Learning Explicit Prosody Models and Deep Speaker Embeddings for Atypical Voice Conversion

Disong Wang[1], *Songxiang Liu*[1], *Lifa Sun*[2], *Xixin Wu*[3], *Xunying Liu*[1], *Helen Meng*[1]

[1]Human-Computer Communications Laboratory
The Chinese University of Hong Kong, Hong Kong SAR, China
[2]SpeechX Limited, Shenzhen, China
[3]Department of Engineering, University of Cambridge, UK

{dswang, sxliu, xyliu, hmmeng}@se.cuhk.edu.hk, lfsun@speechx.cn, xw369@cam.ac.uk

Abstract

Though significant progress has been made for the voice conversion (VC) of typical speech, VC for atypical speech, e.g., dysarthric and second-language (L2) speech, remains a challenge, since it involves correcting for atypical prosody while maintaining speaker identity. To address this issue, we propose a VC system with explicit prosodic modelling and deep speaker embedding (DSE) learning. First, a speech-encoder strives to extract robust phoneme embeddings from atypical speech. Second, a prosody corrector takes in phoneme embeddings to infer typical phoneme duration and pitch values. Third, a conversion model takes phoneme embeddings and typical prosody features as inputs to generate the converted speech, conditioned on the target DSE that is learned via speaker encoder or speaker adaptation. Extensive experiments demonstrate that speaker adaptation can achieve higher speaker similarity, and the speaker encoder based conversion model can greatly reduce dysarthric and non-native pronunciation patterns with improved speech intelligibility. A comparison of speech recognition results between the original dysarthric speech and converted speech show that absolute reduction of 47.6% character error rate (CER) and 29.3% word error rate (WER) can be achieved.

Index Terms: dysarthric speech reconstruction, accent conversion, prosodic modelling, speaker encoder, speaker adaptation

1. Introduction

Voice conversion (VC) is a technique for converting non-linguistic and para-linguistic information, such as speaker identity [1], prosody [2] and accent [3], with potential applications in assistive speech technologies and language acquisition technologies [4, 5]. This work aims to apply VC techniques to convert atypical speech to a typical form. Specifically, we consider two types of atypical speech [6] – dysarthric speech and second-language (L2) speech. Dysarthric speech results from neuromotor disorders [7] that cause disturbances in muscular control during articulation. L2 speech is spoken by L2 learners with non-native accents [8]. Both dysarthric and L2 speech exhibits atypical prosody, imprecise articulation and reduced intelligibility. These may engender substantial communication difficulties for dysarthric patients and hinder the pronunciation clarity of L2 learners.

To enhance the quality of the atypical speech, our previous work [9, 10] presented an end-to-end VC (E2E-VC) method, where a speech-encoder is used to extract linguistic representations, e.g., phoneme embeddings, from the atypical speech, and a text-to-speech (TTS) decoder with attention maps phoneme embeddings to typical speech features. The speaker identity of the converted speech is controlled by the target speaker embedding produced by a speaker encoder [10]. Though high-fidelity speech can be generated, the prosody, speaker similarity and speech intelligibility require further improvement.

In this paper, we propose an improved VC system, where the previous TTS-decoder with attention is broken into a prosody corrector and a conversion model. The prosody corrector contains phoneme duration and pitch predictors that are introduced to explicitly model the prosody for predicting typical phoneme duration and pitch features. The conversion model maps pitch and phoneme embeddings expanded by the duration to mel-spectrograms, conditioned on the target deep speaker embedding (DSE). To obtain effective DSE that captures speaker characteristics, two different approaches are investigated in our work: (1) Speaker encoder, where a speaker classifier trained independently is adopted to extract DSE from the reference target speech; (2) Speaker adaptation, where the DSE is jointly learned and fine-tuned with a pre-trained multi-speaker conversion model by using the target speech. We assume that the DSE obtained using the two approaches contains no prosody cues, so prosody and speaker identity are controlled by individual conditions, i.e., the prosody is controlled by phoneme duration and pitch, and speaker identity is controlled by the DSE. As a result, with the predicted typical prosody features and the effective DSE as conditions, the converted speech has typical pronunciation patterns with high speaker similarity and improved speech intelligibility.

The main advantages of the proposed approach include: (1) Explicit prosody correction to reduce dysarthric or non-native pronunciation patterns; (2) Improvements over previous methods [9, 10] in generating speech with enhanced speaker similarity, naturalness and intelligibility; (3) Potential extensibility to other atypical voice conversion and enhancement tasks.

2. Related work

Dysarthric speech reconstruction (DSR) aims to convert dysarthric speech to be near-normal speech with higher intelligibility and naturalness. Various VC techniques have been applied for DSR. Rule-based VC modifies the temporal or frequency characteristics of speech according to specific rules [11]. Statistical VC builds a mapping function between the acoustic features of dysarthric and normal speech [9, 12–14]. Significant progress has been achieved, but the converted speech has low speaker similarity.

Accent conversion (AC) aims to convert the non-native L2 accented speech to become near-native speech. [15] proposed a GMM based VC by using vocal tract length normalization and linguistic content similarity matching. [16,17] utilized phonetic posteriorgrams (PPG) of the native speaker to generate target acoustic features. Although the non-native accent can be reduced, these methods require native reference utterances that may not be readily available. E2E-VC [10] can effectively solve this issue, but speaker similarity needs to be improved as well.

We intend to convert dysarthric and L2 speech respectively to become near-normal and near-native speech with typical prosody, high speaker similarity and improved intelligibility. We reference multi-speaker TTS [18,19] that uses prosody features for speech synthesis, and DSE obtained via speaker encoder or speaker adaptation to control the speaker identity. Inspired by Deep Voice 2 [18], we introduce predictors of phoneme duration and pitch to attain typical values in order to generate the speech with typical (i.e., normal or native) prosody characteristics.

3. Baseline method

In this paper, we adopt the previously proposed E2E-VC for DSR [9] and AC [10] as the baseline method. E2E-VC is composed of three components: (1) A sequence-to-sequence (seq2seq) based TTS model, e.g., Tacotron [20], is first trained with transcribed *typical* speech. The TTS-decoder with attention implicitly models the prosody, e.g., phoneme duration and pitch, which are inflexible to control during inference. (2) Given the transcribed *atypical* speech, a speech-encoder is trained to produce similar linguistic representations with those produced by the TTS-encoder. (3) By concatenating the speech-encoder and TTS-decoder with attention, an E2E-VC is formed to convert atypical speech to its typical version. Note that speaker similarity issue was not considered in the DSR work [9], so we extend the E2E-VC based DSR with the speaker encoder introduced in AC [10] to preserve speaker identity.

4. Proposed method

This section elaborates on the proposed VC approach with explicit prosodic modelling and DSE learning. The main differences from the baseline E2E-VC approach lie in two aspects: (1) Prosody is modelled in an explicit manner, so the prosody of the converted speech can be effectively controlled and corrected; (2) Speaker adaptation is proposed to obtain more effective DSE that is strongly related with speaker characteristics, leading to higher speaker similarity. As shown in Figure 1, the whole VC system consists of three key modules, i.e., speech-encoder, prosody corrector and conversion model.

4.1. Speech-encoder for phoneme embeddings extraction

To preserve the linguistic content of original atypical speech, a speech-encoder is used to extract robust linguistic representations. Following [9, 10], the speech-encoder adopts a seq2seq network to predict phoneme sequence. The speech-encoder is first pre-trained on large-scale typical speech data, then fine-tuned on the atypical speech of the dysarthric or L2 speaker s_k to improve phoneme prediction accuracy. The pre-trained and fine-tuned speech-encoders are denoted as Φ_p and Φ_{s_k} respectively. We adopt the speech-encoder outputs that denote the phoneme probability distribution as the phoneme embeddings.

4.2. Prosody corrector for explicit prosodic modelling

As atypical speech has atypical prosody, e.g., phoneme duration and pitch values, we propose explicit prosodic modelling by designing a prosody corrector to amend the atypical prosody to its typical version. As shown in Figure 1, the prosody corrector contains the phoneme duration and pitch predictors, where the pitch can be described by fundamental frequency (F_0). Both duration and F_0 predictors are trained with L1 loss by using typical speech of a single speaker: (1) For duration prediction, the inputs are phoneme embeddings extracted by the speech-encoder Φ_p using the teacher-force mode. The targets are ground-truth phoneme durations, which are obtained from pairs of text and audio by forced alignment with Montreal Forced Aligner [21]. (2) For F_0 prediction, the expanded phoneme embeddings **p** by using the ground-truth phoneme duration are used as the inputs, and the targets are the ground-truth F_0, denoted by **v**, which has the same length of **p**. When the duration and F_0 predictors are well-trained, the prosody corrector is expected to infer typical phoneme duration and F_0 values that are used to replace their abnormal counterparts for typical speech generation.

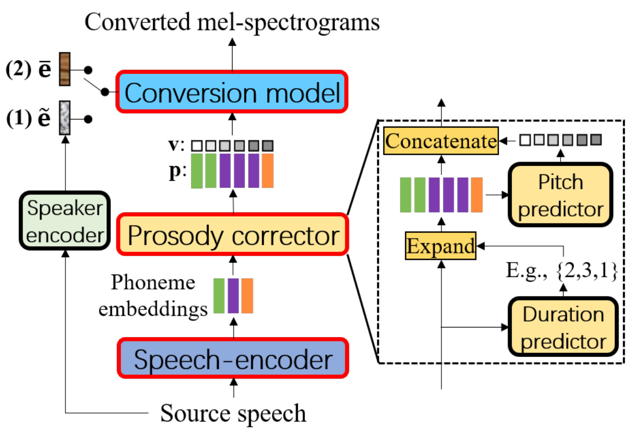


Figure 1: *Diagram of the proposed VC system: (1) DSE $\tilde{\mathbf{e}}$ is extracted from the speaker encoder, which corresponds to Enc-CM; (2) DSE $\bar{\mathbf{e}}$ is obtained by joint learning with the conversion model, which corresponds to Ada-CM.*

4.3. Conversion model for speech generation

As shown in Figure 1, we adopt a conversion model with function f and parameters $\mathbf{W}$ to generate mel-spectrograms $f(\mathbf{p},\mathbf{v};\mathbf{W},\mathbf{e})$, where the spoken content and duration are both controlled by the expanded phoneme embeddings **p**, the pitch and speaker identity are separately controlled by F_0 **v** and DSE **e**. **e** is repeated and concatenated with **p** and **v** for generation. Given the typical speech of a set of speakers S and atypical speech of a dysarthric or L2 speaker s_k, let T_{s_i} and T_{s_k} denote the set of mel-spectrogram features for speaker s_i ($s_i \sim S$) and speaker s_k, respectively. Two DSE learning approaches are investigated and incorporated into the conversion model, i.e., speaker encoder based conversion model (Enc-CM) and speaker adaptation based conversion model (Ada-CM). For clarity, we denote the DSE used in Enc-CM and Ada-CM as $\tilde{\mathbf{e}}$ and $\bar{\mathbf{e}}$ respectively.

4.3.1. Speaker encoder based conversion model

The speaker encoder is a neural network for speaker verification and produces a fixed-dimensional DSE from acoustic feature frames of a speech utterance with variable length. It is trained to optimize a generalized end-to-end (GE2E) loss for DSE learning [22], so that the DSEs extracted from utterances of the same speaker and different speakers have high and low similarity, respectively. The DSE $\tilde{\mathbf{e}}_{s_i}$ derived from the speaker encoder is expected to capture speaker characteristics of s_i. By using typical speech data, Enc-CM is trained to minimize a loss

L (e.g., L1 loss) measuring the distance between the predicted and ground-truth mel-spectrograms:

$$\widetilde{\mathbf{W}}_{SE} = \underset{\mathbf{W}}{argmin} E_{s_i \sim S, \mathbf{a}_{i,j} \sim T_{s_i}} \{L(f(\mathbf{p}_{i,j}, \mathbf{v}_{i,j}; \mathbf{W}, \widetilde{\mathbf{e}}_{s_i}), \mathbf{a}_{i,j})\} \quad (1)$$

where $\mathbf{p}_{i,j}$ are phoneme embeddings extracted by speech-encoder Φ_p and expanded by ground-truth duration, $\mathbf{v}_{i,j}$ and $\mathbf{a}_{i,j}$ are ground-truth F_0 and mel-spectrograms for speaker s_i ($s_i \sim S$), respectively.

At the conversion phase, the atypical speech of the speaker s_k is used as the input of the speaker encoder and the fine-tuned speech-encoder Φ_{s_k} to extract DSE $\widetilde{\mathbf{e}}_{s_k}$ and phoneme embeddings, respectively. The phoneme embeddings are used as the inputs of the prosody corrector to obtain the expanded phoneme embeddings $\hat{\mathbf{p}}$ with typical duration and F_0 $\hat{\mathbf{v}}$. Finally, the system generates converted mel-spectrograms as $f(\hat{\mathbf{p}}, \hat{\mathbf{v}}; \widetilde{\mathbf{W}}_{SE}, \widetilde{\mathbf{e}}_{s_k})$.

4.3.2. *Speaker adaptation based conversion model*

Instead of obtaining the speaker representations from an external network, DSE can be jointly learned with the conversion model. The joint learning enables the DSE to directly capture the speaking characteristics related with speech generation, leading to higher speaker similarity. Specifically, Ada-CM involves two-stage training: First, the conversion model is pre-trained with typical speech data, where the DSE $\mathbf{e}_{s_i}$ for each speaker s_i is randomly initialized and jointly trained with $\mathbf{W}$:

$$\hat{\mathbf{W}}_{SA}, \{\hat{\mathbf{e}}_{s_i}\} = \underset{\mathbf{W}, \{\mathbf{e}_{s_i}\}}{argmin} E_{s_i \sim S, \mathbf{a}_{i,j} \sim T_{s_i}} \{L(f(\mathbf{p}_{i,j}, \mathbf{v}_{i,j}; \mathbf{W}, \mathbf{e}_{s_i}), \mathbf{a}_{i,j})\} \quad (2)$$

Second, with the speech data of multiple speakers for training, the conversion model $\hat{\mathbf{W}}_{SA}$ has good generalization capacity and can be fine-tuned well to unseen speakers for DSE learning. Therefore, for the dysarthric or L2 speaker s_k with the expanded phoneme embeddings $\mathbf{p}_{k,j}$ and F_0 $\mathbf{v}_{k,j}$, speaker adaptation is performed as:

$$\overline{\mathbf{W}}_{SA}, \bar{\mathbf{e}}_{s_k} = \underset{\mathbf{W}, \mathbf{e}_{s_k}}{argmin} E_{\mathbf{a}_{k,j} \sim T_{s_k}} \{L(f(\mathbf{p}_{k,j}, \mathbf{v}_{k,j}; \mathbf{W}, \mathbf{e}_{s_k}), \mathbf{a}_{k,j})\} \quad (3)$$

where $\mathbf{W}$ is initialized by $\hat{\mathbf{W}}_{SA}$ and DSE $\mathbf{e}_{s_k}$ is also randomly initialized. After adaptation, target speaker characteristics that are beneficial for speech generation are encoded into $\bar{\mathbf{e}}_{s_k}$.

Similar with Enc-CM, at the inference phase, we can use the adapted conversion model $\overline{\mathbf{W}}_{SA}$ to generate the converted mel-spectrograms as $f(\hat{\mathbf{p}}, \hat{\mathbf{v}}; \overline{\mathbf{W}}_{SA}, \bar{\mathbf{e}}_{s_k})$ with high speaker similarity achieved by $\bar{\mathbf{e}}_{s_k}$, and typical prosody controlled by predicted typical phoneme duration and F_0.

5. Experiments

5.1. Experimental settings

Experiments are conducted on the LibriSpeech [23], VCTK [24], LJSpeech [25], UASpeech [26] and L2-ARCTIC [27] datasets. Parallel WaveGAN (PWG) [28] is adopted as the vocoder to synthesize the waveform from the converted mel-spectrograms. We use 960h training data of LibriSpeech for pre-training the speech-encoder Φ_p, 105 native speakers of VCTK for the training of PWG, and the training of Enc-CM and Ada-CM to obtain $\widetilde{\mathbf{W}}_{SE}$ and $\hat{\mathbf{W}}_{SA}$, respectively. The typical speech of single female speaker from LJSpeech is used for training duration and F_0 predictors. For atypical speech, we select speaker M05 of UASpeech and speaker LXC of L2-ARCTIC for experiments. M05 has moderate-severe dysarthria with the speech having middle intelligibility. Following [9], we use the speech of blocks 1 and 3 for speech-encoder and Ada-CM fine-tuning, and the speech of block 2 for testing. As the audio of M05 has strong background noise which degrades the speaker adaptation performance, we adopt log-MMSE speech enhancement algorithm [29] to pre-process the audio. LXC speaker is a non-native English speaker with Mandarin accent and has 1131 recorded utterances, which are randomly divided into 1000/66/65 for training/validation/testing, where training and validation data are used for fine-tuning the speech-encoder and Ada-CM. The speech is sampled or resampled to 16kHz, and all speech features are calculated with 25ms Hanning window, 10ms frame shift and 400-point fast Fourier transform.

The speech-encoder has a similar architecture as in [9], including a 6-layer VGG extractor and a 5-layer BLSTM with 512 units per direction in the encoder, 512-dimensional location-aware attention and 2-layer LSTM with 1024 units in the decoder. The inputs to the speech-encoder are 40-band mel-spectrograms appended with delta and delta-delta features. Adadelta optimizer [30] with learning rate of 1 and batch size 8 is applied for the pre-training and fine-tuning of the speech-encoder with 1M and 2k steps, respectively. Duration and F_0 predictors adopt the same structure, which consists of a 3-layer BGRU with 256 units per direction, 3 convolution layers with kernel size of 5, 9 and 19 respectively, and a 1-dimensional fully-connected (FC) layer to predict the duration or F_0 value. Both predictors are trained by the Adam optimizer [31] with a learning rate of 0.001, batch size of 16 and 30k steps. The settings and training of speaker encoder is same as [10], and DSE is set to 256-dimensional for both Enc-CM and Ada-CM. The conversion model is a frame-to-frame network composed of two 512-dimensional FC layers, 4-layer BLSTM with 512 units per direction and one 80-dimensional FC layer to predict mel-spectrograms. Both Enc-CM training and Ada-CM pre-training use the learning rate of 0.001 and batch size of 16 with 50k steps, and Ada-CM fine-tuning takes 3k steps. Readers are encouraged to listen to our audio samples[1].

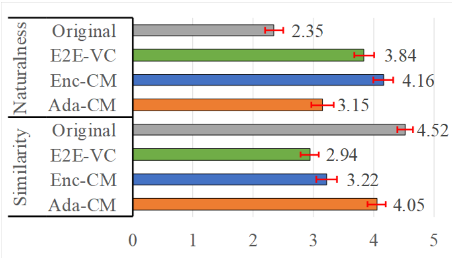
We compare the Enc-CM and Ada-CM with our previously proposed E2E-VC for DSR [9] and AC [10], where the original settings are adopted. To evaluate the performance of all methods, 20 listeners are invited to give subjective evaluations, including mean opinion score (MOS) tests (1-bad, 2-poor, 3-fair, 4-good, 5-excellent) to evaluate speech naturalness and speaker similarity, AB preference tests to evaluate the impact of phoneme duration and F_0, and objective evaluation of speech intelligibility based on a speech recognition model.

5.2. Experimental results

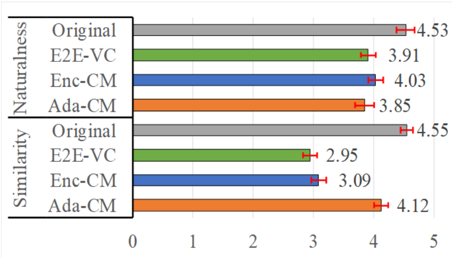
5.2.1. *Speech naturalness and speaker similarity comparison*

Figure 2 shows the MOS results for speech naturalness and speaker similarity, where 'Original' denotes the original dysarthric or L2 speech. We randomly select 15 testing utterances of M05 or LXC for evaluation. For DSR experiments as shown in Figure 2(a), we observe that compared with the original dysarthric speech, the converted speech by all methods achieves improvements in naturalness. The proposed Enc-CM achieves higher naturalness than E2E-VC, indicating that the effectiveness of the proposed prosody corrector to generate the speech with stable and accurate prosody. Ada-CM achieves lower naturalness than E2E-VC, partially due to the speech enhancement algorithm degrading the quality of M05 audio for speaker adaptation. Besides, speaker adaptation also inevitably incorporates the abnormal speaking characteristics of the dysarthric speaker into the converted speech, such as

[1] Audio samples: https://wendison.github.io/VC-DSR-AC-demo/



(a) DSR experiments



(b) AC experiments

Figure 2: *Comparison results of MOS with 95% confidence intervals for speech naturalness and speaker similarity*

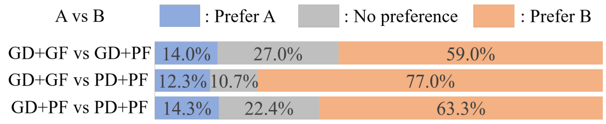
atypical prosody and articulation, which partially contribute to the speaker characteristics. As a result, Ada-CM can achieve highest speaker similarity, followed by Enc-CM and E2E-VC. For AC experiments as shown in Figure 2(b), all methods achieve high and similar speech naturalness, while Ada-CM also achieves the highest speaker similarity. This verifies the effectiveness of explicit prosodic modelling and DSE learned by speaker adaptation for achieving high speaker similarity.

5.2.2. Impact of phoneme duration and F_0

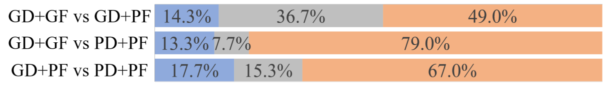
We investigate the impact of the phoneme duration and F_0 on the normality and accentedness of the converted speech. We explore three combinations of phoneme duration and F_0 used by Enc-CM to generate the speech: (1) Ground-truth Duration and Ground-truth F_0 (GD+GF); (2) Ground-truth Duration and Predicted F_0 (GD+PF); (3) Predicted Duration and Predicted F_0 (PD+PF). The AB preference test is conducted, and results are shown in Figure 3. From the comparison 'GD+GF vs GD+PF', we can see that the predicted typical F_0 facilitates the conversion model to generate near-normal or near-native speech. From the comparison 'GD+PF vs PD+PF', we observe that using the predicted typical duration can further improve the quality of the converted speech, this shows that both phoneme duration and F_0 affect speech normality or the degree of accentedness, and the proposed prosody corrector is helpful for attaining typical phoneme duration and F_0 values, which are beneficial for generating speech with normal or native prosody characteristics.

5.2.3. Speech intelligibility comparison

To show the effectiveness of proposed methods to improve the intelligibility of atypical speech, a publicly released automatic speech recognition model, i.e., Jasper [32], is used to test the character error rate (CER) and word error rate (WER) with greedy decoding. The results are illustrated in Table 1, we also report the results for 'Original (Mel+PWG)' that uses the original mel-spectrograms to synthesize the waveform by using PWG vocoder. We can see that 'Original (Mel+PWG)' is inferior to 'Original', which indicates that the PWG vocoder tends to degrade the speech quality. For DSR experiments, we can



(a) DSR experiments



(b) AC experiments

Figure 3: *AB preference test results for different combinations of phoneme duration and F_0.*

Table 1: *Comparisons based on CER (%) and WER (%).*

Methods	DSR experiments		AC experiments	
	CER	WER	CER	WER
Original	90.2	91.0	17.7	35.5
Original (Mel+PWG)	94.3	95.3	22.4	40.8
E2E-VC	50.6	69.8	22.7	41.3
Enc-CM	**42.6**	**61.7**	**15.3**	**31.1**
Ada-CM	56.5	80.5	21.5	40.2

observe that CER and WER of the original dysarthric speech can be significantly reduced by the proposed methods, where Enc-CM performs the best and achieves 47.6% and 29.3% absolute reduction for CER and WER respectively. As the original dysarthric speech used in Ada-CM to perform speaker adaptation contains strong background noise, even though log-MMSE is used for denoising, the pre-processed audio still contains artificial noise that hurts Ada-CM performance, thus smaller CER and WER reduction is achieved for Ada-CM. For AC experiments, the proposed Enc-CM can still achieve 2.4% CER and 4.4% WER reduction compared with Original speech, the baseline E2E-VC and proposed Ada-CM have no improvements over Original speech while achieve similar CER and WER with 'Original (Mel+PWG)', adopting a more powerful vocoder is expected to enhance the speech intelligibility.

6. Conclusions

This paper presents a VC system for converting atypical speech to typical speech, by explicit prosodic modelling and DSE learning. prosodic modelling is proposed to leverage phoneme duration and F_0 predictors to obtain typical values for prosody correction, while speaker encoder and speaker adaptation approaches are separately proposed to obtain effective DSE to capture speaker characteristics. DSR and AC experiments show that proposed methods can achieve reduction of dysarthric and non-native speaking characteristics, where significant intelligibility improvements can be achieved for dysarthric speech. Enc-CM outperforms previously proposed E2E-VC, and Ada-CM achieves the highest speaker similarity. However, for Ada-CM, atypical pronunciation patterns are also incorporated into the converted speech after speaker adaptation. Explicitly modelling more para-linguistic information may be helpful to mitigate this problem. In addition, using better speech denoising algorithms or cleaner audio data is expected to further improve the performance of Ada-CM, this will be studied in the future.

7. Acknowledgements

This research is partially supported by a grant from the HKSARG Research Grants Council General Research Fund (Project Reference No. 14208817).

8. References

[1] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[2] D. Rentzos, S. Vaseghi, E. Turajlic, Q. Yan, and C.-H. Ho, "Transformation of speaker characteristics for voice conversion," in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*. IEEE, 2003, pp. 706–711.

[3] K. Oyamada, H. Kameoka, T. Kaneko, H. Ando, K. Hiramatsu, and K. Kashino, "Non-native speech conversion with consistency-aware recursive network and generative adversarial network," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 182–188.

[4] K. 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.

[5] 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.

[6] J. Shor, D. Emanuel, O. Lang, O. Tuval, M. Brenner, J. Cattiau, F. Vieira, M. McNally, T. Charbonneau, M. Nollstadt *et al.*, "Personalizing asr for dysarthric and accented speech with limited data," *Interspeech*, pp. 784–788, 2019.

[7] 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.

[8] J. Scales, A. Wennerstrom, D. Richard, and S. H. Wu, "Language learners' perceptions of accent," *Tesol Quarterly*, vol. 40, no. 4, pp. 715–738, 2006.

[9] D. Wang, J. Yu, X. Wu, S. Liu, L. Sun, X. Liu, and H. Meng, "End-to-end voice conversion via cross-modal knowledge distillation for dysarthric speech reconstruction," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7744–7748.

[10] S. Liu, D. Wang, Y. Cao, L. Sun, X. Wu, S. Kang, Z. Wu, X. Liu, D. Su, D. Yu *et al.*, "End-to-end accent conversion without using native utterances," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6289–6293.

[11] F. Rudzicz, "Acoustic transformations to improve the intelligibility of dysarthric speech," in *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, 2011, pp. 11–21.

[12] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "Individuality-preserving voice conversion for articulation disorders based on non-negative matrix factorization," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8037–8040.

[13] R. Aihara, T. Takiguchi, and Y. Ariki, "Phoneme-discriminative features for dysarthric speech conversion." in *Interspeech*, 2017, pp. 3374–3378.

[14] C.-Y. Chen, W.-Z. Zheng, S.-S. Wang, Y. Tsao, P.-C. Li, and Y.-H. Lai, "Enhancing intelligibility of dysarthric speech using gated convolutional-based voice conversion system," *Interspeech*, pp. 4686–4690, 2020.

[15] S. Aryal and R. Gutierrez-Osuna, "Can voice conversion be used to reduce non-native accents?" in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 7879–7883.

[16] G. Zhao, S. Sonsaat, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Accent conversion using phonetic posteriorgrams," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5314–5318.

[17] G. Zhao and R. Gutierrez-Osuna, "Using phonetic posteriorgram based frame pairing for segmental accent conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1649–1660, 2019.

[18] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *Advances in neural information processing systems*, 2017, pp. 2962–2970.

[19] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 019–10 029.

[20] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *Interspeech*, pp. 4006–4010, 2017.

[21] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[22] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[24] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.

[25] K. Ito *et al.*, "The lj speech dataset," 2017.

[26] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[27] G. Zhao, S. Sonsaat, A. O. Silpachai, I. Lucic, E. Chukharev-Khudilaynen, J. Levis, and R. Gutierrez-Osuna, "L2-arctic: A non-native english speech corpus," *Perception Sensing Instrumentation Lab*, 2018.

[28] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[29] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443–445, 1985.

[30] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[32] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J. M. Cohen, H. Nguyen, and R. T. Gadde, "Jasper: An end-to-end convolutional neural acoustic model," *Interspeech*, pp. 71–75, 2019.