Two-stage and Self-supervised Voice Conversion for Zero-Shot Dysarthric Speech Reconstruction

Dong Liu *, Yueqian Lin *, Hui Bu †, Ming Li *

*Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems,
Data Science Research Center, Duke Kunshan University, Kunshan, China

†AISHELL Foundation, Beijing, China

***Abstract*—Dysarthria is a motor speech disorder commonly associated with conditions such as cerebral palsy, Parkinson's disease, amyotrophic lateral sclerosis, and stroke. Individuals with dysarthria typically exhibit significant speech difficulties, including imprecise articulation, lack of fluency, slow speech rate, and decreased volume and clarity, which can hinder their ability to communicate effectively with others. We propose a two-stage Voice Conversion method to enhance the reconstruction of dysarthric speech. In the first stage, we develop a KNN-VC approach based on a same-gender-retrieval strategy to preliminarily repair the dysarthric speech. In this stage, we match the dysarthric speech only with normal speech of the same gender. In the second stage, we adapt so-vits-svc to restore the speaker's timbre and improve the sound quality of the speech repaired in the first stage. Both objective and subjective evaluations were conducted on the dataset of the Low Resource Dysarthria Wake-Up Word Spotting Challenge (LRDWWS Challenge) shows that the proposed approach can achieve some improvements in terms of speaker similarity, speech intelligibility and naturalness for unknown speakers, and these evaluations also show our method has a good Zero-shot performance. Our audio samples can be accessed online [1].**

***Index Terms*—dysarthric speech reconstruction, Any-to-any, Zero-shot, voice conversion**

I. Introduction

Dysarthria, a motor speech disorder arising from neurological conditions such as cerebral palsy or amyotrophic lateral sclerosis, significantly impairs an individual's ability to communicate effectively. Characterized by weak muscle control in speech production, dysarthric speech often exhibits instability in prosody and imprecision in articulation, leading to reduced intelligibility and substantial communication barriers. [14] Articulation disorders primarily affect the production of speech sounds, resulting in difficulties such as substitutions, omissions, distortions, and additions of phonemes. These disruptions can significantly impair speech intelligibility, making it challenging for others to understand the affected individual. Common issues include replacing one sound with another (e.g., "wabbit" for "rabbit"), omitting sounds within words (e.g., "ca" for "cat"), distorting sounds (producing sounds in an atypical manner), and adding extra sounds.

To enhance the naturalness and intelligibility of affected speech, the development of dysarthric speech reconstruction (DSR) techniques has become a critical area of research. DSR is among the effective solutions to assist communication by converting dysarthric speech to normal speech with higher intelligibility and naturalness. The approaches to DSR can be classified into two types [25]: Voice banking and voice conversion(VC). Voice banking involves creating personalized text-to-speech (TTS) systems using recordings from patients before their speech deteriorates, which requires building a database of utterances when patients are in good health, imposing significant constraints. VC has been widely applied to convert certain acoustic domains, such as speaker identity [17], speaking rate [21], accent [19] and emotion [2], while keeping the same linguistic content. VC also has the potential of improving the speech intelligibility of surgical patients with partial articulators removed [6]. For DSR, there are also multiple VC methods that could be applied here. Rule-based methods modify the temporal or frequency characteristics of speech signals according to predefined rules, while statistical VC creates a mapping function between the acoustic features of dysarthric and normal speech, often using models like Gaussian mixture models (GMM) [14], non-negative matrix factorization (NMF) [1], [3], [11], and partial least squares (PLS) [4].

With the development of deep learning, sequence-to-sequence (seq2seq) based TTS synthesis [24] and knowledge distillation (KD) techniques [13], [27] have emerged as promising solutions for DSR. Wang et al. [23] proposed an end-to-end voice conversion (E2E-VC) based DSR system with cross-modal knowledge distillation. A seq2seq TTS model is trained on normal speech, and its text-encoder guides a speech-encoder to extract linguistic features from dysarthric speech. This speech-encoder is then combined with the TTS model's attention and decoder to convert dysarthric speech to normal speech. This approach has shown promising results in improving the prosody and speaker similarity of reconstructed speech.

Chen et al. [8] proposed an innovative multimodal framework for improving DSR, integrating audio and visual information by utilizing a pre-trained AV-HuBERT model to extract more accurate semantic representations. This significantly enhances the intelligibility and naturalness of speech, particularly for patients with more severe symptoms.

The methods previously discussed have primarily focused on enhancing the naturalness and intelligibility of speech for

Corresponding Author, E-mail: ming.li369@dukekunshan.edu.cn

[1] https://raydonld.github.io/Two-stage-and-Self-supervised-Voice-Conversion-for-Zero-Shot-DSR/

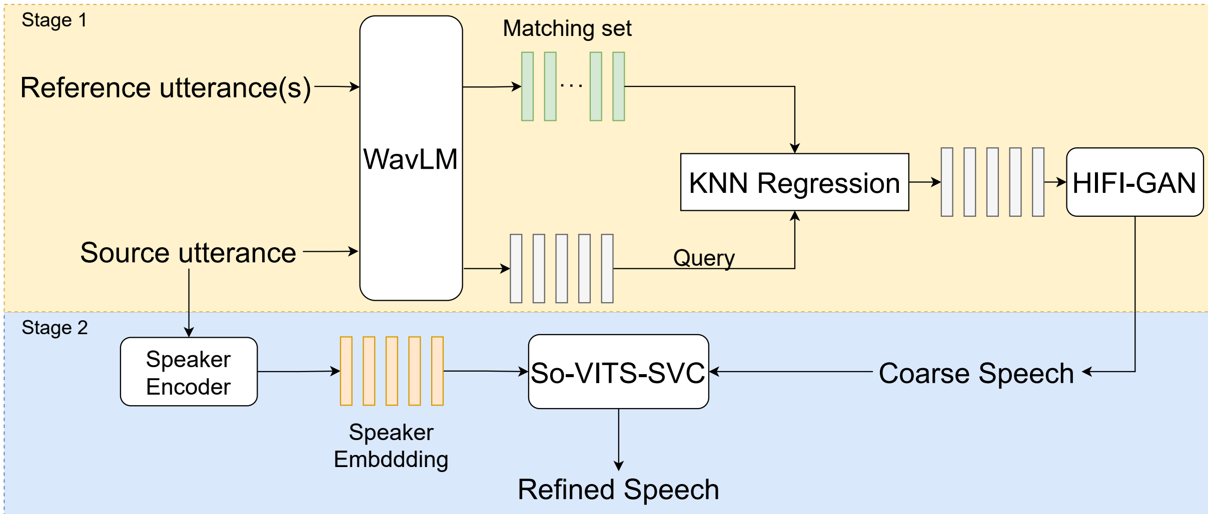

Fig. 1. Architecture of our model.

specific dysarthria patients. However, these approaches exhibit significant limitations when dealing with unseen dysarthric speakers, particularly in zero-shot learning and timbre restoration. They have difficulties to effectively repair the speech of previously unseen dysarthria patients while simultaneously recovering their unique timbral characteristics. Consequently, there is a need to improve these methods' capabilities in zero-shot learning and timbral fidelity for the speech of unknown dysarthria patients.

Our approach to DSR introduces a novel method that addresses these challenges through a two-stage process. For the first stage, we propose a same-gender-retrieval matched KNN-VC approach [5]. For male patients with dysarthria, the KNN-VC approach references the speech of healthy male speakers, while for female patients, it references the speech of healthy female speakers. The design of this strategy is intended to emphasize the lexical content of the speech while mitigating the impact caused by timbral differences. In this stage, we can generate the coarse reconstruction speech. And then, for the second stage, these coarse voices are fed into the so-vits-svc model[2], which has been fine-tuned with our dataset, to restore the original patient's timbre and enhance speech clarity.

II. Proposed Methods

We propose a two-stage speech disorder reconstruction (DSR) model, as shown in Figure 1. The model consists of two parts: a coarse-grained speech converter, which extracts features related to text content from speech with impediments, and a fine-grained speech converter, which restores the speaker's voice to its original state and optimizes speech in detail. This hierarchical processing mechanism aims to achieve efficient conversion from impaired speech to normal speech.

A. Stage1: the coarse reconstruction speech generator

To restore the semantic content of speech with articulation disorders to normal speech, we see this issue as an any-to-any speech conversion problem. We proposed a same-gender-retrieval KNN-VC method, an any-to-any speech conversion model, consisting of three components: a WavLM module, which is a self-supervised model, a KNN matcher module, and a HiFiGAN vocoder. First, we use a pre-trained WavLM model to extract feature sequences of the patient and same-gender normal speakers. Next, we replace each frame's representation in the patient speaker's feature sequence with the average of the K nearest neighbor feature sequences from the same-gender normal speaker's speech. Finally, we use the HiFiGAN model, trained on our dataset, to synthesize the coarse repaired speech.

To achieve the same-gender-retrieval method, we emphasize retaining information related to text content to improve speech intelligibility. Specifically, for male patients with articulation disorders, we only use speech data from normal male speakers for matching and replacement, and for female patients with articulation disorders, we only use data from from normal female speakers. This replacement strategy helps train the WavLM features to achieve good performance in speech restoration for articulation disorders.

B. Stage2: Restoring timbre and enhancing sound quality

To restore the original speaker's timbre and improve the quality of speech generated in the first stage, we further adopted a pre-trained so-vits-svc model for converting the output of the first stage back to the source speaker's timbre. So-vits-svc is an open-source voice conversion tool based on VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech) [15]. VITS is an advanced text-to-speech (TTS) model that combines the characteristics of autoregressive and flow-based models to generate highly natural speech. So-vits-svc leverages VITS's capabilities to convert one speaker's voice into another's while maintaining the original speaker's intonation and content.

During the training, the speech generated in the first stage was used to fine-tune the pre-trained so-vits-svc model to

[2]Code: https://github.com/PlayVoice/so-vits-svc-5.0

TABLE I

DATA DESCRIPTION OF THE LRDWWS CHALLENGE

Type	Num of Speaker	Duration(min)	Num of Utterances
Training set	Control: 25	454	10114
	Dysarthria:17	449	6885
Developing set	Dysarthria:4	116	1620
Testing set	Dysarthria:10	296	4380

optimize the model's ability to represent the target speaker's timbre, thereby improving speech conversion quality. During the inference, we performed voice conversion using the fine-tuned so-vits-svc model. The coarse repaired speech in Step 1 was converted to the refined speech, which retains the linguistic content of the coarse speech while adopting the unique timbral properties of the patient's voice.

III. EXPERIMENTS

A. Experimental Settings

To verify the effectiveness of our proposed method, we used the same dataset and partition settings as the Low Resource Dysarthria Wake-Up Word Spotting Challenge (LRDWWS Challenge)[3]. The LRDWWS Challenge focuses on low-resource speech recognition with speech impairment. The dataset is recorded in a relatively quiet environment using devices such as mobile phones at a sampling rate of 16kHz, eliminating the interference of complex environmental noise on speech signals. Inaccurate pronunciation and reduced fluency in patients with articulation disorders can lead to significant phonetic variation. Even among patients with the same type of articulation disorder, pronunciation of the same word can vary under different conditions. Due to the physical condition of patients with articulation disorders, long-term recording is usually not possible; thus, this dataset mainly comprises specific short-term speech samples. For a detailed description of the dataset, please refer to Table I.

In the first stage of speech restoration for articulation disorders, we adopted a pre-trained WavLM-large model [7] as the audio feature extraction module and maintained its pre-training state without further training. At this stage, we need to train was the HiFiGAN [16] vocoder, which receives features extracted by WavLM as input. The normal and abnormal speech of the training set are used to train the HIFIGAN vocoder. In the second stage, we fine-tuned the so-vits-svc model using the coarse repaired speech data generated in the first stage and the normal speech in the training set to further improve the quality of speech conversion.

B. Experimental Results and Analysis

To quantify the effectiveness of our method in repairing speech with articulation disorders, we used Word Error Rate (WER) to measure the recognition of the repaired speech. When calculating WER, we used the open-source WeNet model [26] for automatic speech recognition of repaired speech and compared the recognition results with the actual text to obtain objective evaluation indicators of speech intelligibility. To evaluate the similarity in timbre between the repaired speech and the speech of patients with original articulation disorders, we calculated the speaker similarity. Additionally, we subjectively evaluated the naturalness of speech restoration using the Mean Opinion Score (MOS) to validate its performance in terms of auditory naturalness.

To validate the effectiveness of the same-gender-retrieval strategy employed in the first stage (Stage 1), we constructed two systems that only involve the first stage for comparative analysis. System S1 represents the first stage system that utilizes a cross-gender-retrieval strategy, while system S2 represents the first stage system that employs the same-gender-retrieval strategy. Systems S1 and S2, Which only include Stage 1, are focused on restructuring the lexical content of speech to improve speech intelligibility.

Stage 2 will restore the original speaker's timbre and improve the quality of speech generated in Stage 1. To further verify the effectiveness of the second stage (Stage 2), we built systems S3 and S4 on the basis of S1 and S2, respectively, which included the second stage method. Through this staged construction and comparison, we could accurately assess the performance of each stage as well as the overall system under different strategies.

1) Objective Evaluation:

a) Speaker similarity

To evaluate the Speaker similarity, we use the d-vector method, as outlined in the literature [22], [28]. The d-vector represents speaker characteristics for each audio segment. The speech speaker verifier was employed to obtain the d-vector for each audio segment. In the speech speaker verifier, we used a speaker verification extraction model employing the ResNet101-ASP architecture from [20] for the speaker representation extraction system. The ResNet101 model [12] with residual module channel settings $[32, 64, 128, 256]$ serves as the front-end feature extractor. It is followed by an Attentive Statistics Pooling (ASP) layer [18], and a 256-dimensional fully connected layer is used as the speaker representation layer. ArcFace [10] ($s = 32, m = 0.2$) is utilized as the classifier. The model is trained on the VoxCeleb2 dataset [9] (5994 speakers with 1092009 utterances).

Subsequently, we calculated the average d-vector for each speaker, obtaining a speaker-level d-vector (denoted as V). We performed distance measurements on speaker-level embedding V from two different types:

- **Speaker-level d-vector for ground-truth target speaker utterances(gt)**: we compute V_i^{gt} by averaging the d-vectors of all testing utterances from speaker i using the ground-truth target speech.
- **Speaker-level d-vector for synthesized speech utterances using conventional speaker embeddings(syn)**: we compute V_i^{syn} by averaging the d-vectors of all testing utterances from speaker i using the synthesized speech.

We differentiated between different speakers by computing the cosine distance between different V. The cosine distance is defined as $d(V_1, V_2) = 1 - \frac{V_1}{\|V_1\|} \cdot \frac{V_2}{\|V_2\|}$. We define the set of testing speakers as T, and the following metrics are the

[3]LRDWWS Challenge Website: http://lrdwws.org/

three metrics that were computed to evaluate performance in speaker similarity:

TABLE II
COSINE DISTANCE BETWEEN DIFFERENT SPEAKERS

Methods	syn2syn-near	syn2gt-same	syn2gt-near	Original speaker timbre restoration
S1	0.115	0.672	0.509	×
S2	0.145	0.515	0.479	×
S3	0.231	0.577	0.518	✓
S4	0.231	0.499	0.524	✓

TABLE III
THE WER OF THESE FOUR SYSTEMS

System	WER(%)	Original speaker timbre restoration
Original speech	66.06%	–
S1	78.79%	×
S2	76.02%	×
S3	85.13%	✓
S4	82.68%	✓

TABLE IV
THE NATURALNESS MOS WITH 95% CONFIDENCE INTERVALS

System	Naturalness MOS score	Original speaker timbre restoration
Original speech	2.502 ± 0.03	–
S1	2.524 ± 0.03	×
S2	2.676 ± 0.03	×
S3	2.386 ± 0.04	✓
S4	2.694 ± 0.05	✓

- **Syn2gt-same**: Compute the speaker-level d-vector distance between synthetically generated speech V_i^{syn} and the corresponding ground truth V_i^{gt} for the same speaker in the Testing set. The smaller, the better.

$$\underset{i \in T}{\text{mean}}\, d(V_i^{syn}, V_i^{gt}) \tag{1}$$

- **Syn2gt-near**: Calculate the average distance between the synthetic testing speaker V_i^{syn} and the closest ground truth testing speaker V_i^{gt}. This metric is used to assess the differences between synthesized speech and the ground true speech of other speakers in the Testing set. The larger, the better.

$$\underset{i \in T}{\text{mean}} \min_{j \in T, i \neq j} d(V_i^{syn}, V_j^{gt}) \tag{2}$$

- **Syn2syn-near**: Compute the average minimum distance between synthesized speech V_i^{syn} for different speakers in the testing set. This metric is applied to measure the worst-case performance in audio synthesis for different speakers within the Testing set. The larger, the better.

$$\underset{i \in T}{\text{mean}} \min_{j \in T, i \neq j} d(V_i^{syn}, V_j^{syn}) \tag{3}$$

The speaker similarity of these four systems is shown in Table II. We found that among the systems containing only the first stage, S1 and S2, the S2 system, which employed the same-gender-retrieval strategy, achieved a higher degree of similarity between the synthetic speech of the same speaker and the speaker's actual speech compared to the S1 system. Additionally, the S2 system demonstrated better discriminability between the speeches of different speakers. This result laid the foundation for the restoration of the original speaker's timbre in the second stage (Stage 2).

Upon further comparison between the S3 and S4 systems, we observed that the S4 system, which also utilized the same-gender-retrieval strategy, outperformed the S3 system in terms of speaker timbre restoration and exceeded the performance of the S2 system. This is because the S3 system had to perform gender conversion during the timbre restoration process, an additional conversion step that increased the difficulty of timbre restoration. This also indicates that the second stage (Stage 2) can effectively restore the timbre to a certain extent.

b) Speech intelligibility

To demonstrate the content intelligibility improvement of the final reconstructed speech compared with the original dysarthric speech, we used the WeNet ASR model to obtain the word error rate (WER). The result is shown in Table III. Through comparative analysis of these four systems, it is evident that systems adhering to the same-gender-retrieval strategy outperform those employing other retrieval strategies. System S2 achieves the lowest Word Error Rate (WER), and it is approved that the same-gender-retrieval strategy makes the system focus on the lexical content to improve speech intelligibility. However, these systems that only include this strategy can't restore the original speaker's timbre.

Notably, although the two-stage system S4, which follows the same-gender-retrieval strategy, could restore the original speaker's timbre, the WER of system S4 is higher than that of system S2, which only includes Stage 1. For the two-stage systems, The WER of the S4 system is lower than the S3 system.

2) Subjective Evaluation:
MOS tests were conducted to verify the effectiveness of the proposed method in improving the intelligibility and naturalness of reconstructed speech. Ten subjects participated in the MOS tests (1-bad, 2-poor, 3-fair, 4-good, 5-excellent) to evaluate speech intelligibility and naturalness. We randomly selected ten testing utterances from unseen speakers DF0026, DF0028, DM0025, DM0029, and DM0031, respectively. The Naturalness MOS Score of these four systems is shown in Table IV. It shows that S2 and S4 systems, which adopt the same-gender-retrieval strategy, achieve higher Mean Opinion Scores (MOS) for speech naturalness compared to systems utilizing other strategies. The naturalness MOS score of System S3 is lower than that of System S1, a decrement attributed to the necessity of gender conversion in the second stage (Stage 2), which increases the complexity of the voice conversion process. System S4, however, not only restores the original speaker's timbre but also achieves the highest level of speech naturalness among all the systems evaluated.

IV. CONCLUSIONS

This paper proposes a two-stage speech conversion method aimed at improving the reconstruction quality of dysarthric speech. In the first stage, we employ a KNN-VC method based on a same-gender-retrieval strategy to perform initial repair of the dysarthric speech. Subsequently, in the second stage, we utilize the so-vits-svc model to further restore the speaker's timbre and optimize the sound quality of the speech repaired in the first stage. To evaluate the zero-shot performance of the proposed method, we conducted tests on speech data from unknown speakers. Both objective and subjective experimental results show that the proposed approach can achieve some improvements in terms of speaker similarity, speech intelligibility and naturalness.

ACKNOWLEDGMENT

This research is funded by the DKU foundation project "Interdisciplinary Signal Processing Technologies".

REFERENCES

[1] Aihara, R., Takashima, R., Takiguchi, T., Ariki, Y.: 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. pp. 1–4. IEEE (2012)

[2] Aihara, R., Takashima, R., Takiguchi, T., Ariki, Y.: Gmm-based emotional voice conversion using spectrum and prosody features. American Journal of Signal Processing **2**(5), 134–138 (2012)

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

[4] Aihara, R., Takiguchi, T., Ariki, Y.: Phoneme-discriminative features for dysarthric speech conversion. In: Interspeech. pp. 3374–3378 (2017)

[5] Baas, M., van Niekerk, B., Kamper, H.: Voice conversion with just nearest neighbors. arXiv preprint arXiv:2305.18975 (2023)

[6] Chen, L.W., Lee, H.Y., Tsao, Y.: Generative adversarial networks for unpaired voice transformation on impaired speech. arXiv preprint arXiv:1810.12656 (2018)

[7] Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., et al.: Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing **16**(6), 1505–1518 (2022)

[8] Chen, X., Wang, Y., Wu, X., Wang, D., Wu, Z., Liu, X., Meng, H.: Exploiting audio-visual features with pretrained av-hubert for multi-modal dysarthric speech reconstruction. arXiv preprint arXiv:2401.17796 (2024)

[9] Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622 (2018)

[10] Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)

[11] Fu, S.W., Li, P.C., Lai, Y.H., Yang, C.C., Hsieh, L.C., Tsao, Y.: Joint dictionary learning-based non-negative matrix factorization for voice conversion to improve speech intelligibility after oral surgery. IEEE Transactions on Biomedical Engineering **64**(11), 2584–2594 (2016)

[12] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

[13] Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)

[14] Kain, A.B., Hosom, J.P., Niu, X., Van Santen, J.P., Fried-Oken, M., Staehely, J.: Improving the intelligibility of dysarthric speech. Speech communication **49**(9), 743–759 (2007)

[15] Kim, J., Kong, J., Son, J.: Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In: International Conference on Machine Learning. pp. 5530–5540. PMLR (2021)

[16] Kong, J., Kim, J., Bae, J.: Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. Advances in neural information processing systems **33**, 17022–17033 (2020)

[17] Mohammadi, S.H., Kain, A.: An overview of voice conversion systems. Speech Communication **88**, 65–82 (2017)

[18] Okabe, K., Koshinaka, T., Shinoda, K.: Attentive statistics pooling for deep speaker embedding. arXiv preprint arXiv:1803.10963 (2018)

[19] Oyamada, K., Kameoka, H., Kaneko, T., Ando, H., Hiramatsu, K., Kashino, K.: 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). pp. 182–188. IEEE (2017)

[20] Qin, X., Li, N., Lin, Y., Ding, Y., Weng, C., Su, D., Li, M.: The dku-tencent system for the voxceleb speaker recognition challenge 2022. arXiv preprint arXiv:2210.05092 (2022)

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

[22] Stanton, D., Shannon, M., Mariooryad, S., Skerry-Ryan, R., Battenberg, E., Bagby, T., Kao, D.: Speaker generation. In: IEEE ICASSP. pp. 7897–7901. IEEE (2022)

[23] Wang, D., Yu, J., Wu, X., Liu, S., Sun, L., Liu, X., Meng, H.: 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). pp. 7744–7748. IEEE (2020)

[24] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., et al.: Tacotron: Towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135 (2017)

[25] Yamagishi, J., Veaux, C., King, S., Renals, S.: Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction. Acoustical Science and Technology **33**(1), 1–5 (2012)

[26] Yao, Z., Wu, D., Wang, X., Zhang, B., Yu, F., Yang, C., Peng, Z., Chen, X., Xie, L., Lei, X.: Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. In: Proc. Interspeech. IEEE, Brno, Czech Republic (2021)

[27] Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4133–4141 (2017)

[28] Zhang, Y., Liu, G., Lei, Y., Chen, Y., Yin, H., Xie, L., Li, Z.: Promptspeaker: Speaker generation based on text descriptions. arXiv preprint arXiv:2310.05001 (2023)