSTM, has a drop of 12.18, corresponding to an improvement of 41.72% with the model of SVINING2E. When the model of SVINING2E is trained on the LibriSpeech training set, the EER evaluated on

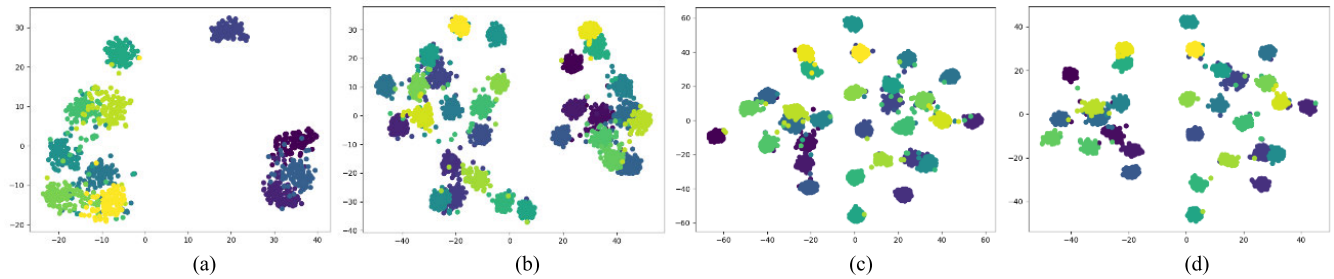


FIGURE 4. Dimension reduction graphs of speaker representations extracted by SVINING2E model trained with different training numbers of (a) 98, (b)1200, (c)2300 and (d) 3500.

the VCTK data set is 5.68, and compared with the 6.26 of the basic model, decreased by 0.58, corresponding to an improvement rate of 9.26%; while the EER tested on the LibriSpeech testing set is 5.11, rise by a marginal amount of 0.03,. From Table 1 we can know that our model SVINGE2E is better than 3L-LSTM.

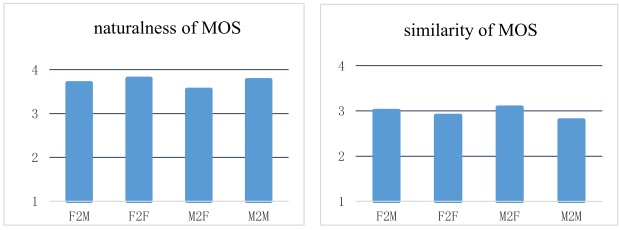
It can be seen from Table 2 that as the number of speakers (the trainees) in the training set increases, the lower the EER of the model in each testing set, the better the model performance. At the same time, it can be seen from the model experiments in this paper that the performance of VoxCeleb1 is worse than the other two testing sets. From Table 2, it can be concluded that when the training number of speakers reaches 3,500, the model performs best on VCTK, and the EER is 3.15.

When the EER is smaller, the speaker representations of the same speaker are better, and the distance between the speaker representations of different speakers is larger. Therefore, the speaker representations extracted from the trained SVINGE2E model data set are displayed on the two-dimensional image (Fig. 4) which is clustered. In this paper, the model SWINGE2E is trained on a dataset with trainees of 98(Fig.4(a)), 1,200(Fig.4(b)), 2,300(Fig.4(c)), and 3,500(Fig.4(b)) respectively. After training, 30 speakers are randomly picked up and 100 utterances are randomly selected from each speaker in the VCTK dataset to extract the speaker representation. In the Fig.4(a) we use VCTK test dataset which only contains 10 speakers to extract the speaker representation of each utterance in this dataset. t-SNE (t-distributed stochastic neighbor embedding) [23] is used for dimension reduction display, as shown in Figure 4.

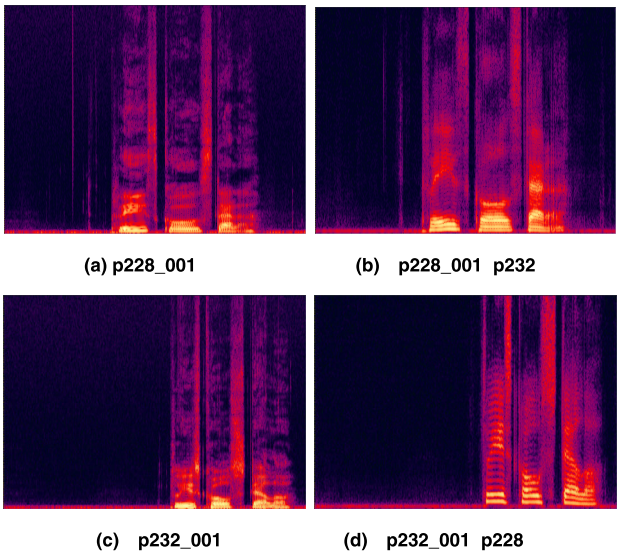
C. RS-VC EXPERIMENT

The RS-VC experiment here is divided into two parts according to the test data: the in-set speaker test and the out-set speaker test. The in-set speaker test means that the source speaker and the target speaker in test speeches have other speeches as training set to train the model. The out-set speaker test means that neither the source speaker nor the target speaker in test speeches has any speech as a training set to train the model. The test data consists of 8 speakers: 4 from the in-set (two males and two females) and other 4 from the out-set (two males and two females). 5 voices are selected from each speaker for the test. There are total 40 voices that need scoring. During scoring, 10 volunteers are selected to conduct the in-set speaker test and the out-set speaker test respectively. First the in-set speaker test is conducted and then comparison is done with other models. When comparing scores with the other models, the volunteers are first taught the scoring criteria of the two evaluation methods, then given the comparison model samples and the corresponding scores of the samples, and finally score the test results (converted voices) on RS-VC experiment. When conducting out-set speaker test, because the models under comparison are unable to realize the speech conversion between out-set speakers, the testing scores are given according to the criteria for the in-set speaker test. The main purpose of this research is to realize the direct speech conversion in the out-set speakers.

TABLE 3. Converted voice MOS scores.



In this RS-VC experiment, this paper uses two evaluation methods: MOS and ABX Test. Both of these evaluation methods are subjective evaluation methods. After the model is trained, the test set is used for voice conversion testing. The test set focuses on the source speaker and the target speaker, while the training set does not include them. According to the gender of speakers, the conversion is divided into male to female (M2F), female to male (F2M), male to male (M2M), and female to female (F2F). We train SVINGE2E with a training set of 3500 trainees and then use it as the speaker encoder of RS-VC after the training. When neither the source speaker nor the target speaker belongs to the training set, the naturalness and similarity of the voices after voice conversion between different genders and within the same genders are listed below Table 3.



(a) p228_001 (b) p228_001 p232

(c) p232_001 (d) p232_001 p228

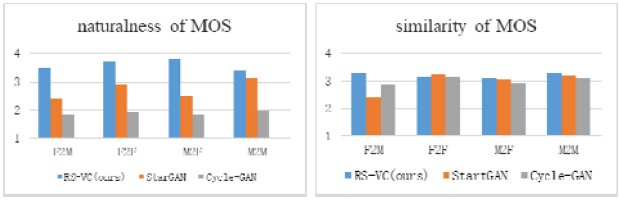
FIGURE 5. Voice spectrum chart of converted voice.

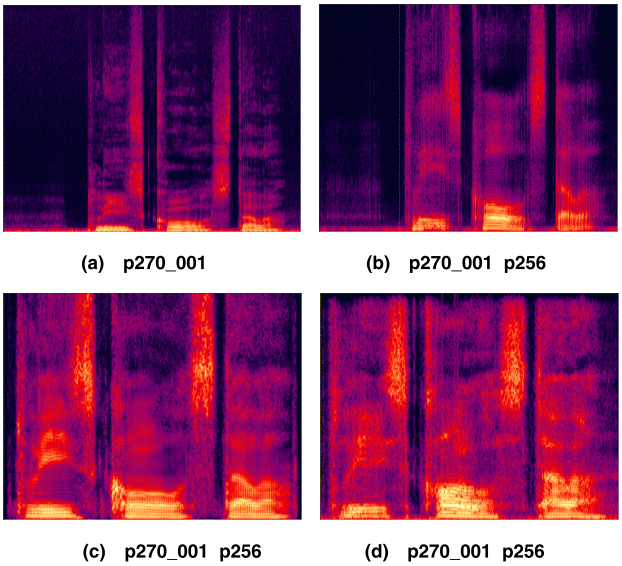
It can be seen from the Fig.5 that (a) is the speech spectrum of the speech p228_001 of the female source speaker p228; (b) is the speech of the male speaker p228_001×p232 obtained after voice conversion of the original speech (a) p228_001 via the male target speaker p232; (c) is the male speaker speech spectrum p232_001 of the male source speaker p232; (d) is the voice p232_001×p228 of the female speaker obtained after the voice conversion of (c) p232_001 via the female target speaker p228. It can be seen that the voice content before and after conversion is basically unchanged, and the energy distribution changes.

This paper also compares the RS-VC model with other voice conversion models: Cycle-GAN [16] model and the StarGAN model [15] respectively. The two models compared with RS-VC are called comparison models. In the comparative experiment, the similarity of MOS and the naturalness of MOS are used as the evaluation criteria. Because the two comparison models can only conduct the conversion between speakers in the training set during the voice conversion test, while in this paper, upon performing the experimental evaluation with the two comparison models, the voice conversion is done between the speakers from the training set, but the test speech did not appear in the training set during the conversion, that is, the speaker's other speech is used for model training, but the speaker's test speech is never used for model training. The evaluation results are shown in Table 4. From Table 4, it can be seen that the naturalness of the converted speech of this model is significantly higher than that of the two comparison models, and the similarity of the converted speech is slightly higher than that of the two comparison models in most cases. In this paper, some examples of the converted speech of the two comparative models and RS-VC model is shown in Fig. 6.

In Fig. 6, (a) is the source speaker's speech p270_001; (b) is the converted speech p270_001×p256 of the RS-VC

TABLE 4. Comparison on MOS scores between RS-VC model and other models.





(a) p270_001 (b) p270_001 p256

(c) p270_001 p256 (d) p270_001 p256

FIGURE 6. The speech spectrum chart of the speech converted by RS-VC and the comparison models.

model where the target speaker is p256; (c) is the converted speech p270_001×p256 of the StarGAN voice conversion model where the target speaker is p256; (d) is the converted speech p270_001×p256 of the Cycle-GAN voice conversion model where the target speaker is p256.

IV. CONCLUSION

In this paper, the speaker verification model is applied to the field of VC, and a one-shot voice conversion algorithm based on Representation Separation (RS-VC) is designed and implemented. This algorithm can realize voice conversion between any speakers. We improved the speaker verification model and obtained the optimized speaker verification model called SVINGE2E which reduced the equal error rate (EER) and enhanced its capability to extract speaker representation purity. In the RS-VC model, the speaker representation extracted by SVINGE2E is used for training and testing the RS-VC. RS-VC model has encoder-decoder structure, in which the encoder is composed of two encoders: a content encoder and a speaker encoder, and the decoder is composed of a decoder and a post network. The decoder synthesizes the representations generated by the two encoders and generates new speech, and the post-network complements the speech generated by the decoder. The reconstruction loss function is used to help the decoder and the post-network to reconstruct

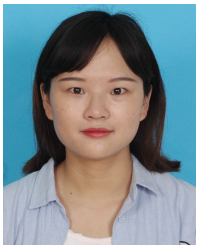
the speech. The filter of content encoder and content loss function are set to help the content encoder effectively remove the speaker information, and generate the content representation. The progressive training method is used to train RS-VC model with a view to improving the quality of the generated speech.

Further work could include (1) improving the quality of model speech Mel spectrum reconstruction; (2) merging the speaker encoder into the voice conversion model for training to realize a simple end-to-end voice conversion; (3) further improving the effect of representation separation; and (4) expanding the scope of voice conversion, not only for timbre conversion, but also for prosody and rhythm.

REFERENCES

[1] M. Abe, K. Shikano, and H. Kuwabara, "Statistical analysis of bilingual speaker's voice for cross-language voice conversion," *J. Acoust. Soc. Amer.*, vol. 90, no. 1, pp. 76–82, 1991.
[2] K. Shikano, S. Nakamura, and M. Abe, "Speaker adaptation and voice conversion by codebook mapping," in *Proc. IEEE Int. Sympoisum Circuits Syst.*, Jun. 1991, pp. 594–597.
[3] M. Eslami, H. Sheikhzadeh, and A. Sayadiyan, "Quality improvement of voice conversion systems based on trellis structured vector quantization," in *Proc. Conference Int. Voice Commun. Assoc.*, 2011, pp. 665–668.
[4] H. Fang, X. Ning, and L. Haiyan, "Improvement of voice conversion algorism based on codebook mapping," *Microprocessors*, vol. 43, pp. 35–38, 2015.
[5] T. Toda, J. Lu, H. Saruwatari, and K. Shikano, "Straight-based voice conversion algorithm based on Gaussian mixture model," in *Proc. Conf. Int. Voice Commun. Assoc.*, 2000, pp. 279–282.
[6] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum," *Trans. Inst. Electron. Inf. Commun. Eng.*, vol. 2, no. 2, pp. 841–844, 2001.
[7] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," in *Proc. 9th Int. Conf. Spoken Lang. Process. INTERSPEECH (ICSLP)*, vol. 5, 2006, pp. 2266–2269.
[8] J. Zhihua and Y. Zhen, "A method for voice conversion based on Viterbi algorithm," *Acta Electronica Sinica*, vol. 37, no. 7, pp. 1470–1475, 2009.
[9] L.-H. Chen, Z.-H. Ling, W. Guo, and L.-R. Dai, "GMM-based voice conversion with explicit modelling on feature transform," in *Proc. 7th Int. Symp. Chin. Spoken Lang. Process.*, Nov. 2010, pp. 364–368.
[10] E.-K. Kim, S. Lee, and Y.-H. Oh, "Hidden Markov model based voice conversion using dynamic characteristics of speaker," in *Proc. Conf. Int. Voice Commun. Assoc.*, 1997, pp. 2519–2522.
[11] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*. [Online]. Available: http://arxiv.org/abs/1609.03499
[12] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 3, pp. 631–644, Mar. 2019.
[13] L. Liu, Z. Ling, Y. Jiang, M. Zhou, and L. Dai, "Wavenet vocoder with limited training data for voice conversion," in *Proc. Interspeech*, 2018, pp. 1983–1987. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1190
[14] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," 2017, *arXiv:1704.00849*. [Online]. Available: http://arxiv.org/abs/1704.00849
[15] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2018, pp. 266–273.
[16] T. Kaneko and H. Kameoka, "Parallel-Data-Free voice conversion using cycle-consistent adversarial networks," 2017, *arXiv:1711.11293*. [Online]. Available: http://arxiv.org/abs/1711.11293
[17] J.-C. Chou, C.-C. Yeh, H.-Y. Lee, and L.-S. Lee, "Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations," 2018, *arXiv:1804.02812*. [Online]. Available: http://arxiv.org/abs/1804.02812
[18] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "AUTOVC: Zero-shot voice style transfer with only autoencoder loss," 2019, *arXiv:1905.05879*. [Online]. Available: http://arxiv.org/abs/1905.05879
[19] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4879–4883.
[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.
[21] A. Nagrani, J. Son Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," 2017, *arXiv:1706.08612*. [Online]. Available: http://arxiv.org/abs/1706.08612
[22] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 4480–4490.
[23] L. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

CHUNHUI DENG received the bachelor's degree from Xidian University, Xi'an, China, in July 1985, and the master's degree from Shanghai Jiao Tong University, in March 1991. Since January 1995, he has been with the Computing Center and the School of Information, Hainan University, Haikou, China, where he also served as the Deputy Director and as an Associate Professor for the Computing Center. From 2004 to 2006, he served as the Secretary General of Hainan Electronic Society. Since July 2009, he has been working with the Guangzhou College, South China University of Technology, Guangzhou, China. In 2016, he was appointed as a Professor and the Dean of the School of Computer Engineering, and elected as a member of Guangdong Software Engineering Teaching Steering Committee. While working at the Guangzhou College, he has published more than 15 articles and chaired four provincial-level teaching and research projects. His research interests include information processing, data mining, the Internet of Things, cloud computing, and big data.



YING CHEN received the bachelor's degree in computer science and technology from the Hefei University of Technology, Anhui, China, in 2017, and the master's degree from the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China, in 2020, under the supervision of Prof. H. Deng. Her research interests include information processing, neural networks, data mining, voice conversion, and machine learning.

HUIFANG DENG received the B.Sc. and M.Sc. degrees in China, and the Ph.D. degree from University College London (UCL), U.K.

From 1989 to 2004, he was studying, working, and living in the U.K. (for nearly 16 years). From 2001 to 2004, he served as the Chief Technical Officer and Chief Scientist at Sunrise Systems Limited, Cambridge, U.K. In September 2004, he was the Dean of the Software School, South China University of Technology, Guangzhou, China, where he is currently a Full Professor. So far, he has published over 140 articles and holds several patents for invention. His research interests include RFID technology and applications, the Internet of Things, cloud computing, big data, AI, machine learning, data mining, data fusion, neural networks, social computing, computer modeling, high-performance computing, and scientific computing.