Dysarthric Speech Conversion by Learning Disentangled Representations with Non-parallel Data *

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

1 Introduction

Voice conversion (VC) aims to convert speaker-specific information in source speech while keeping linguistic information unchanged. There are many approaches for VC. In general, VC techniques are broadly grouped into parallel and non-parallel methods. Parallel voice conversion methods can learn the frame-wise mapping between source and target spectral features. Among these approaches, the Gaussian mixture model (GMM)-based mapping method [8] is the most widely used.

Non-parallel voice conversion techniques are certainly more attractive as parallel data is not easily available in practice. Non-parallel VC methods based on generative adversarial networks (GANs) have been studied [3]. The model using auto-encoder [5] is also attractive in that it can work for input speech of unknown speakers with disentangled linguistic and speaker representations.

However, most of the above-mentioned approaches cannot implement well in a dysarthric speech conversion task. Because conventional VC methods usually focus on non-dysarthric speech (normal speech). As for dysarthric speech, their phonetic structures are difficult to discriminate. Dysarthria is a motor speech disorder affecting multiple aspects of speech production. In this study, we focus on dysarthria resulting from cerebral palsy or amyotrophic lateral sclerosis, which are two of the prevalent causes of speech disability. Most people with dysarthria cannot communicate by sign language or writing, so there is a great need for voice conversion systems for them.

Inspired by [5, 11], in this paper, we present a non-parallel VC method for dysarthric speech with disentangled linguistic and speaker representations. In this method, a speaker encoder is constructed for extracting speaker information from acoustic features. An auxiliary reference encoder is adopted to enforce a phoneme recognition encoder to extract linguistic information effectively from dysarthric speech. A decoder is built for reconstruct dysarthric speech from the combination of linguistic and speaker representations.

2 Related work

Speaking assistance is one of the most important and meaningful tasks of VC. Previously, we have published several statistical VC approaches to improve the intelligibility of dysarthric speech. In [1], we proposed NMF-based VC method for articulation disorders. Some pairs of parallel utterances are needed in order to build parallel dictionaries. In [2], we proposed a dysarthric speech conversion method that utilizes phoneme-discriminative feature associated with a VC approach based on partial least square (PLS). The above-mentioned methods were effective for dysarthric speech conversion. However, the drawback is that these two methods require sufficient parallel data, which is difficult to collect. Especially for severe dysarthria, collecting such parallel data could be tedious. Recently, CycleGAN was shown that it is a promising method to transform dysarthric speech to normal speech without parallel data in [10]. Therefore, we compare our proposed method with the baseline approach based on CycleGAN.

3 Proposed method

In order to achieve dysarthric speaker conversion, we should be able to improve the intelligibility of source dysarthric speech and make the speaker identity conversion effectively. So it is important to estimate linguistic and speaker-related features because the phonetic structure of dysarthric speech fluctuates.

3.1 Model Architecture

Our proposed model, illustrated in Fig. 1, consists of four major modules: a reference encoder E_r, a

* Dysarthric Speech Conversion by Learning Disentangled Representations with Non-parallel Data, 陳訓泉 [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)

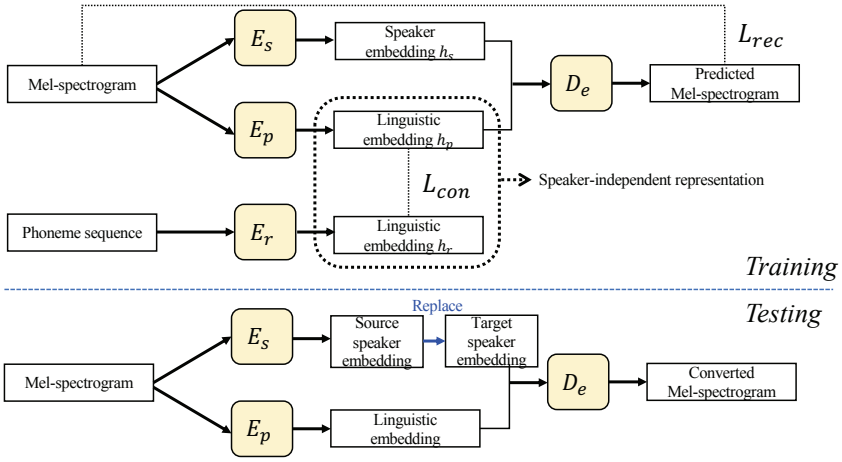

Fig. 1 Diagram of the proposed dysarthric speech conversion

phoneme recognition encoder E_p, a speaker encoder E_s, and a decoder D.

In the training stage, the speaker encoder E_s accepts the acoustic features as input, and learns the reasonable representations relating to speaker in the embedding space. E_s consists of two bi-directional LSTM layers, an average pooling and a fully connected layer. The phoneme recognition encoder E_p is used to recognize the phoneme sequence from input acoustic features. In order to align the acoustic and phoneme sequences automatically, E_p is composed of an bi-directional LSTM based encoder and a LSTM based decoder with an attention mechanism. The reference encoder E_r is adopted to enforce E_p to predict reasonable linguistic representation. E_r is composed of convolution layers, transforms phoneme sequences into linguistic embedding. Finally, the decoder D_e takes the concatenation of linguistic embeddings and speaker embeddings to predict mel-spectrogram features. D_e follows the structures of the Tacotron model [7]. The proposed model is pretrained on a normal speech dataset and then fine-tuned on the dysarthric corpus

When all the components are well trained, we use the system to perform dysarthric speech conversion on dysarthric corpus. In the conversion stage, the reference encoder E_r is not used. The trained speaker encoder E_s accepts the acoustic features of source dysarthric speech as input and generates the speaker-related representations. Then we can replace the speaker embeddings with target one. The source dysarthric speech is also fed into the trained E_p to have linguistic embedding extracted. Finally, the decoder generates the converted speech based on the linguistic information in the source dysarthric speech and the speaker information in the target speech.

3.2 Objective Function

Let M be the input mel-spectrograms. The training losses for the proposed VC are described as follows.

Contrastive loss: During training process, we will get two linguistic representations, h_p and h_r, from encoder E_p and E_r. Both h_p and h_r are extracted to represent the speaker-independent linguistc information. So we expect that h_p and h_r share the same linguistic space. In order to enforce E_p to learn a reasonable linguistic information, we constrain the linguistic embeddings from mel-spectrograms to be close to linguistic embeddings from phoneme sequence. Inspired by [11], we adopt the contrastive loss to increase the similarity between h_p and h_r.

$$L_{\text{con}} = \sum_{n=1,m=1}^{N,N} (I_{mn} d_{mn} + (1 - I_{mn}) \max(1 - d_{mn}, 0)) \quad (1)$$

Here, $h_p^{(m)}$ and $h_p^{(m)}$ denote mth and nth sequence of h_p and h_r.$I_{mn} = 0$ if $m = n$, otherwise $I_{mn} = 1$. And

$$d_{mn} = \left\| h_p^{(m)} - h_r^{(n)} \right\|_2$$

denotes the distance between h_p and h_r.

A reconstruction loss is also adopted to generate reasonable speech using disentangled representations.

$$L_{rec} = \left\| D_e\left(h_s, h_p\right) - M \right\|_2 \quad (2)$$

The full objective function can be summarized as follows:

$$L_{full} = L_{con} + \lambda_{rec} L_{rec} \tag{3}$$

where λ_{rec} is a trade-off parameter to control the weights.

4 Experiments

4.1 Experiments Conditions

The VCTK corpus [9] is used for pretraining. We carried out dysarthric speech conversion on the TORGO database [6], which contained recordings of 8 dysarthric speakers (3 females, 5 males) and 7 control/non-dysarthric speakers (3 females, 4 males) with 16kHZ sampling rate. The dysarthric speech resulting from either cerebral palsy or amyotrophic lateral sclerosis are provided. The severity of dysarthria is varied from severe to mild among different speakers.

The proposed method was evaluated in a dysarthric speaker conversion task. One male (M05) and one female (F03) with dysarthric speech were stored as the source speakers. The target speakers are chosen from the non-dysarthric speech group (we adopt MC01 and FC03). Fifty sentences are used for testing and the other sentences are used for training.

We adopted 80-dimensional mel-spectrograms as acoustic features and used WaveNet vocoder [4] to generate waveforms from the converted mel-spectrograms.

4.2 Objective Evaluations

Mel-cepstral distortion (MelCD) is commonly used as an objective metric to measure the global structural differences between converted speech and target speech. It is calculated as Eq. (4) by converted Mel-cepstral coefficients (MCCs) and target MCCs.

$$\text{MelCD} = (10/\log 10)\sqrt{2\sum_{d}^{24}(mc_d^{conv} - mc_d^{tar})^2} \tag{4}$$

Here, mc_d^{conv} and mc_d^{tar} denote the dth dimension of the converted MCCs and target MCCs, respectively. A smaller value indicates that the target and converted features are more similar.

Figure 2 shows the average MelCD values. M, F, MC and FC denote males with dysarthria, females with dysarthria, male controls without dysarthria, and female controls without dysarthria, respectively. M-MC, M-FC, F-MC and F-FC denote male-to-male conversion, male-to-female conversion, female-to-male conversion and female-to-female conversion, respectively. These results show that the proposed method achieves a better score than CycleGAN.

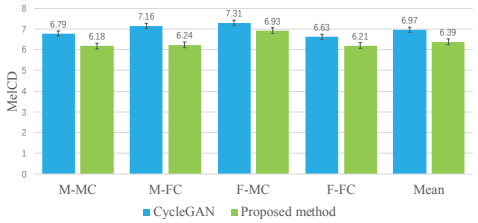

Fig. 2 Comparison of $MelCD$ [dB]

4.3 Subjective Evaluations

We note that a lower value indicates better performance in objective evaluation. However, the value of MelCD is not always correlated with human perception. Therefore, the subjective evaluation was conducted on "intelligibility" and "speaker similarity" for the task of dysarthric speech conversion. For the subjective evaluation, 20 sentences for each conversion pair were evaluated by 9 listeners.

For the evaluation of speech quality, we performed a Mean Opinion Score (MOS) test. The opinion score was set to a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). For the similarity evaluation, the XAB test was carried out. For both tests, a higher value indicates a better result.

The results of MOS are shown in Figure 3. It shows that the proposed method achieved a better score than CycleGAN. The MOS indicates that the proposed method works relatively well in terms of sound quality for every speaker pair.

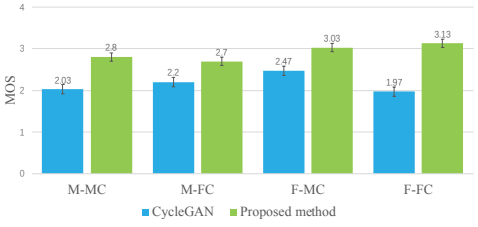

Fig. 3 MOS for intelligibility with 95% confidence intervals.

Figure 4 shows the results of XAB tests on speaker similarity. The difference between the two methods

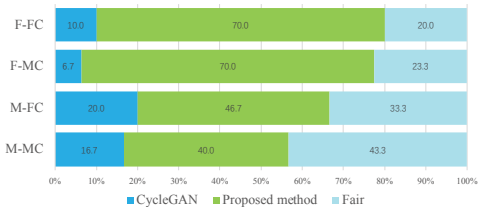

Fig. 4 Average preference score (%) on speaker similarity.

is significant. These results did not contradict the results of objective evaluation and show the effectiveness of our proposed method.

5 Conclusions

In this paper, we present a non-parallel VC method for dysarthric speech with disentangled linguistic and speaker representations. The model key feature is that we utilize a phoneme-guided reference encoder to enforce the phoneme recognition encoder to learn reasonable linguistic representation of dysarthric speech. Experimental results show that, in the task of dysarthric speaker conversion, our proposed method makes it possible to obtain higher intelligibility and better similarity compared to baseline VC. In this study, we mainly consider dysarthric speech with low intelligibility. So in future experiments, we will increase the number of subjects and further examine the effectiveness of our method according to different severities of dysarthric speech.

References

[1] Ryo Aihara *et al.*, "A preliminary demonstration of exemplar-based voice conversion for articulation disorders using an individuality-preserving dictionary". EURASIP Journal on Audio, Speech, and Music Processing, 2014(1):1–10, 2014.

[2] Ryo Aihara *et al.*, "Phoneme-discriminative features for dysarthric speech conversion," in Interspeech, pages 3374–3378, 2017.

[3] Takuhiro Kaneko and Hirokazu Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," ArXiv, abs/1711.11293, 2017.

[4] Oord *et al.*, "Wavenet: A generative model for raw audio," arXiv preprint arXiv:1609.03499, 2016.

[5] Kaizhi Qian *et al.*, "Autovc: Zero-shot voice style transfer with only autoencoder loss," arXiv preprint arXiv:1905.05879, 2019.

[6] Frank Rudzicz *et al.*, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," Language Resources and Evaluation, 46(4):523–541, 2012.

[7] Jonathan Shen *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4779–4783. IEEE, 2018.

[8] Y. Stylianou *et al.*, "Continuous probabilistic transform for voice conversion," IEEE Transactions on Speech and Audio Processing, 6(2):131–142, 1998.

[9] Christophe Veaux *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2019.

[10] Seung Hee Yang and Minhwa Chung, "Improving dysarthric speech intelligibility using cycle-consistent adversarial training," arXiv preprint arXiv:2001.04260, 2020.

[11] Hang Zhou *et al.*, "Talking face generation by adversarially disentangled audio-visual representation," In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 9299–9306, 2019.