Towards Expressive Speech Conversion based on StarGANv2 *

☆ Shangyang Mou[1], Jinhui Chen[2], Ryoichi Takashima[1], Tetsuya Takiguchi[1]
[1] Kobe University, [2]Prefectural University of Hiroshima

1 Introduction

Expressive Speech Conversion is a technique that aims to convert the emotional state of the utterance from one to another while preserving the linguistic information and speaker identity. Emotional expression is essential in our daily life for conveying intentions and social attitude. Therefore, expressive speech conversion is an important objective in voice conversion field. In addition, it is difficult for people with dysarthria to express themselves clearly. The expressive speech conversion can be useful as their auxiliary expressions, allowing them to express as well as people without dysarthria. In this study, we explore the StarGANv2's ability of expressive speech conversion for dysarthric speech.

Recently, the research in voice conversion has attracted much attention in the field of speech processing. This technique can be applied in multiple domains, for example, emotion conversion [1], speech assistance [2], dysarthric speech conversion [3] and other such applications. Therefore, it has continued to motivate related studies each year. The expressive speech conversion, for example, has been an enabling technology for many applications including expressive text-to-speech [4], speech emotion recognition (SER) [5].

Expressive speech conversion is a kind of voice conversion where previous studies have focused on frame-based mapping of source and target spectral features, including using statistical methods [6] and deep learning methods, such as CycleGAN and Sequence-to-Sequence [7]. Recently, StarGAN models have attracted much interests in expressive voice conversion. It was originally proposed for image-to-image translations in multiple domains, and existing research shows that it is capable of speech conversion. In StarGANv2 [8], a variant of StarGAN proposed a scalable approach that can generate diverse images across multiple domains. For expressive speech conversion, a method based on StarGANv2 called StarGANv2-VC [1] has been shown to convert the speaker identity and emotion from one speaker to others. Although StarGANv2-VC has generalized to a variety of voice conversion tasks, such as any-to-many, cross-lingual and singing conversion, its capability for dysarthric emotional speech conversion has not been applied. In this paper, we investigate the application of StarGANv2-VC to the expressive conversion task for converting dysarthric speech and we focus on fundamental frequency (F0) transformation only. Our model allows multiple expressive speech conversion for dysarthric speaker, such as neutral to happy, sad and angry.

The remainder of this paper is organized as follows. Section 2 describes the framework of our method and details of each objective function. Section 3 describes the experiment. Finally, Section 4 concludes this paper.

2 Proposed Method

2.1 Model Architecture

StarGANv2-VC is a non-parallel and many-to-many voice conversion method based on the StarGANv2, which consists of a single discriminator and generator to generator multiple converted voice. We adopt the same architecture to expressive dysarthric speech conversion, treating each dysarthric speaker as an individual domain, and using a pre-trained joint direction and classification (JDC) F0 extraction network [9] to enable F0-consistent conversion. Besides, a style encoder is applied to extract the style code from the normal emotional speech. The overview of the method is shown in Figure 1.

Generator. The generator converts the input dysarthric speech into an emotional output speech which reflects the reference style. The generator gets the style code from a style encoder, and the fundamental frequency (F0) is provided by the convolution layers in the F0 Network.

F0 Network. In this framework the F0 Network adopts the pre-trained JDC Network, that consists of convolutional layers with residual connec-

* Towards Expressive Speech Conversion based on StarGANv2, 牟尚泱 [1], 陳金輝 [2], 高島遼一 [1], 滝口哲也 [1] ([1] 神戸大, [2] 広島県立大)

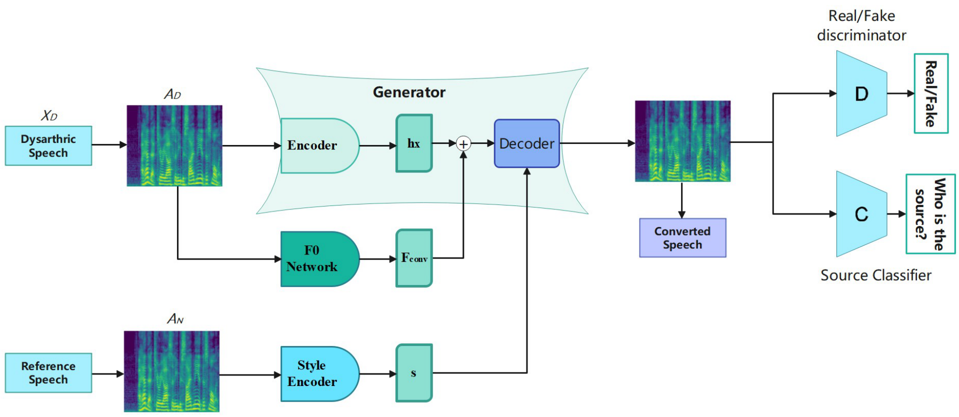

Fig. 1 Overview of the architecture, the input source speech X_D is neutral dysarthric speech, the reference speech is normal emotional speech, and the converted speech is the dysarthric speech with reference emotional expression. h_x, F_{conv} and s denote the latent feature of the source, the F0 pitch features and the emotional style code of the reference speaker, respectively. The latent feature h_x and F0 feature F_{conv} are concatenated by channel as the input of decoder, and the style code s is injected into the decoder by the adaptive instance normalization (AdaIN). Two classifiers form the discriminators that D determines whether a generated sample is real or fake and C determines who the source speaker is.

tions and bi-directional long short-term memory Bi-LSTM layers. It extracts the fundamental frequency from the input dysarthric speech, which is in the form of mel-spectrogram.

Style Encoder. Given a reference emotional mel-spectrogram, the encoder can extract the style code of multiple domains. Using different reference emotional speech, we can produce diverse emotional style code that allow the converted speech reflecting different emotions.

Discriminator. The discriminator consists of two classifiers, one to distinguish whether the generated speech is real or fake, the other to determine who is the source speaker of the dysarthric speech.

2.2 Objective Function

Our aim is to convert the neutral dysarthric speech into expressive speech, enabling the speakers with dysarthria to express their emotions like normal speaker. During training, we sample a neutral dysarthric speech $x_D \in X_D$ and an emotional style code $s \in S_N$ randomly, we train our model with the following loss function.

Adversarial objective. The generator is trained to convert the neutral dysarthric speech into emotional, here the input is mel-spectrogram A_D. At the same time a domain-specific style code s which extracted by style encoder is injected into the generator. Here we present the target domain $\widetilde{y} \in Y$ and source domain $y \in Y$ respectively, the generator learns to generate an output speech via an adversarial loss

$$L_{adv} = E_{X_D,y}[logD(X_D, y)] + E_{X_D,\widetilde{y},s}[log(1 - D(G(X_D, s), \widetilde{y}))] \tag{1}$$

where $D(., \widetilde{y})$ denotes the output of real/fake classifier in a specific-domain $\widetilde{y} \in Y$.

Style reconstruction objective. To better utilize the style code and ensure the generator to learn the style features fully, we employ a style reconstruction objective

$$L_{sty} = E_{X_D,\widetilde{y},s}[||s - S_N(G(X_D, s), \widetilde{y})||_1] \tag{2}$$

Adversarial domain classifier objective. For the other domain classifier C shown in Figure 1 we adopt the additional adversarial loss to make the classifier find the corresponding speaker as much as possible

$$L_{advcls} = E_{X_D,\widetilde{y},s}[CE(C(G(X_D, s)), \widetilde{y})] \tag{3}$$

where $CE(.)$ denotes the cross-entropy loss function.

Style diversification objective. To express diverse emotions, we apply the diversity sensitive loss on the generator which is maximized to enforce the generator to produce diverse emotions. Besides of maximizing the mean absolute error (MAE) of generated samples, the MAE of the F0 features between samples generated with different style code also be maximized here

$$L_{ds} = E_{X_D,s_1,s_2,\widetilde{y}}[||G(X_D,s_1) - G(X_D,s_2)||_1] + E_{X_D,s_1,s_2,\widetilde{y}}[||F_{conv}(G(X_D,s_1)) - F_{conv}(G(X_D,s_2))||_1] \tag{4}$$

where s_1, $s_2 \in S_N$ are two randomly sampled style code from domain $\widetilde{y} \in Y$ and $F_{conv}(.)$ comes from the F0 network.

F0 consistency objective. We use the F0-consistent loss to optimize the F0-consistent results, for an input mel-spectrogram A_D, $F_0(A_D)$ provides the absolute F0 value in Hertz for each frame of A_D

$$L_{f0} = E_{A_D,s}[||\hat{F}(A_D) - \hat{F}(G(A_D,s))] \tag{5}$$

where $\hat{F}(A_D) = \frac{F(A_D)}{||F(A_D)||_1}$, we normalize the absolute F0 value $F(A_D)$ by its temporal mean.

Speech consistency objective. In order to keep the linguistic content consistent after converted, we employ a speech consistency loss using convolutional features from a pre-trained joint CTC-attention VGG-BLSTM network [10]. The h_{asr} denotes linguistic features getting from the output of intermediate layer before the LSTM layers

$$L_{asr} = E_{X_D,s}[||h_{ars}(X_D) - h_{ars}(G(X_D,s))||_1] \tag{6}$$

Norm consistency objective. To preserve the speech/silence intervals of the converted emotional speech, we use the absolute column-sum norm for the input mel-spectrogram to obtain consistency objective

$$L_{norm} = E_{X_D,s}[\frac{1}{T}\sum_{t=1}^{T} | \, ||X_D,t|| - ||G(X_D,s),t|| \, |] \tag{7}$$

where $t \in 1,...,T$ is the frame index

Cycle consistency objective. We employ the cycle consistency loss to preserve the style domain characteristics

$$L_{cyc} = E_{X_D,y,\widetilde{y},s}[||X_D - G(G(X_D,s),\widetilde{s})||_1] \tag{8}$$

where $\widetilde{s} = S_N(X_D,y)$ is the extracted style code of the input in source domain $y \in Y$.

Full objective. The full objective function can be summarized as follows:

$$\begin{aligned} \min_{G,S,M} \quad & L_{adv} + \lambda_{advcls}L_{advcls} + \lambda_{sty}L_{sty} \\ & - \lambda_{ds}L_{ds} + \lambda_{f0}L_{f0} + \lambda_{asr}L_{asr} \\ & + \lambda_{norm}L_{norm} + \lambda_{cyc}L_{cyc} \end{aligned} \tag{9}$$

where $\lambda_{advcls}, \lambda_{sty}, \lambda_{ds}, \lambda_{f0}, \lambda_{asr}, \lambda_{norm}$ and λ_{cyc} are hyperparameters for each term.
The full discriminator objective is given by:

$$\min_{C,D} -L_{adv} + \lambda_{cls}L_{cls} \tag{10}$$

where λ_{cls} is the hyperparameter for source classifier loss L_{cls}.

3 Experiments

This section we describe the details of the training process and explain the subjective evaluation results. We do not conduct any objective experiment because there is no ground truth.

In the training process, for the source dysarthric speech, we use the UA-Speech [11] which is audio-visual isolated-word recordings of talkers with spastic dysarthria. We compare the converted speech of speakers with different degrees of dysarthria then select two speakers whose speech are almost intelligible. On the other hand, for the emotional reference speech, we use the public Emotional Speech Dataset (ESD) [12] and select 10 English speakers with three emotional states (angry, sad and happy).

We train the model for 150 epochs with a batch size of 5. The dysarthric speech of two speakers are randomly converted into angry, sad or happy, while the corresponding emotion style come from the 10 normal speakers. In the validation process, we conduct emotion conversion from dysarthric female speech to angry, sad and happy with the reference female emotional speech. Similarly, we perform the same with the other male speech.

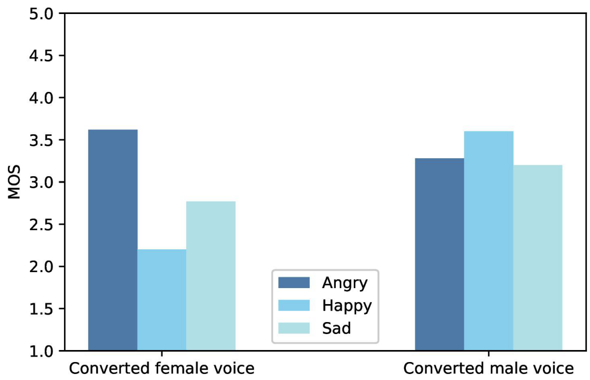

Fig. 2 Result of average on MOS

3.1 Evaluation

We conduct listening test to evaluate the emotion style and speech quality, and the results are shown in Figure 2. There are 6 participants in subjective experiment, and they are asked to listen 20 original dysarthric speeches, each speech corresponding to three converted emotional speech angry, sad and happy. Then they rate the results on a scale of 1 to 5, where 1 indicates that the converted speech has no emotion and 5 indicates that the converted speech is completely emotional.

The results show the mean opinion scores (MOS) and we can see that the MOS exceeds 3 for most of the target domains. The highest MOS is 3.62 and only two MOS are less than 3 when the female dysarthric speeches are converted to happy and sad. This means the overall quality of converted voice is good, and there is still room for improvement in some domains.

4 Conclusions

This study conducts expressive speech conversion for dysarthric speech using StarGANv2, i.e., StarGANv2-VC. We evaluate the capability of emotional speech conversion applying to dysarthric speech through a subjective listening test. The result shows that StarGANv2-VC can perform emotions conversion for dysarthric speech using the specified emotion style. However, for people with different degrees of dysarthria, we can observe significantly different conversion effects depending on the severity of the dysarthria.

Therefore, to enable the dysarthria speaker express emotionally in speech, we need to ensure the intelligibility of their speech. We believe that the promising dysarthric speech conversion can be used in emotional speech conversion. Simultaneously we can use other methods to process the semantic intelligibility, and we will take that as the future work.

References

[1] Yinghao Aaron Li, Ali Zare, Nima Mesgarani, "StarGANv2-VC: A Diverse, Unsupervised, Non-parallel Framework for Natural-Sounding Voice Conversion," in *INTERSPEECH*, pp. 1349-1353, 2021.

[2] Ryo Aihara, Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki, "A preliminary demonstration of exemplar-based voice conversion for articulation disorders using an individuality-preserving dictionary," in *EURASIP Journal on Audio, Speech, and Music Processing*, 2014.

[3] Xunquan Chen, Atsuki Oshiro, Jinhui Chen, Ryoichi Takashima, Tetsuya Takiguchi, "Phoneme-guided Dysarthric speech conversion With non-parallel data by joint training," in *Signal, Image and Video Processing*, 16, pp. 1641–1648, 2022.

[4] Rui Liu, Berrak Sisman, Guanglai Gao, Haizhou Li, "Expressive TTS training with frame and style reconstruction loss," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1806-1818, 2021.

[5] Babak Joze Abbaschian, Daniel Sierra-Sosa, Adel Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," in *Sensors*, 21(4):1249, 2021.

[6] Berrak Sisman, Mingyang Zhang, Haizhou Li, "Group sparse representation with wavenet vocoder adaptation for spectrum and prosody conversion," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 1085-1097, 2019.

[7] Kun Zhou, Berrak Sisman, Haizhou Li, "Limited data emotional voice conversion leveraging text-to-speech: Two-stage sequence-to-sequence training," in *INTERSPEECH*, pp. 811-815, 2021.

[8] Yunjey Choi, Youngjung Uh, Jaejun Yoo, Jung-Woo Ha, " StarGAN v2: Diverse Image Synthesis for Multiple Domains," in *CVPR*, pp. 8188-8197, 2020.

[9] Sangeun Kum, Juhan Nam, "Joint detection and classification of singing voice melody using convolutional recurrent neural networks," in *Appl. Sci.* 9(7):1324, 2019.

[10] Suyoun Kim, Takaaki Hori, Shinji Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *ICASSP*, pp. 4835-4839, 2017.

[11] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas Huang, Kenneth Watkin, Simone Frame, "Dysarthric speech database for universal access research," in *INTERSPEECH*, pp. 1741-1744, 2008.

[12] Kun Zhou, Berrak Sisman, Rui Liu, Haizhou Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *ICASSP*, pp. 920-924, 2021.