[1]Chenghuan Xie

[1]Aimin Zhou

SVCGAN: Speaker Voice Conversion Generative Adversarial Network for Children's Speech Conversion and Recognition

Abstract: Automatic speech recognition (ASR) refers to a technological process that entails the conversion of spoken language into written text. However, the acoustic distinctions between children's speech and adult speech are substantial, rendering the automatic speech recognition system trained on adult speech inadequate for effectively recognizing children's speech. To overcome this issue, in this study, we propose speaker conversion generative adversarial network (SVCGAN). SVCGAN is a novel non-parallel voice conversion model, which enhances three key areas: log-cosh loss, semantic-similarity loss, and third adversarial loss. Therefore, the incorporation of these losses better protects semantic information for young children during voice conversion process and improves the quality of the converted speech. Additionally, the character error rate (CER) of children's speech recognition can benefit from children's speech transformed into adult speech. Experimental results suggest that SVCGAN demonstrates superior performance across multiple dimensions compared to both CycleGAN-VC3 and MaskCycleGAN-VC models. It encompasses training efficiency, semantic information similarity, voice type similarity, sound naturalness and intelligibility, which leads to a reduction in the CER of speech recognition for young children.

Keywords: Children's speech conversion, Voice conversion (VC), Generative adversarial network (GAN), Children's speech recognition

INTRODUCTION

Automatic speech recognition (ASR) refers to a technological process that converts spoken audio signals into written text. Traditional ASR systems employ hidden Markov models (HMMs) while contemporary ASR systems predominantly rely on machine learning and deep learning techniques. These systems utilize popular computer tools like Kaldi[1], PyTorch[2], and TensorFlow[3] to construct and train models using extensive datasets. ASR is extensively employed in several domains such as voice assistants, smart homes, autonomous driving, meeting note

[1] School of Computer Science and Technology, East China Normal University，Shanghai, 200062，China

Corresponding Author : Chenghuan Xie

Email：xchPeter@126.com

Aimin Zhou： Email: amzhou@cs.ecnu.edu.cn

transcription, telephone customer support, and educational applications.

The application of speech recognition technology in the context of children's education holds significant potential. By utilizing speech recognition and analyzing the language used by young children in their daily interactions, it is possible to enhance their language proficiency and facilitate the acquisition of language skills. Nevertheless, there are numerous factors that impact the precision of speech recognition in children. It is common practice to train automatic speech recognition systems using adult speech data. However, significant disparities exist between the vocal characteristics of children and adults[4][5][6][7], which might result in diminished accuracy when applying speech recognition to children's speech[8][9][10][11][12]. Furthermore, the amount of data on children's voice is very limited[13][14]. The adult voice dataset that is accessible to the general public has a substantial amount of training data, exceeding 1,000 hours[15] and potentially even reaching 10,000 hours[16]. In contrast, the kid voice dataset is limited to only a few hours of speech data.

To address the aforementioned issues, a potential approach involves adjusting the tone and pace of children's speech to align it more closely with adult speech[17][18][19][20]. This adjustment aims to mitigate the rate of recognition errors. An alternative methodology is employed to investigate the relationship between age and the modulation of formant frequency in sound[21][22][23]. The empirical evidence indicates that there exists a disparity in the formant frequency between children's and adult's speech, which can be attributed to the variation in vocal tract length between these two age groups[24]. Linear prediction is utilized to modify the formant of children's voice[25][26], thereby aligning it with the formant of adult speech in order to mitigate the occurrence of recognition errors. However, the above methods only modify the voice type of children's speech from a single aspect, and can not make the modified speech close to the adult's speech well, so the effect is very limited.

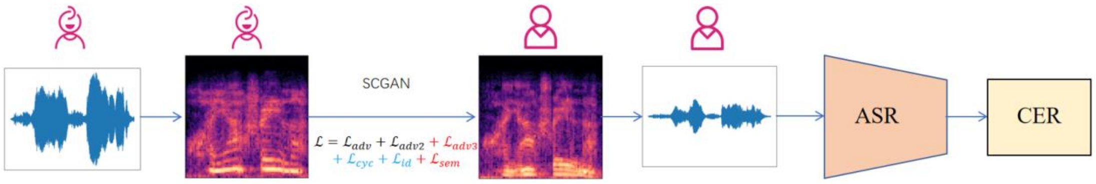

SCGAN
$\mathcal{L} = \mathcal{L}_{adv} + \mathcal{L}_{adv2} + \mathcal{L}_{adv3} + \mathcal{L}_{cyc} + \mathcal{L}_{id} + \mathcal{L}_{sem}$
ASR
CER

Figure 1. The flow chart of the whole experiment. SVCGAN transforms children's mel-spectrograms into adult mel-spectrograms. The utilization of ASR is employed to obtain the CER for adult speech signals. The blue objectives cycle-consistency loss ($\mathcal{L}_{cyc}$) and identity-mapping loss ($\mathcal{L}_{id}$) have been enhanced. The red objectives semantic-similarity loss ($\mathcal{L}_{sem}$) and third adversarial loss ($\mathcal{L}_{adv3}$) are proposed.

In this paper, we employ voice conversion (VC) technology to transform children voice type into adult voice type and enhance the accuracy of speech recognition in young children. VC is a method to change the voice type of the speaker without changing the semantic information of the speech. MaskCycleGAN-VC[27] is a widely used VC model to convert speech between adults. However, MaskCycleGAN-VC is not effective in the task of converting children's speech into adult speech. There is a huge difference between children's speech and adult speech, so MaskCycleGAN-VC cannot well protect the semantic information of the original children's speech, resulting in a decline in the quality of the converted speech. In order to solve this problem and enhance the accuracy of speech recognition for young children, we propose the utilization of a speaker voice conversion generative adversarial network (SVCGAN) as an advancement to the existing MaskCycleGAN-VC model. The model's performance is enhanced through the utilization of three key components: (1) an improved objective known as log-cosh loss, (2) an improved objective semantic-similarity loss, and (3) an improved objective third adversarial loss. The impacts of each enhanced aim are examined in relation to the conversion of young children's speech to adult speech and the accuracy of young children's speech recognition. Fig.1 illustrates the complete procedure of the experiment. The results of an objective evaluation indicate that our proposed enhancement offers several advantages over MaskCycleGAN-VC. Specifically, it enhances the training efficiency of the model, improves the semantic information similarity between the converted speech and the source speech, enhances the voice type similarity between the converted speech and the target speech, and improves the accuracy rate of speech recognition for young children. The subjective assessment indicates that our model exhibits superiority over MaskCycleGAN-VC in terms of semantic information similarity, sound naturalness and intelligibility. Additionally, our model demonstrates comparable performance in voice type similarity. We summarize our contributions below:

(1) We point out limitation of the existing model, which lacks of ability to retain the semantic information of children's speech and leads to poor converted speech quality.

(2) Present the SVCGAN model as a proposed approach for enhancing the conversion of children's speech into adult speech. SVCGAN improves the quality of the converted speech, especially in semantic information extraction aspect.

(3) Experiments show that three improved objectives augment the semantic information similarity, voice type similarity, sound naturalness and intelligibility of converted speeches. The model demonstrates a decreased character error rate (CER) in the converted children's speech as compared to the original children's speech.

RELATED WORK

Voice Conversion (VC) is a method utilized to transform the vocal characteristics of one individual into those of another individual, while ensuring the retention of semantic information. VC is a commonly employed technique in the fields of voice assistance[28][29], speech enhancement[30][31], and accent modification[32][33]. Neural networks are extensively employed in the field of VC, encompassing recurrent neural networks[34], attention networks[35][36][37], and generative adversarial networks (GANs).

Parallel VC approaches were commonly employed in the early stages, utilizing supervised training generators based on parallel corpora. One drawback of the parallel VC model is its reliance on parallel corpus databases, which are often challenging to acquire. In contrast, non-parallel corpus databases are more commonly utilized in everyday contexts.

One potential enhancement is non-parallel VC training, a technique that enables the training of generators utilizing non-parallel corpora without the need for supplementary data or pre-trained models. Non-parallel VC techniques employ GANs[38][39][40] and variational autoencoders (VAEs)[41][42], with notable success demonstrated by CycleGAN-VC[43] and StarGAN-VC[44][45][46]. CycleGAN-VC2[47] enhances the generator and discriminator components of CycleGAN-VC, while incorporating a secondary adversarial loss to further enhance overall performance. However, it has been observed that CycleGAN-VC and CycleGAN-VC2 fail to effectively capture the time-frequency structure, limiting its use to mel-cepstrum conversion alone. The challenge at hand is addressed by the introduction of CycleGAN-VC3[48] and MaskCycleGAN-VC methodologies. CycleGAN-VC3 and MaskCycleGAN-VC have been offered as potential solutions for addressing this particular issue. CycleGAN-VC3 incorporates a time-frequency adaptive normalizing module to preserve the harmonic structure. In contrast, MaskCycleGAN-VC does not rely on an extra module and instead employs a filling in frames (FIF) approach during training to facilitate the converter in learning how to fill in missing frames.

METHOD

The MaskCycleGAN-VC model is employed to train a function $G_{X \to Y}$ that converts source acoustic features $x \in X$ into target acoustic features $y \in Y$ in the absence of parallel voice recordings. MaskCycleGAN-VC draws inspiration from CycleGAN[49], an algorithm designed for unpaired image-to-image translation. The MaskCycleGAN-VC model incorporates various loss functions, including adversarial loss $\mathcal{L}_{adv}^{X \to Y}$ and $\mathcal{L}_{adv}^{Y \to X}$, cycle-consistency loss[50] $\mathcal{L}_{cyc}$, identity-mapping loss[51] $\mathcal{L}_{id}$, second adversarial loss $\mathcal{L}_{adv2}^{X \to Y \to X}$ and $\mathcal{L}_{adv2}^{Y \to X \to Y}$, in order to tackle the aforementioned problem. We present SVCGAN, which enhances the performance of MaskCycleGAN-VC in three specific areas.

Improved Objective: Log-Cosh Loss

In the MaskCycleGAN-VC framework, the cycle-consistency loss and identity-mapping loss are both implemented using the L1 loss function. However, the utilization of L1 loss results in a deficiency of smoothness and reduces efficiency when applies to extensive datasets during the training phase. To address these issues, the SVCGAN algorithm employs log-cosh loss as an alternative to L1 loss. In contrast to L1 loss, log-cosh loss exhibits the desirable properties of being smooth and differentiable across its whole domain. Moreover, log-cosh loss has been seen to facilitate faster convergence during the training phase. The cycle-consistency loss $\mathcal{L}_{cyc}$ in the SVCGAN framework is expressed as

$$\mathcal{L}_{cyc} = \mathbb{E}_{x \sim P_X}[\log(\cosh(G_{Y \to X}(G_{X \to Y}(x)) - x))] + \mathbb{E}_{y \sim P_Y}[\log(\cosh(G_{X \to Y}(G_{Y \to X}(y)) - y))],$$

and the identity-mapping loss $\mathcal{L}_{id}$ in SVCGAN framework is expressed as

$$\mathcal{L}_{id} = \mathbb{E}_{y \sim P_Y}[\log(\cosh(G_{X \to Y}(y) - y))] + \mathbb{E}_{x \sim P_X}[\log(\cosh(G_{Y \to X}(x) - x))].$$

Improved Objective: Semantic-Similarity Loss

MaskCycleGAN-VC can’t well catch semantic information when there is a substantial variance between different speakers’ voice types. In order to alleviate this problem, we apply a semantic-similarity loss on the converted feature $G_{X \to Y}(x)$, as

$$\mathcal{L}_{sem} = \mathbb{E}_{x \sim P_X}[\log(\cosh(G_{X \to Y}(x) - x))] + \mathbb{E}_{y \sim P_Y}[\log(\cosh(G_{Y \to X}(y) - y))].$$

The utilization of a semantic-similarity loss promotes the conversion of the feature $G_{X \to Y}(x)$ to exhibit a higher degree of semantic similarity to the original data x, while also encourages the converted feature $G_{Y \to X}(y)$ to display a greater degree of semantic similarity to the original data y. The aforementioned loss function contributes to the enhancement of training efficiency in SVCGAN and facilitates the attainment of a higher degree of semantic information. The incorporation of a hyper-parameter λ_{sem} is necessary in determining the relative significance of the semantic-similarity loss within the full objective. An excessively high hyper-parameter λ_{sem} leads to difficulty on voice type converted.

Improved Objective: Third Adversarial Loss

A third adversarial loss is employed to quantify the distinguishability between the identity-mapping data $G_{Y\to X}(x)$ and the source data x. An additional discriminator, denoted as D_X'', is introduced for the identity-mapping data. Third adversarial loss is expressed as

$$\mathcal{L}_{adv3}^{Y\to X} = \mathbb{E}_{x\sim P_X}[\log D_X''(x)] + \mathbb{E}_{x\sim P_X}[\log(1 - D_X''(G_{Y\to X}(x)))].$$

The discriminator D_X'' aims to maximize the third adversarial loss to effectively distinguish between genuine data samples x and identity-mapped data samples $G_{Y\to X}(x)$. The generator $G_{Y\to X}$ generates $G_{Y\to X}(x)$ with the objective of deceiving the discriminator D_X'' by minimizing the third adversarial loss.

Full Objective

The complete objective $\mathcal{L}$ is written as

$$\mathcal{L} = \mathcal{L}_{adv}^{X\to Y} + \mathcal{L}_{adv}^{Y\to X} + \mathcal{L}_{adv2}^{X\to Y\to X} + \mathcal{L}_{adv2}^{Y\to X\to Y} + \mathcal{L}_{adv3}^{X\to Y} + \mathcal{L}_{adv3}^{Y\to X} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{id}\mathcal{L}_{id} + \lambda_{sem}\mathcal{L}_{sem}$$

The weighing parameters are denoted as λ_{cyc}, λ_{id}, and λ_{sem}. The generators $G_{X\to Y}$ and $G_{Y\to X}$ are required to minimize the complete objective function $\mathcal{L}$, whereas the discriminators D_X, D_Y, D_X', D_Y', D_X'', D_Y'' are expected to maximize the complete objective function $\mathcal{L}$.

EXPERIMENTS

Experimental conditions

Dataset. We assess the efficacy of SVCGAN by employing the AISHELL-2021C-EVAL and AISHELL-ASR0009-OS1[52] datasets. These datasets consist of recordings of standard Chinese pronunciation from both children and adults, and are devoid of any background noise. To ensure that the children's recordings in the AISHELL-2021C-EVAL dataset accurately reflect the language environment and language expression level of young children in their daily lives, we have introduced environmental noise and selectively removed certain frames from the children's speech files. In our methodology, we allocate 90 percent of the recordings for training purposes, while the remaining 10 percent are reserved for testing. The kid and adult datasets do not contain identical statements, resulting in a training process that is entirely non-parallel.

Conversion and synthesis process. The implementation of the conversion and synthesis process closely resembles that of MaskCycleGAN-VC, enabling a comparison of performances. The MaskCycleGAN-VC model is employed to transform the mel-spectrogram of recorded audio samples, followed by the synthesis of the corresponding waveform using the pretrained MelGAN vocoder[53].

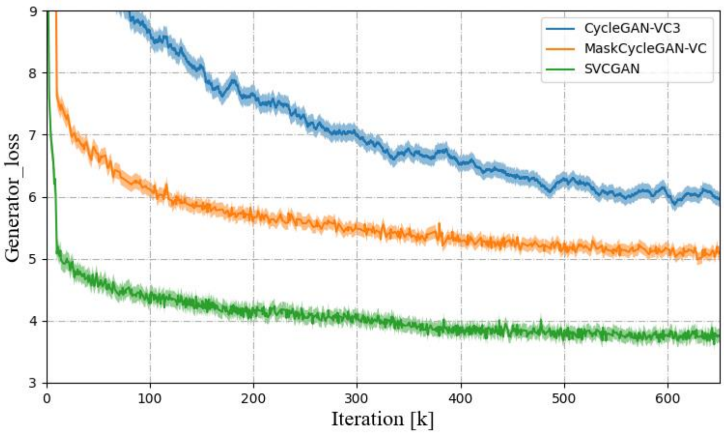
Network architectures. In order to facilitate performance comparison with MaskCycleGAN-VC, the network architectures of SVCGAN are identical to those of MaskCycleGAN-VC. The generator employs a 2-1-2D CNN, whereas the discriminator utilizes PatchGAN[54].

Training settings. In order to provide a meaningful comparison of performance with MaskCycleGAN-VC, the training settings of SVCGAN are aligned with those applied in MaskCycleGAN-VC. In the preprocessing stage, the mel-spectrograms are subjected to normalization. The employed GAN objective is a least-squares

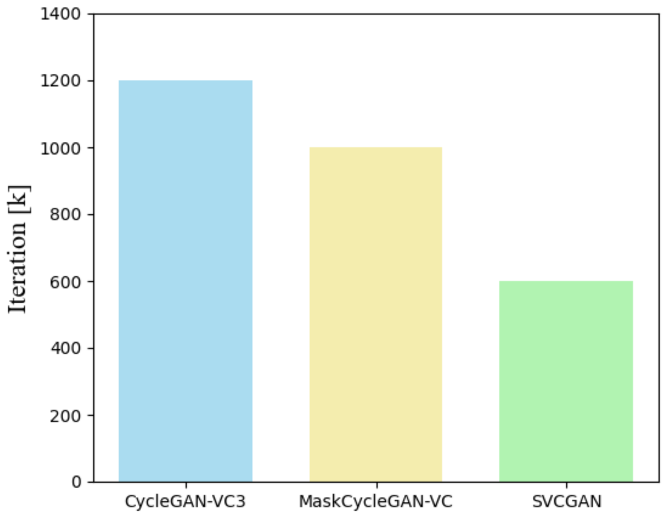
GAN[55], and the generator and discriminator are trained using an Adam optimizer. The learning rates for the generator and discriminator are set to 0.0002 and 0.0001, respectively. The momentum terms β_1 and β_2 are initialized to 0.500 and 0.999, respectively. The batch size is configured as 8, and the training samples comprise 64 frames that have been randomly cropped. The hyper-parameter λ_{cyc} is assigned a value of 10. The hyper-parameters λ_{id} and λ_{sem} are assigned a value of 5, and they are exclusively utilized during the initial 10,000 iterations in order to enhance the efficacy of voice type transformation.

Objective evaluation

Four measures are employed for the purpose of objective evaluation: (1) **Training efficiency.** The training iterations necessary for achieving model convergence are employed as a metric for evaluating the training efficiency. When the model has reached convergence, fewer training iterations means more efficient in training efficiency. (2) **Mel-cepstral distortion (MCD).** The MCD is commonly employed for evaluating the difference between converted mel-cepstral features and target mel-cepstral features. Mean squared error is employed as a metric for computing the MCD. It is imperative to note that there exists no identical statement inside the child and adult datasets. Consequently, a direct comparison between the converted mel-cepstral features and the target mel-cepstral features is unattainable. As an alternative, we solely assess the similarity of semantic information in MCD by quantifying the MCD between converted mel-cepstral features and the original child mel-cepstral features. A lower value of the MCD indicates a higher degree of semantic information proximity. (3) **Voiceprint recognition (VPR).** The aforementioned MCD solely evaluates the semantic information similarity. The VPR technique is employed to demonstrate the degree of voice type similarity between the converted speeches and the target recordings. A VPR model[56][57][58][59][60] is trained using the target recordings, after which the voice type similarity between the converted speeches and the target recordings is computed using the pre-trained VPR model. The VPR value is a continuous numerical value ranging from 0 to 1. A higher VPR similarity score indicates a greater degree of similarity in voice type between the converted speeches and the target recordings. (4) **Automatic speech recognition (ASR).** Open source ASR is employed to transcribe spoken language into written text. Subsequently, the character error rate (CER) is computed by comparing the transcribed text with a reference text. A lower CER is indicative of superior performance.



(a)Generator loss



(b)Convergence iterators

Figure 2. Training efficiency of CycleGAN-VC3, MaskCycleGAN-VC and SVCGAN. Fig(a) shows the generator loss of CycleGAN-VC3, MaskCycleGAN-VC and SVCGAN. SVCGAN is convergence in 600k iterations while CycleGAN-VC3 and MaskCycleGAN-VC are not. Fig(b) shows the generator convergence iterators of CycleGAN-VC3, MaskCycleGAN-VC and SVCGAN. CycleGAN-VC3 converges in 1200k iterations, MaskCycleGAN-VC converges in 1000k iterations, and SVCGAN converges in 600k iterations. This means SVCGAN is more efficient.

Fig.2 illustrates the training efficiency of CycleGAN-VC3, MaskCycleGAN-VC and SVCGAN. Both CycleGAN-VC3 and MaskCycleGAN-VC models contain 2 generators and 4 discriminators, while the SVCGAN model contains 2 generators and 6 discriminators. Each model trains with same number of iteratiorns on generators and discriminators. The CycleGAN-VC3 model demonstrates convergence of the generator network after approximately 1200k iterations; the MaskCycleGAN-VC model demonstrates convergence of the generator network after approximately 1000k iterations; whereas the SVCGAN model achieves generator network convergence in 600k iterations. This implies that SVCGAN model exhibits around a 40% increase in efficiency compared to MaskCycleGAN-VC and a 50% increase in efficiency compared to CycleGAN-VC3 during the training phase.

TABLE I

Comparison of different hyper-parameters λ_{id} and λ_{sem} using (a) MCD, smaller value is better, (b) VPR, higher value is better, (c) CER of the speech texts, smaller value is better. Boldface indicates the best result.

Hyper-parameter	MCD	VPR	CER
$\lambda_{id} = 1, \lambda_{sem} = 1$	7.51	0.871	29.14%
$\lambda_{id} = 3, \lambda_{sem} = 1$	7.51	0.871	28.87%
$\lambda_{id} = 1, \lambda_{sem} = 3$	7.47	0.875	26.89%
$\lambda_{id} = 3, \lambda_{sem} = 3$	7.46	0.876	26.03%

$\lambda_{id} = 5, \lambda_{sem} = 3$	7.46	0.877	26.11%
$\lambda_{id} = 3, \lambda_{sem} = 5$	7.43	0.881	23.75%
$\lambda_{id} = 5, \lambda_{sem} = 5$	7.42	**0.883**	**23.39%**
$\lambda_{id} = 7, \lambda_{sem} = 5$	7.42	0.882	23.83%
$\lambda_{id} = 5, \lambda_{sem} = 7$	7.39	0.880	24.58%
$\lambda_{id} = 7, \lambda_{sem} = 7$	7.39	0.880	24.74%
$\lambda_{id} = 10, \lambda_{sem} = 7$	7.39	0.878	25.43%
$\lambda_{id} = 7, \lambda_{sem} = 10$	**7.36**	0.873	26.38%
$\lambda_{id} = 10, \lambda_{sem} = 10$	**7.36**	0.874	25.81%

TABLE II

Performance ablation of different improved objectives using (a) MCD, smaller value is better, (b) VPR, higher value is better, (c) CER of the speech texts, smaller value is better. SVCGAN uses all objectives. Boldface indicates the best result.

Method	MCD	VPR	CER
SVCGAN	**7.42**	**0.883**	**23.39%**
SVCGAN without -Log-Cosh Loss	7.44	0.881	24.04%
SVCGAN without -Semantic-Similarity Loss	7.50	0.878	27.20%
SVCGAN without -Third Adversarial Loss	7.45	0.881	24.16%
SVCGAN without -Log-Cosh Loss -Semantic-Similarity Loss	7.52	0.877	28.22%
SVCGAN without -Log-Cosh Loss -Third Adversarial Loss	7.47	0.880	25.31%
SVCGAN without -Semantic-Similarity Loss -Third Adversarial Loss	7.53	0.877	28.65%
SVCGAN without -Log-Cosh Loss -Semantic-Similarity	7.54	0.876	29.73%

Loss
-Third Adversarial Loss

TABLE III

Comparison of different methods using (a) MCD, smaller value is better, (b) VPR, higher value is better, (c) CER of the speech texts, smaller value is better. Boldface indicates the best result.

Method	MCD	VPR	CER
SVCGAN	**7.42**	**0.883**	**23.39%**
MaskCycleGAN-VC	7.54	0.876	29.73%
CycleGAN-VC3	7.71	0.861	32.06%
Original children's recordings			29.01%

Comparison among MCDs. In Table I, it is observed that the MCD algorithm yields the most favorable outcome when the hyper-parameter λ_{sem} is set to 10. The proximity between the converted speeches and the original kid recordings increases as the hyper-parameter λ_{sem} is augmented. In Table II, it is observed that all objectives result in a decrease in the MCD as compared to the MaskCycleGAN-VC. Notably, the objective incorporating semantic-similarity loss exhibits the most significant reduction in MCD. The rationale behind incorporating the semantic-similarity loss in the MaskCycleGAN-VC framework is to enhance the model's ability to capture and learn semantic information. Consequently, this leads to a significant reduction in the MCDs between the converted speeches and the original recordings of children. In Table III, our model demonstrates superior performance in MCD when compared to MaskCycleGAN-VC and CycleGAN-VC3.

Comparison among VPRs. In Table I, it is observed that the VPR achieves the highest performance when the hyper-parameters λ_{id} and λ_{sem} are both set to 5. In Table II, it is observed that all objectives result in an increase in the VPR when compared to the MaskCycleGAN-VC. Notably, the VPR associated with the semantic-similarity loss exhibits the highest rise. The outcomes beyond our initial expectations, as semantic-similarity loss would have made the conversion of voice types more challenging. One possible explanation is that the inclusion of a semantic-similarity loss function enhances the acquisition of semantic knowledge and improves the efficiency of model training. The model has the capacity to acquire the ability to preserve semantic information within a limited training duration, hence enhancing its proficiency in transforming kid voice types into adult voice types during subsequent training sessions. In Table III, our model demonstrates superior performance in VPR when compared to MaskCycleGAN-VC and CycleGAN-VC3.

Comparison among CERs. In Table I, it is observed that the optimal outcome for the CER is achieved when the hyper-parameters λ_{id} and λ_{sem} are both set to 5. In Table II, it is observed that all improved objectives result in a drop in the CER. Among these objectives, the semantic-similarity loss exhibits the most significant reduction in CER. The findings obtained from the MCD and VPR indicate that our model exhibits enhancements in two aspects: the acquisition of semantic information and the conversion of voice types. Consequently, our model achieves a lower CER when transcribing converted speech. In Table III, the CER of our model exhibits significant improvements of 8.67%, 6.34% and 5.62% compared to the CycleGAN-VC3 model, MaskCycleGAN-VC model and the original child recordings. This finding suggests that voice type conversion has the potential to enhance the accuracy rate of spoken language recognition for young children in their daily activities. However, the CER of the MaskCycleGAN-VC model and the CycleGAN-VC3 model is higher than the original children's recordings. The reason is because the semantic information of young children is easily destroyed during the voice conversion process, and these two models do not sufficiently protect it. In SVCGAN, we use log-cosh loss, semantic-similarity loss, and third adversarial loss to protect the semantic information of young children and better transform the voice type of young children into adult. Consequently, our model demonstrates superior performance in terms of CER when compared to MaskCycleGAN-VC and CycleGAN-VC3.

Subjective evaluation

A listening test was undertaken in order to examine the manner in which various listeners assessed the quality of the converted speeches. A total of 23 individuals were invited to assess the converted speeches produced by the CycleGAN-VC3, MaskCycleGAN-VC and SVCGAN models across several areas: (1) **Speech semantic information similarity.** In this assessment, listeners are required to evaluate two speech samples in terms of their semantic information resemblance to the original kid speech. Thirty sets of original kid speeches and transformed speeches from the CycleGAN-VC3, MaskCycleGAN-VC and SVCGAN models are randomly selected. (2) **Voice type similarity.** In this assessment, listeners are required to evaluate two speech samples based on their closeness to adult speech in terms of voice type. Thirty sets of adult speeches and converted speeches from the CycleGAN-VC3, MaskCycleGAN-VC and SVCGAN models are selected in a random manner. (3) **Speech naturalness and intelligibility.** In this assessment, listeners are required to evaluate two speech samples based on their naturalness and intelligibility. Thirty groups of the converted speeches from the CycleGAN-VC3, MaskCycleGAN-VC and SVCGAN models are selected in a random manner. In all assessments, the model of the speech samples is not disclosed to the listeners, and the resulting score is a continuous numerical value ranging from 0 to 5. In order to maintain fairness, the sequence of speech samples is randomized.

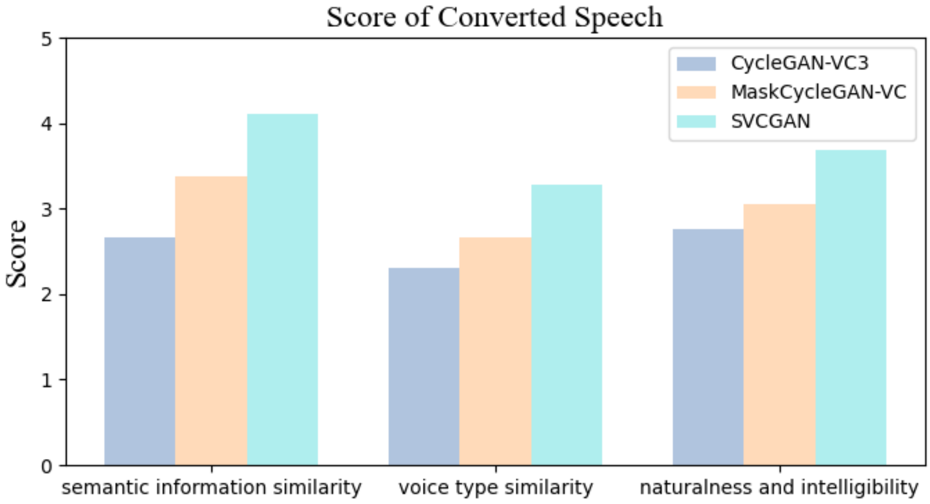


Figure 3. Speech semantic information similarity test, voice type similarity test, speech naturalness and intelligibility test of CycleGAN-VC3, MaskCycleGAN-VC and SVCGAN models.

Fig.3 illustrates the outcomes of the assessment conducted on speech semantic information similarity, voice type similarity, speech naturalness and intelligibility. The results of our evaluation indicate that the SVCGAN model outperformed the CycleGAN-VC3 and MaskCycleGAN-VC models in all terms of semantic information similarity, voice type similarity, speech naturalness and intelligibility. This means that the SVCGAN model exhibits superiority over the CycleGAN-VC3 and MaskCycleGAN-VC models.

CONCLUSIONS

We provide SVCGAN as a potential solution to enhance the quality of converted speech and the accuracy of speech recognition in the context of young children. Our model incorporates three novel techniques: (1) log-cosh loss, (2) semantic-similarity loss, and (3) third adversarial loss. These techniques aim to improve the effectiveness of training and enhance several aspects of converted speech, including semantic information similarity, voice type similarity, speech naturalness and intelligibility. The results demonstrate that our model outperforms CycleGAN-VC3 and MaskCycleGAN-VC models in terms of both objective and subjective assessment metrics, leading to enhanced converted speech quality and speech recognition accuracy in young children.

ACKNOWLEDGMENT

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, *et al.*, “The kaldi speech recognition toolkit,” in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF, IEEE Signal Processing Society, 2011.

[2] Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., “Pytorch: An imperative style, high-performance deep learning library,” Advances in neural information processing systems, vol. 32, 2019.

[3] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, “Tensorflow: a system for large-scale machine learning,” in *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pp. 265–283, 2016.

[4] S. Shahnawazuddin, A. Dey, and R. Sinha, “Pitch-adaptive front-end features for robust children’s asr.,” in *Interspeech*, pp. 3459–3463, 2016.

[5] C. Yadav, S. Shahnawazuddin, D. Govind, and G. Pradhan, “Spectral smoothing by variationalmode decomposition and its effect on noise and pitch robustness of asr system,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5629–5633, IEEE, 2018.

[6] H. K. Kathania, S. Shahnawazuddin, N. Adiga, and W. Ahmad, “Role of prosodic features on children’s speech recognition,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5519–5523, IEEE, 2018.

J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80

[7] P. G. Shivakumar and P. Georgiou, "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," *Computer speech & language*, vol. 63, p. 101077, 2020.

[8] S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 65–78, 2002.

[9] Potamianos and S. Narayanan, "Robust recognition of children's speech," IEEE Transactions on speech and audio processing, vol. 11, no. 6, pp. 603–616, 2003.

[10] P. Cosi, "On the development of matched and mismatched italian children's speech recognition systems," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[11] G. Yeung and A. Alwan, "On the difficulties of automatic speech recognition for kindergarten-aged children," *Interspeech 2018*, 2018.

[12] P. G. Shivakumar and S. Narayanan, "End-to-end neural systems for automatic children speech recognition: An empirical study," *Computer Speech & Language*, vol. 72, p. 101289, 2022.

[13] F. Claus, H. G. Rosales, R. Petrick, H.-U. Hain, and R. Hoffmann, "A survey about databases of children's speech.," in *INTERSPEECH*, pp. 2410–2414, 2013.

[14] Fainberg, P. Bell, M. Lincoln, and S. Renals, "Improving children's speech recognition through out-of-domain data augmentation.," in *Interspeech*, pp. 1598–1602, 2016.

[15] Du, X. Na, X. Liu, and H. Bu, "Aishell-2: Transforming mandarin asr research into industrial scale," *arXiv preprint arXiv:1808.10583*, 2018.

[16] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. Zeng, *et al.*, "Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6182–6186, IEEE, 2022.

[17] W. Ahmad, S. Shahnawazuddin, H. K. Kathania, G. Pradhan, and A. B. Samaddar, "Improving children's speech recognition through explicit pitch scaling based on iterative spectrogram inversion.," in *Interspeech*, pp. 2391–2395, 2017.

[18] S. Shahnawazuddin, N. Adiga, and H. K. Kathania, "Effect of prosody modification on children's asr," *IEEE Signal Processing Letters*, vol. 24, no. 11, pp. 1749–1753, 2017.

[19] H. K. Kathania, W. Ahmad, S. Shahnawazuddin, and A. B. Samaddar, "Explicit pitch mapping for improved children's speech recognition," *Circuits, Systems, and Signal Processing*, vol. 37, pp. 2021–2044, 2018.

[20] H. K. Kathania, S. Shahnawazuddin, W. Ahmad, N. Adiga, S. K. Jana, and A. B. Samaddar, "Improving children's speech recognition through time scale modification based speaking rate adaptation," in *2018 International Conference on Signal Processing and Communications (SPCOM)*, pp. 257–261, IEEE, 2018.

[21] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[22] J. E. Huber, E. T. Stathopoulos, G. M. Curione, T. A. Ash, and K. Johnson, "Formants of children, women, and men: The effects of vocal intensity variation," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1532–1542, 1999.

[23] S. Yildirim, S. Narayanan, D. Byrd, and S. Khurana, "Acoustic analysis of preschool children's speech," in *Proc. 15th ICPhS*, pp. 949–952, 2003.

J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80

[24] G. P. Scukanec, L. Petrosino, and K. Squibb, “Formant frequency characteristics of children, young adult, and aged female speakers,” *Perceptual and motor skills*, vol. 73, no. 1, pp. 203–208, 1991.

[25] H. K. Kathania, S. R. Kadiri, P. Alku, and M. Kurimo, “Study of formant modification for children asr,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7429–7433, IEEE, 2020.

[26] H. K. Kathania, S. R. Kadiri, P. Alku, and M. Kurimo, “A formant modification method for improved asr of children’s speech,” *Speech Communication*, vol. 136, pp. 98–106, 2022.

[27] [27] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, “Maskcycleganvc: Learning non-parallel voice conversion with filling in frames,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5919–5923, IEEE, 2021.

A. B. Kain, J.-P. Hosom, X. Niu, J. P. Van Santen, M. Fried-Oken, and J. Staehely, “Improving the intelligibility of dysarthric speech,” *Speech communication*, vol. 49, no. 9, pp. 743–759, 2007.

[28] Nakamura, T. Toda, H. Saruwatari, and K. Shikano, “Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech,” *Speech communication*, vol. 54, no. 1, pp. 134–146, 2012.

[29] Z. Inanoglu and S. Young, “Data-driven emotion conversion in spoken english,” *Speech Communication*, vol. 51, no. 3, pp. 268–283, 2009.

[30] T. Toda, M. Nakagiri, and K. Shikano, “Statistical voice conversion techniques for body-conducted unvoiced speech enhancement,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2505–2517, 2012.

[31] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, “Foreign accent conversion in computer assisted pronunciation training,” *Speech communication*, vol. 51, no. 10, pp. 920–932, 2009.

[32] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, “Sequence-tosequence voice conversion with similarity metric learned using generative adversarial networks.,” in *Interspeech*, vol. 2017, pp. 1283–1287, 2017.

[33] Sun, S. Kang, K. Li, and H. Meng, “Voice conversion using deep bidirectional long short-term memory based recurrent neural networks,” in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4869–4873, IEEE, 2015.

[34] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, “Sequenceto-sequence acoustic modeling for voice conversion,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 631–644, 2019.

[35] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, “Atts2s-vc: Sequenceto-sequence voice conversion with attention and context preservation mechanisms,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6805–6809, IEEE, 2019.

[36] H. Kameoka, K. Tanaka, D. Kwasny, T. Kaneko, and N. Hojo, “Convs2s-vc: Fully convolutional sequence-to-sequence voice conver-sion,” *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 28, pp. 1849–1863, 2020.

[37] Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in neural information processing systems*, vol. 27, 2014.

[38] 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,” *arXiv preprint arXiv:1704.00849*, 2017.

J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80

[39] T. Kaneko and H. Kameoka, “Parallel-data-free voice conversion using cycle-consistent adversarial networks,” *arXiv preprint arXiv:1711.11293*, 2017.

[40] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, “Voice conversion from non-parallel corpora using variational auto-encoder,” in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 1–6, IEEE, 2016.

[41] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, “Acvae-vc: Non-parallel voice conversion with auxiliary classifier variational autoencoder,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 9, pp. 1432–1443, 2019.

[42] T. Kaneko and H. Kameoka, “Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks,” in *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 2100–2104, IEEE, 2018.

[43] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, “Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks,” in *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 266–273, IEEE, 2018.

[44] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, “Stargan-vc2: Rethinking conditional methods for stargan-based voice conversion,” *arXiv preprint arXiv:1907.12279*, 2019.

[45] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, “Nonparallel voice conversion with augmented classifier star generative adversarial networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2982–2995, 2020.

[46] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, “Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6820–6824, IEEE, 2019.

[47] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, “Cyclegan-vc3: Examining and improving cyclegan-vcs for mel-spectrogram conversion,” *arXiv preprint arXiv:2010.11672*, 2020.

[48] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2223– 2232, 2017.

[49] T. Zhou, P. Krahenbuhl, M. Aubry, Q. Huang, and A. A. Efros, “Learning dense correspondence via 3d-guided cycle consistency,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 117–126, 2016.

[50] Y. Taigman, A. Polyak, and L. Wolf, “Unsupervised cross-domain image generation,” *arXiv preprint arXiv:1611.02200*, 2016.

[51] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, “Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline,” in *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*, pp. 1–5, IEEE, 2017.

[52] Kumar, R. Kumar, T. De Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. De Brebisson, Y. Bengio, and A. C. Courville, “Melgan: Generative adversarial networks for conditional waveform synthesis,” *Advances in neural information processing systems*, vol. 32, 2019.

J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80
J. Electrical Systems 20-3s (2024): 71-80

[53] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14,* pp. 702–716, Springer, 2016.

[54] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision,* pp. 2794–2802, 2017.

[55] M. C. Martinez, C. Spille, J. Roßbach, B. Kollmeier, and B. T. Meyer, "Prediction of speech intelligibility with dnn-based performance measures," Computer Speech & Language, vol. 74, p. 101329, 2022.

[56] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 652–662, 2019.

[57] Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 7132–7141, 2018.

[58] H. Wang, S. Zheng, Y. Chen, L. Cheng, and Q. Chen, "Cam++: A fast and efficient network for speaker verification using context-aware masking," *arXiv preprint arXiv:2303.00332,* 2023.

[59] Y. Chen, S. Zheng, H. Wang, L. Cheng, Q. Chen, and J. Qi, "An enhanced res2net with local and global feature fusion for speaker verification," *arXiv preprint arXiv:2305.12838,* 2023.