End-To-End Voice Conversion Via Cross-Modal Knowledge Distillation for Dysarthric Speech Reconstruction

Conference Paper · May 2020

DOI: 10.1109/ICASSP40776.2020.9054596

7 authors, including:

Jianwei Yu

76 PUBLICATIONS **710** CITATIONS

SEE PROFILE

Xixin Wu

The Chinese University of Hong Kong

77 PUBLICATIONS **751** CITATIONS

SEE PROFILE

Songxiang Liu

Tencent

40 PUBLICATIONS **465** CITATIONS

SEE PROFILE

Lifa Sun

SpeechX Limited

16 PUBLICATIONS **789** CITATIONS

SEE PROFILE

END-TO-END VOICE CONVERSION VIA CROSS-MODAL KNOWLEDGE DISTILLATION FOR DYSARTHRIC SPEECH RECONSTRUCTION

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

[1]Human-Computer Communications Laboratory
Department of System Engineering and Engineering Management
The Chinese University of Hong Kong, Hong Kong SAR, China
[2]SpeechX Limited, Shenzhen, China
{dswang, jwyu, wuxx, sxliu, lfsun, xyliu, hmmeng}@se.cuhk.edu.hk

ABSTRACT

Dysarthric speech reconstruction (DSR) is a challenging task due to difficulties in repairing unstable prosody and correcting imprecise articulation. Inspired by the success of sequence-to-sequence (seq2seq) based text-to-speech (TTS) synthesis and knowledge distillation (KD) techniques, this paper proposes a novel end-to-end voice conversion (VC) method to tackle the reconstruction task. The proposed approach contains three components. First, a seq2seq based TTS is first trained with the transcribed normal speech. Second, with the text-encoder of this trained TTS system as "teacher", a teacher-student framework is proposed for cross-modal KD by training a speech-encoder to extract appropriate linguistic representations from the transcribed dysarthric speech. Third, the speech-encoder of the previous component is concatenated with the attention and decoder of the first component (TTS) to perform the DSR task, by directly mapping the dysarthric speech to its normal version. Experiments demonstrate that the proposed method can generate the speech with high naturalness and intelligibility, where the comparisons of human speech recognition between the reconstructed speech and the original dysarthric speech show that 35.4% and 48.7% absolute word error rate (WER) reduction can be achieved for dysarthric speakers with low and very low speech intelligibility, respectively.

Index Terms— Dysarthric speech reconstruction, voice conversion, seq2seq, cross-modal, knowledge distillation

1. INTRODUCTION

Dysarthria denotes a set of speech disorders related with neurological conditions and diseases such as traumatic brain injury or stroke, Parkinson's disease or amyotrophic lateral sclerosis, which cause disturbances in muscular control over the speech production [1]. Therefore, dysarthria may result in unnatural and unintelligible speech with unstable prosody and imprecise articulation, which engender substantial communication difficulties for dysarthric patients. To enhance the quality of the dysarthric speech, various speech reconstruction techniques have been proposed and can be divided into two categories [2]: voice banking and voice conversion (VC). The former employs the speech recordings of patients to build personalized TTS systems before their speech deteriorates, while the latter adjusts the dysarthric speech signals to be more natural and intelligible, which is the focus of this paper.

VC has been widely applied to convert certain acoustic domains, such as speaker identity [3], speaking rate [4], accent [5] and emotion [6], while keeping the same linguistic content. VC also has the potential of improving the speech intelligibility of surgical patients with partial articulators removed [7]. For DSR, rule-based and statistical VC have been investigated. The rule-based VC modifies the temporal or frequency characteristics of speech according to specific rules [8, 9]. Though speech intelligibility can be improved, the rules are not stable as different dysarthric patients need different rules. Contrarily, statistical VC creates a mapping function between the acoustic features of dysarthric and normal speech. Popular approaches contain Gaussian mixture model (GMM) that converts formant and vowel features [1], non-negative matrix factorization (NMF) that builds dictionaries using mel-cepstral or spectrogram features [10-12], and partial least square (PLS) using phoneme-discriminative features [13]. Though significant progress has been made, the existing approaches have two main drawbacks. First, the prosody recovery performance is limited, e.g., the speech rate cannot be adjusted to normal as the frame-based mapping retains the abnormal speaking rate. Second, when the dysarthria becomes severe, the articulation correction is difficult to achieve due to the lack of training data and limited capacity of conversion models, which leads to small speech intelligibility improvement.

To tackle these two issues, this paper proposes a novel end-to-end VC method for DSR, which is inspired by the success of the seq2seq based TTS [14] and KD techniques [15, 16], where the latter is used in the teacher-student framework that transfers the distilled knowledge of the "teacher" to the "student". The approach consists of three components. First, a seq2seq TTS composed of the encoder, attention and decoder modules is trained with the transcribed normal speech, and the text-encoder of the TTS is used in the next component. Second, we employ the teacher-student framework to perform cross-modal KD by using the transcribed dysarthric speech training data. The text-encoder trained from the previous component serves as the "teacher" and guides the leaning of a speech-encoder which is the "student". In other words, the idea is to transfer the distilled knowledge across modalities from text to speech by forcing the outputs of the speech-encoder to be similar with those of the text-encoder. When the speech-encoder is well-trained, it can replace the text-encoder to generate appropriate linguistic representations. Then the speech-encoder is used in the next (i.e. third) component. The third component is an end-to-end VC based DSR system that concatenates the speech-encoder with the attention and decoder

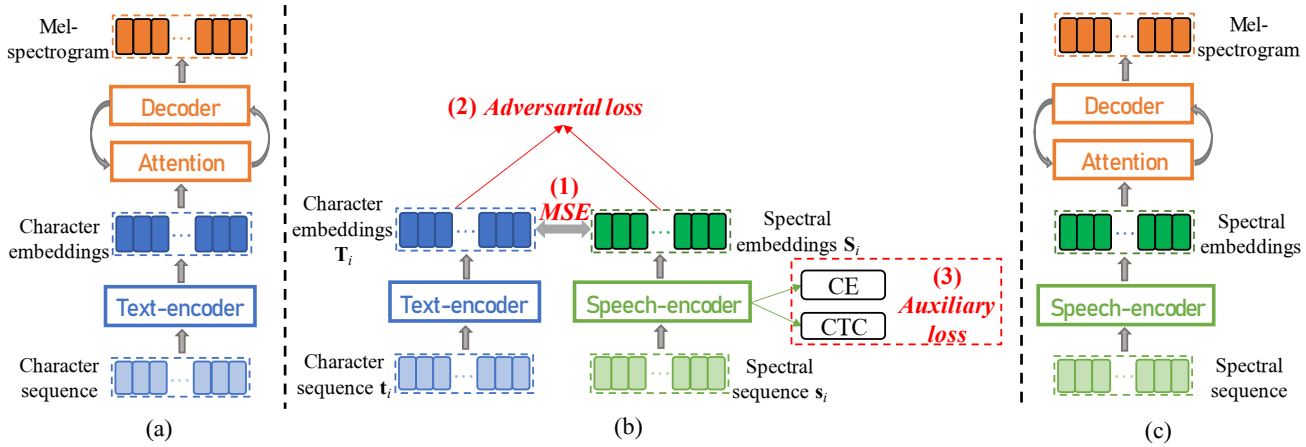

Fig. 1. Diagram of end-to-end VC for dysarthric speech reconstruction: (a) TTS system trained on the transcribed normal speech; (b) Cross-modal KD with the transcribed dysarthric speech; (c) End-to-end DSR system.

modules of TTS (i.e. first component). Dysarthric speech features are directly used as input to the VC system to generate the reconstructed normal speech features. Previously, the end-to-end approach has been used in VC for normal speech [17]. However, to the best of our knowledge, the proposed architecture that combines end-to-end and cross-modal KD approaches is among the first to be used for the reconstruction of dysarthric speech.

2. PROPOSED METHOD

Fig. 1 shows the diagram of the proposed method, which contains three components: (a) TTS system is first trained with the transcribed normal speech; (b) Cross-modal KD is performed with the transcribed dysarthric speech to train a speech-encoder to generate appropriate linguistic representations; (c) End-to-end DSR system that directly converts the dysarthric speech to its normal version.

2.1. TTS system

Recent developments in TTS show that attention based seq2seq models can yield high-fidelity speech synthesis with compact structures. Therefore, we propose to utilize Tacotron [14], which is a state-of-the-art architecture, as the TTS model that contains encoder, attention and decoder modules as shown in Fig. 1(a). The input to the text-encoder is the character sequence, where each character is denoted as a one-hot vector, which is processed by the encoder to derive robust linguistic representations, namely, character embeddings. Then for each time step, the attention module takes the character embeddings as inputs to produce the context vectors that are fed into the decoder to generate the mel-spectrogram features, which are used for waveform synthesis via neural-network-based vocoder. In this paper, WaveRNN [18] is used as it can synthesize the high-quality waveform with fast inference. We use the transcribed speech of one normal speaker to train the TTS, so the TTS can be used to generate normal speech with stable prosody and precise articulation. In the following, the parameters of well-trained TTS are frozen during the KD process.

2.2. Cross-modal KD for Speech-encoder training

This paper strives to improve the speech quality of dysarthric patients, thus how to extract appropriate linguistic information that can be mapped to normal speech is important. As the text-encoder of TTS is used to derive character embeddings which are only associated with linguistics, we hope to build a speech-encoder to mimic the text-encoder to extract linguistic-related information from the dysarthric speech. KD is an effective technique applied in the teacher-student framework, where the knowledge is distilled from a teacher to guide the learning of a student [15, 16]. Therefore, we propose the cross-modal KD for speech-encoder training as shown in Fig. 1(b). By treating the well-trained text-encoder of TTS as the "teacher", the "student" (speech-encoder) is trained to make its outputs to be similar with those of the "teacher".

Assume there are N dysarthric speech-text pairs used for training, the i^{th} pair is denoted as $\{\mathbf{s}_i, \mathbf{t}_i\}$, where $\mathbf{s}_i$ and $\mathbf{t}_i$ are the spectral and character sequence respectively. By taking $\mathbf{t}_i$ as the input of the text-encoder, and $\mathbf{s}_i$ as the input of the speech-encoder, the character embeddings $\mathbf{T}_i$ and spectral embeddings $\mathbf{S}_i$ can be obtained, respectively. To align the spectral and character embeddings automatically, the speech-encoder also adopts the attention based seq2seq architecture [17] as shown in Fig. 2. We transfer the knowledge from the text-encoder to the speech-encoder by forcing the character and spectral embeddings to be similar. Three types of losses are considered during the training, i.e. mean square error (MSE) loss, adversarial loss and auxiliary loss as highlighted in Fig. 1(b).

MSE loss: This constrains spectral embeddings to be close to corresponding character embeddings, the MSE loss L_{MSE} is used:

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{T}_i - \mathbf{S}_i\|_2^2 \quad (1)$$

Adversarial loss: The spectral and character embeddings have different statistical distributions as they are generated from two different modalities. To alleviate the mismatch in training, a text/speech domain discriminator f_D is introduced and trained via adversarial learning [19]. On one hand, the discriminator is trained to classify whether the input to the discriminator is the character embedding or spectral embedding by minimizing the cross entropy (CE) loss:

$$L_D = \frac{1}{N}\sum_{i=1}^{N}\left(-\log f_D(\mathbf{T}_i) - \log\left(1 - f_D(\mathbf{S}_i)\right)\right) \quad (2)$$

On the other hand, the speech-encoder is trained to 'fool' the discriminator to make its outputs to be the same either when the input are character or spectral embeddings by minimizing the loss:

$$L_{ADV} = \frac{1}{N}\sum_{i=1}^{N}\left(\left\|\frac{1}{2}\boldsymbol{e} - f_D(\mathbf{T}_i)\right\|_2^2 + \left\|\frac{1}{2}\boldsymbol{e} - f_D(\mathbf{S}_i)\right\|_2^2\right) \quad (3)$$

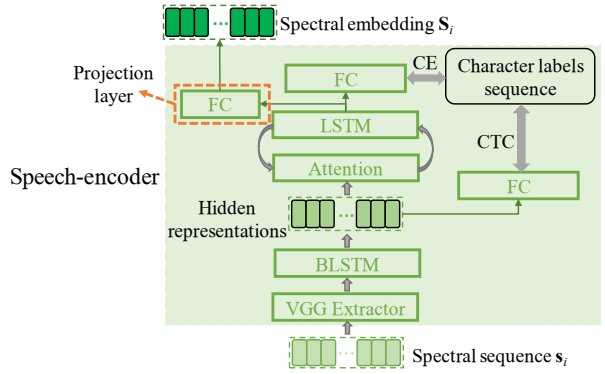

Fig. 2. Diagram of speech-encoder

where $\boldsymbol{e}$ is a two-dimensional all-one vector. As a result, the statistical characteristics of spectral and character embeddings are expected to be similar, which further reduces discrepancies between them.

Auxiliary loss: The seq2seq approach with connectionist temporal classification (CTC) for automatic speech recognition (ASR) [20] is adopted for the speech-encoder, which is composed of the encoder (including 6-layer VGG extractor and 5-layer bidirectional long short-term memory (BLSTM) with 320 units per direction), location-aware attention [21] and decoder (one-layer LSTM with 320 units) as shown in Fig. 2. To steer the spectral embeddings towards a pure linguistic space and stabilize the training process, similar with [20], we use the ASR loss as the auxiliary loss containing the CE and CTC losses of character prediction:

$$L_{AUX} = L_{CE} + L_{CTC} \tag{4}$$

where L_{CE} is the CE calculated between the decoder outputs of the speech-encoder and the ground-truth character labels sequence, L_{CTC} is the CTC calculated between the encoder outputs of the speech-encoder and all possible sequences that can be mapped to ground-truth character labels sequence by inserting blanks and repeating characters [20]. The characters include 26 alphabets (A~Z) and one end-of-sentence token.

Therefore, the training loss for the speech-encoder is a combination of the above losses through weighting factors:

$$L_{SE} = \alpha_1 L_{MSE} + \alpha_2 L_{ADV} + \alpha_3 L_{AUX} \tag{5}$$

where α_1, α_2 and α_3 are the weights empirically set to 0.1, 1 and 0.5, respectively. During the training process, parameters of speech-encoder and text/speech domain discriminator are updated alternatively via adversarial learning.

2.3. End-to-end DSR system

Given the dysarthric speech-text pair, when the speech-encoder is well-trained, it can generate the spectral embeddings similar with the text embeddings produced by the text-encoder. Therefore, the speech-encoder can be concatenated with the attention and decoder modules of the TTS system (i.e. first component) to form an end-to-end DSR system, as shown in Fig. 1(c). At the conversion phase, the speech-encoder takes the dysarthric spectral sequence as the input for spectral embeddings generation, which is terminated when the speech-encoder predicts the end-of-sentence token. Then the spectral embeddings are fed into the attention and decoder modules of TTS to predict a normal mel-spectrogram, which is utilized for waveform synthesis with the same WaveRNN vocoder used by the TTS.

3. EXPERIMENTS

3.1 Experimental settings

The experiments are implemented on the LJSpeech [22] and UASpeech [23] datasets. The TTS and WaveRNN training are conducted on the LJSpeech dataset which contains around 24 hours transcribed speech of a normal female speaker, and the cross-modal KD is conducted on the UASpeech dataset which contains 15 dysarthric speakers with cerebral palsy and 13 normal speakers. For UASpeech, each speaker has 3 blocks of utterances, where each block consists of 10 digits, 26 alphabets, 19 computer commands, 100 common words and 100 uncommon words, which are not repeated across blocks. All speech data are sampled at 16 kHz. As the severity of dysarthria is varied among different patients, which increases the modeling difficulties to build one DSR system for all patients, speaker-dependent DSR systems are built for 4 dysarthric speakers (F05, M05, M07, F03) with speech having high, middle, low and very low intelligibility, respectively, where block 1 and 3 are used for training and block 2 is used for testing. To improve the training performance, data augmentation is performed by modifying the tempo of all normal speech via *Sox* [24] with 0.6 ratio, then all normal speech and its modified version are added as training data for each dysarthric speaker.

The TTS and vocoder adopt the original architecture of Tacotron [14] and WaveRNN [18] respectively, following the same training settings in [14, 18], where the TTS output is 80-band mel-spectrogram. The speech-encoder has the similar architecture as the ASR model in [20], except for that one 256-dimensional fully-connected (FC) projection layer is added after the decoder output as shown in Fig. 2, which ensures the spectral and character embeddings have the same dimension. The text/speech domain discriminator takes each frame of spectral or character embeddings as input and is a 4-layer FC neural network (256→512→512→2). The input of speech-encoder is 40-band mel-spectrogram appended with delta and delta-delta features, which are calculated by using a 25ms Hanning window, 10ms frame shift and 400-point fast Fourier transform (FFT). Adadelta method [25] is applied for speech-encoder and discriminator training with learning rate of 1, batch size of 16 and 50k training steps.

Three baseline methods are compared with our proposed method: (1) Joint dictionary learning NMF (JDNMF) based VC [12] used for impaired speech reconstruction; (2) Deep BLSTM (DBLSTM) based VC system [26]; (3) ASR-TTS system which takes the ASR results as the input of the proposed TTS to synthesize normal speech. The ASR is the CUHK Dysarthric Speech Recognition System [27], which is a state-of-the-art systems on the UASpeech dataset. The first two baseline methods map the dysarthric spectral features to normal spectral features, following the original training and testing pipelines in [12, 26], where the parallel training data is required. The proposed TTS is used to obtain the target speech with the text transcriptions for fair comparison. It is noted that the ASR-TTS system is a strong baseline, as it achieves state-of-the-art recognition results and generates high-quality speech with the proposed TTS.

3.2 Experimental results

To verify the effectiveness of the proposed method to improve the quality of dysarthric speech, we conducted three subjective evaluation tests including two mean opinion score (MOS) tests and one by human speech recognition. 10 listeners are asked to conduct 5-scale MOS tests (1-bad, 2-poor, 3-fair, 4-good, 5-excellent) in

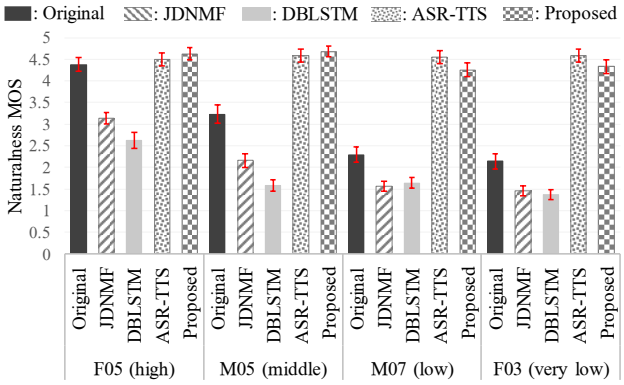

Fig. 3. MOS comparison based on naturalness. F05 (high) denotes F05 speaker with high speech intelligibility.

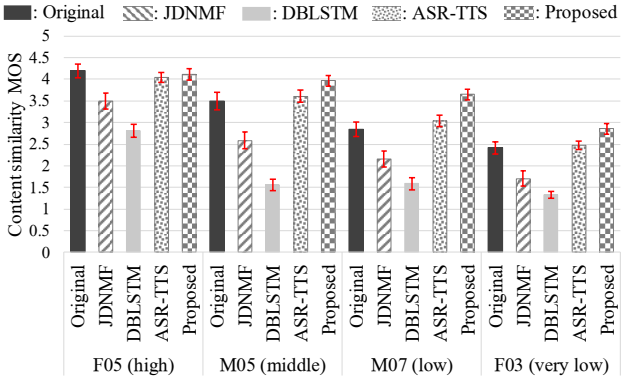

Fig. 4. MOS comparison based on Content similarity.

terms of speech naturalness and content similarity between the converted speech and reference speech of CF02, who is a normal speaker selected from the UASpeech. The content similarity can be used to measure how much linguistic content can be preserved after the transformation [7]. With reference to previous work with UASpeech [23], our experiments engage 5 listeners to perform speech recognition and word error rates (WER) are reported. 15 and 30 randomly selected sentences of each dysarthric speaker are used for MOS tests and human speech recognition, respectively. Readers are encouraged to listen to our audio samples[1].

Fig. 3 and Fig. 4 show the naturalness and content similarity MOS respectively, where 'Original' denotes the original dysarthric speech. Compared with the original speech, we can observe that the reconstructed speech of both JDNMF and DBLSTM is less natural and preserves less original content, due to the inadequate dysarthric speech training data and limited capacity of VC models trained by direct spectral mapping. Contrarily, with more training data by adding the normal speech, the ASR-TTS and proposed method can achieve quality improvements for almost all cases. Fig. 4 shows that the proposed method outperforms ASR-TTS consistently with more content preservation, Fig. 3 shows that both ASR-TTS and proposed methods can generate more natural speech, and the proposed method outperforms and underperforms ASR-TTS when the original speech intelligibility is relatively high (F05 & M05) and low (M07 & F03), respectively. ASR-TTS can generate natural speech irrelevant to ASR results, but when the dysarthric speech intelligibility decreases, the speech-encoder of the proposed system may produce inaccurate spectral embeddings used for speech reconstruction, which lowers the quality of the output speech. Besides, the proposed method strives to preserve as much content as possible by using spectral embeddings, which results in higher content similarity.

[1] https://wendison.github.io/E2E-DSR-demo/

Table 1. Human speech recognition performance across different dysarthric speakers and systems. Δ (%) denotes the absolute WER reduction with respect to the "Original" dysarthric speech

Speaker	Systems	WER / Δ	Speaker	Systems	WER / Δ
F05 (high)	Original	**8.67** / -	M05 (middle)	Original	44.00 / -
	JDNMF	31.32 / −22.65		JDNMF	78.00 / −34.00
	DBLSTM	56.00 / −47.33		DBLSTM	92.67 / −48.67
	ASR-TTS	10.67 / −2.00		ASR-TTS	35.33 / 8.67
	Proposed	9.33 / **−0.66**		Proposed	**34.67 / 9.33**
M07 (low)	Original	78.67 / -	F03 (very low)	Original	93.32 / -
	JDNMF	94.67 / −16.00		JDNMF	97.33 / −4.01
	DBLSTM	82.00 / −3.33		DBLSTM	98.67 / −5.35
	ASR-TTS	46.67 / 32.00		ASR-TTS	59.33 / 33.99
	Proposed	**43.32 / 35.35**		Proposed	**44.67 / 48.65**

Table 1 demonstrates human speech recognition results. Similarly, compared with the original speech, the reconstructed speech intelligibility of both JDNMF and DBLSTM are lower with larger WER. Besides, we can see that no systems can reduce the WER for F05 original speech. We suspect that F05 original speech is intelligible enough and the reconstructed speech has some artifacts that degrade the human speech recognition performance. However, the proposed method still outperforms ASR-TTS and achieves the lowest WER for all speakers among all reconstruction systems, with 9.33%, 35.35% and 48.65% absolute WER reduction for M05 (middle), M07 (low) and F03 (very low), respectively. This shows that the speech-encoder of the proposed system can derive appropriate linguistic representations, which can be used to generate speech with high intelligibility.

4. CONCLUSIONS

This paper makes a first attempt to apply end-to-end VC based on KD to the DSR task. With the attention based encoder-decoder architecture of TTS, the idea is to transfer the cross-modal knowledge from the text-encoder of a well-trained TTS to a speech-encoder, forcing the speech-encoder to mimic the text-encoder in producing appropriate linguistic representations from the dysarthric speech. Then the linguistic representations can be fed into the attention and decoder modules of TTS to generate normal speech with stable prosody and precise articulation. Extensive experiments show that significant speech quality improvements can be achieved, especially for patients with severe dysarthria. It is noted that the proposed method can be extended to other conversion tasks, such as speaker identity, emotion, speaking style and accent. For instance, by replacing the proposed single-speaker TTS with multi-speaker TTS [28], the proposed system can generate the high-quality speech that preserves both speaker identity and content, which is our future work.

5. ACKNOWLEDGEMENTS

This project is partially supported by the General Research Fund from the Research Grants Council of Hong Kong SAR Government (Project No. 14208817).

6. REFERENCES

[1] 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.

[2] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction," *Acoustical Science and Technology,* vol. 33, no. 1, pp. 1-5, 2012.

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

[4] 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)*, 2003: IEEE, pp. 706-711.

[5] 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)*, 2017: IEEE, pp. 182-188.

[6] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "GMM-based emotional voice conversion using spectrum and prosody features," *American Journal of Signal Processing,* vol. 2, no. 5, pp. 134-138, 2012.

[7] L.-W. Chen, H.-Y. Lee, and Y. Tsao, "Generative Adversarial Networks for Unpaired Voice Transformation on Impaired Speech," *arXiv preprint arXiv:1810.12656,* 2018.

[8] 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: Association for Computational Linguistics, pp. 11-21.

[9] S. A. Kumar and C. S. Kumar, "Improving the intelligibility of dysarthric speech towards enhancing the effectiveness of speech therapy," in *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2016: IEEE, pp. 1000-1005.

[10] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "Consonant enhancement for articulation disorders based on non-negative matrix factorization," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2012: IEEE, pp. 1-4.

[11] 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*, 2013: IEEE, pp. 8037-8040.

[12] S.-W. Fu, P.-C. Li, Y.-H. Lai, C.-C. Yang, L.-C. Hsieh, and Y. Tsao, "Joint dictionary learning-based non-negative matrix factorization for voice conversion to improve speech intelligibility after oral surgery," *IEEE Transactions on Biomedical Engineering,* vol. 64, no. 11, pp. 2584-2594, 2016.

[13] R. Aihara, T. Takiguchi, and Y. Ariki, "Phoneme-Discriminative Features for Dysarthric Speech Conversion," in *Interspeech*, 2017, pp. 3374-3378.

[14] Y. Wang *et al.*, "Tacotron: Towards End-to-End Speech Synthesis," *Proc. Interspeech 2017,* pp. 4006-4010, 2017.

[15] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531,* 2015.

[16] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4133-4141.

[17] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Non-Parallel Sequence-to-Sequence Voice Conversion with Disentangled Linguistic and Speaker Representations," *arXiv preprint arXiv:1906.10508,* 2019.

[18] N. Kalchbrenner *et al.*, "Efficient Neural Audio Synthesis," in *International Conference on Machine Learning*, 2018, pp. 2415-2424.

[19] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672-2680.

[20] A. H. Liu, H.-y. Lee, and L.-s. Lee, "Adversarial training of end-to-end speech recognition using a criticizing language model," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019: IEEE, pp. 6176-6180.

[21] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577-585.

[22] K. Ito, "The lj speech dataset, 2017a. URL ttps," *keithito. com/LJ-Speech-Dataset*.

[23] H. Kim *et al.*, "Dysarthric speech database for universal access research," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[24] SoX, audio manipulation tool. http://sox.sourceforge.net/ (accessed September 1, 2019).

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

[26] L. 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)*, 2015: IEEE, pp. 4869-4873.

[27] J. Yu *et al.*, "Development of the CUHK Dysarthric Speech Recognition System for the UA Speech Corpus," in *Interspeech*, 2018, pp. 2938-2942.

[28] Y. Jia *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Advances in neural information processing systems*, 2018, pp. 4480-4490.