Two-stage Voice Conversion for Dysarthric Speech Reconstruction with Speaker Identity Preservation

Devin Chang[1], Ming-Chi Yen[2], Hsin-Te Hwang[1], Fo-Rui Li[1,3], Ching-Feng Liu[4], Yu Tsao[1], and Hsin-Min Wang[2]

[1] Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
devchang918@gmail.com, {hwanght,yu.tsao}@citi.sinica.edu.tw
[2] Institute of Information Science, Academia Sinica, Taipei, Taiwan
{ymchiqq,whm}@iis.sinica.edu.tw
[3] National Central University, Taoyuan, Taiwan
f323red@gmail.com
[4] Chi Mei Hospital, Tainan, Taiwan
wtcen@stust.edu.tw

Abstract. Dysarthria, a speech disorder that hinders speech production, can cause difficulties in daily communication for patients. To assist dysarthric patients, voice conversion (VC) methods for dysarthric speech reconstruction (DSR) have been implemented to improve the intelligibility of dysarthric speech. Despite the success of various VC methods, the identity of the dysarthric speaker is often lost during conversion. In this study, we proposed a two-stage VC framework. For Stage 1, we evaluated the ability of a sequence-to-sequence Voice Transformer Network (VTN) and a locally linear embedding (LLE)-based VC method to enhance speech intelligibility by converting dysarthric speech to four normal reference speakers. We tested the LLE conversion with different types of content features (Chinese-HuBERT, WavLM, and Whisper). For Stage 2, we investigated five state-of-the-art approaches for restoring speaker identity (Seed-VC, FreeVC, SPARC, MKL-VC, Vevo) and compared them with a previously tested vector-quantized variational autoencoder (VQ-VAE) baseline model. Our evaluation found that the VTN and LLE using Chinese-HuBERT as content features achieved the most significant improvements in intelligibility. Furthermore, Seed-VC considerably outperformed the other state-of-the-art models and the baseline in preserving speaker similarity without sacrificing intelligibility, especially when applied to VTN results.

Keywords: Dysarthric Speech Reconstruction, Voice Conversion, Speaker Identity.

1 Introduction

Dysarthria is a speech disorder characterized by muscle impairments involved in speech production, arising from neurological diseases such as Parkinson's dis-

ease (PD) and amyotrophic lateral sclerosis (ALS), as well as physical disabilities such as traumatic brain injury [10]. Due to impaired muscular structures, patients with dysarthria suffer from inaccuracy in pronunciation, difficulty in articulation, and slowness in speech, affecting their daily interaction and communication with others. Therefore, it is crucial to explore effective methods to facilitate communication and improve the quality of life of individuals with dysarthria.

To assist dysarthric patients, dysarthric speech reconstruction (DSR) serves as an innovative and popular approach to improve the intelligibility and naturalness of dysarthric speech. Current advances in DSR involve voice conversion (VC) techniques, which convert speech properties of a source speaker to that of the target speaker while maintaining linguistic content [22]. Among VC systems, rule-based and statistical VC models are often investigated in DSR tasks [22]. Rule-based VC models perform VC by adjusting speech frequency or temporal properties under a set of given rules, which have limitations since adjustments to rules are needed to account for individual variations [20]. In contrast, traditional statistical VC, such as the Gaussian mixture model [9], converts dysarthric to normal speech via acoustic feature mapping, but the method did not significantly increase speech intelligibility. Many recent VC research in DSR focus on deep learning (DL) methods. For instance, DL techniques such as sequence-to-sequence (seq2seq) based text-to-speech (TTS) synthesis [24] and generative adversarial networks (GAN) based VC models [3] were evaluated on DSR tasks, yielding considerable improvements in intelligibility of dysarthric speech. However, one drawback with DL methods is that they require large amount of training data, which is often unavailable for dysarthric speech in real settings.

Besides increasing speaker intelligibility, another difficulty to apply VC technique in DSR is speaker identity preservation. Due to the lack of normal speech from dysarthric patients, voices of dysarthric patients are frequently converted to another normal speaker in VC, thereby changing the speaker identity. To address this issue, a two-stage VC framework was proposed to enhance the intelligibility and naturalness of dysarthric speech while preserving speaker identity [7]. A seq2seq-based Voice Transformer Network (VTN) [6] first performs dysarthric-to-normal VC to improve intelligibility and naturalness, and then a vector-quantized variational autoencoder (VQ-VAE) [17] restores the dysarthric speaker's identity without altering the improved intelligibility. However, the experimental results indicate that the VQ-VAE conversion in Stage 2 substantially decreases the intelligibility gained from Stage 1, while providing limited improvement in speaker similarity.

In this study, we surveyed various VC approaches suitable for the two-stage VC framework that enhances intelligibility while preserving the speaker identity in the DSR task. In Stage 1, we re-evaluated the seq2seq-based VTN model [6] on our dataset for DSR and examined locally linear embedding (LLE) [13] as a new potential method for comparison. For LLE, we converted dysarthric speech into voices of multiple reference speakers with content features extracted from Chinese-HuBERT Large [5], WavLM Large [1], or Whisper Large v3 [19] and

reconstructed the speech via a HiFi-GAN vocoder [12]. Unlike DL methods, LLE does not require a large training dataset, which is advantageous in experiments with scarce dysarthric speech data. For Stage 2, we surveyed six different VC approaches: Seed-VC [15], FreeVC [14], SPARC [2], MKL-VC [16], Vevo [31], and a baseline VQ-VAE [17], to restore the identity of the dysarthric speaker.

As a result, we found that LLE conversion with content features extracted from Chinese-HuBERT has the lowest character error rate (CER), whereas VTN results have the lowest syllable error rate (SER). Since LLE with Chinese-HuBERT features performed the best in improving the intelligibility of dysarthric speech compared to other content features, we selected its converted speech to conduct Stage 2 experiments along with VTN results. Objective and subjective evaluation of Stage 2 results showed that VTN followed by Seed-VC conversion outperformed other combinations in preserving speaker similarity with minimal impact on improved CER and SER.

2 Related Work

2.1 Locally Linear Embedding-based Voice Conversion (LLE-VC)

In our experiment, we implemented an improved locally linear embedding-based VC (LLE-VC) [13], which is originally designed for electrolaryngeal VC (ELVC). Different from many statistical VC methods, LLE-VC does not require large training data due to the absence of model training. The LLE-VC consists of four stages: 1.)neighbor searching in the source speech domain, 2.)LLE weight estimation, 3.)feature reconstruction from aligned normal speech exemplars, and 4.)waveform synthesis. In their study, prosody-rich WavLM-6th layer features were used for weight estimation and reconstruction, whereas features extracted from the same or different model (Chinese-HuBERT, Whisper, etc.) were used for alignment and neighbor search. The integration of various foundation models contributed to the flexibility that increases intelligibility and stability of ELVC. Through experiments with a Mandarin EL/NL corpus, objective and subjective evaluation showed that using Chinese-HuBERT or Whisper as content features for LLE-VC significantly reduced CER and improved MOSA-Net metrics. The effectiveness of using Chinese-HuBERT features implied the importance of Mandarin-specific pretraining in Mandarin VC tasks. The results confirmed that using content features from robust models for neighbor searching enhanced LLE-VC framework with better content representation.

Similar to EL speech, dysarthric speech often exhibits reduced intelligibility and abnormal speech properties. Despite the success observed in ELVC tasks, little to no research has been done on DSR with exemplar-based VC, so it is worth assessing its effectiveness on DSR tasks. Therefore, we adopted LLE-VC in our study and selected the top 3 best content features (Chinese-HuBERT, WavLM, and Whisper) from [13] for DSR. With content-rich representation for neighbor searching while keeping prosody-rich features for weight estimation, we intended to improve the intelligibility of dysarthric speech through LLE-VC without the need for large training data.

2.2 Sequence-to-sequence voice conversion (seq2seq VC)

In Stage 1, a seq2seq-based VTN model transfers knowledge learned from a pretrained TTS model, enabling VC with high intelligibility and fidelity [6]. The model has the ability to capture long-term dependencies, generate variable length output, and perform any-to-one (A2O) VC with wav2vec [8]. Besides the model's notable performance in other VC tasks, previous research experiments showed that seq2seq-based VTN model could considerably increase intelligibility of dysarthric speech and account for the abnormal speech rate [7]. Thus, it is worth reaffirming the capability of the seq2seq-based VTN model in reconstructing dsyarthric speech and evaluating the results when combined with state-of-the-art methods for speaker identity conversion. In this paper, we refer to the conversion using seq2seq-based VTN model as VTN-VC.

2.3 Two-stage VC for preserving speaker similarity in DSR

The experimental design in this study is mainly based on the two-stage VC framework proposed by [7], which aims to improve the intelligibility of dysarthric speech while preserving speaker identity. In the first stage, VTN-VC increases speech intelligibility by performing dysarthric VC with a reference normal speaker. Then, in the second stage, a nonparallel frame-wise VQ-VAE model maps the speaker's identity back to the dysarthric speaker's voice without altering the intelligibility and naturalness gained from Stage 1. However, the experimental results show that although VTN-VC reduces SER from around 90% to 65%, SER rises back to around 80% after VQ-VAE conversion, indicating the loss of intelligibility gained from Stage 1. Furthermore, VQ-VAE conversion only increases speaker similarity to about 45%, revealing room for improvement in speaker similarity. Therefore, a survey of other possible VC approaches suitable for preserving speaker identity is needed.

2.4 Approaches for Speaker Similarity Preservation

In this study, we surveyed five state-of-the-art VC approaches for speaker similarity preservation in Stage 2: Seed-VC, FreeVC, SPARC, MKL-VC, and Vevo, while VQ-VAE [17] served as the baseline model in our experiment. Often used as voice changers, the other five selected VC approaches were zero-shot or one-shot VC methods proposed after VQ-VAE. Information about these approaches is as follows:

Seed-VC. Seed-VC is a zero-shot VC model with a diffusion transformer that allows the model to capture fine-grained speaker characteristics [15]. It enables semantic features to be extracted from timbre-shifted utterances during training and minimizes timbre leakage when content and timbre come from different speakers. Experimental results show that Seed-VC produces higher speaker similarity than existing methods such as OpenVoice and CosyVoice, highlighting its advantage in maintaining high similarity between timbres.

FreeVC. Built upon the VITS architecture, FreeVC is a one-shot VC model intended for high-quality waveform reconstruction without requiring text annotations [14]. It first extracts content information from self-supervised WavLM features, and removes residual speaker characteristics through an information bottleneck. With a speaker encoder supplying speaker identity, FreeVC is able to perform one-shot VC with a single reference utterance. Experimental results show that FreeVC outperforms existing text-based and text-free models in terms of naturalness and speaker similarity.

SPARC. Speech Articulatory Coding (SPARC) is a neural encoding-decoding model that represents speech through vocal tract kinematics [2]. The articulatory analysis model infers the position of vocal tract structures and other properties such as pitch based on the source speech, and the articulatory synthesis model reconstructs the speech from articulatory features. A speaker identity encoder is trained to account for individual differences. The study shows that SPARC is capable of preserving intelligibility and high speech quality while being generalizable and interpretable.

MKL-VC. Unlike the kNN-VC pipeline, MKL-VC is a training-free model utilizing Monge-Kantorovich Linear optimal transport map instead of kNN regression [16]. This allows the model to deal with uneven variances and improves feature transformation. Experimental results show that MKL-VC achieves better performance than kNN-VC in robustness and content preservation even with a 5-second target speech.

Vevo. Vevo is a VC framework designed for zero-shot imitation of timbre, style, or both [31]. The framework achieves voice imitation by decoupling timbre, style, and linguistic content. Firstly, the source speech is passed to a VQ-VAE [17] to filter out timbre or style information and generate content tokens. The framework contains an autoregressive transformer and an flow-matching transformer to imitate style and timbre, respectively, and each of which can be used independently. The study shows that Vevo matches or outperforms existing methods for accent and emotion conversion. Since the goal for Stage 2 is to restore speaker identity, we used only the flow-matching transformer from Vevo in this study.

3 Proposed Method

In this study, we implemented two-stage VC for DSR. Stage 1 converts dysarthric speech to reference normal speech with either VTN-VC [6] or LLE-VC [13], improving dysarthric speech intelligibility. Then, Stage 2 restores speaker similarity by converting the speaker identity back to the dysarthric speaker with various VC approaches. An illustration of the two-stage VC for DSR is shown in Fig. 1.

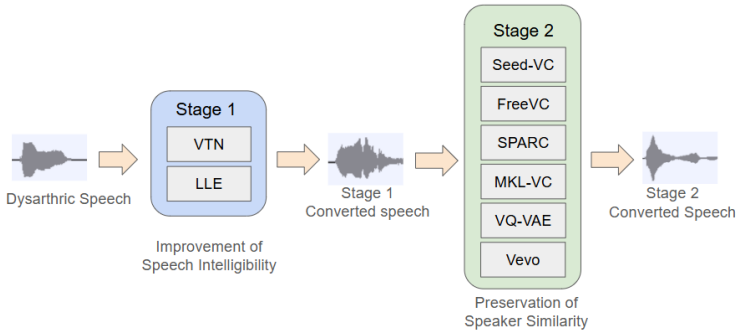


Fig. 1. Illustration of the two-stage VC for DSR that improves speech intelligibility while preserving speaker similarity.

3.1 Datasets

For the dysarthric speech dataset, we obtained audio recordings from a female patient reading 320 ten-character mandarin sentences from the TMHINT dataset [28]. For the reference dataset, we selected two male speakers (SP04, SP09) and two female speakers (SP07, SP13) from the TMSV dataset [4] since they yielded the poorest evaluation results when serving as target speakers in previous experiments [7]. Each reference speaker from the TMSV dataset also has 320 audio recordings of 10-word sentences from the TMHINT corpus. We implemented a 240/40/40 split on the source and reference datasets for training/development/evaluation, and all audio recordings were downsampled to 16kHz. Besides the datasets mentioned above, we used the Sinica-COSPRO dataset [23], 17 speakers from the TMSV dataset, and the dysarthric patient's recording to train the VQ-VAE baseline in Stage 2. Furthermore, the syllable recognition system for SER was trained with the 196-hour MATBN corpus [25].

3.2 Stage 1: Dysarthric Speech Reconstruction with VTN and LLE

In the first stage, we replicated the experiment using VTN-VC with TTS pretraining [6]. The TTS system was pretrained with the multi-speaker, 44-hour long Sinica-COSPRO Mandarin dataset. The implementation of the VTN model was based on ESPnet [27, 26] with configuration posted online[5]. The VTN model used Mel-spectrogram features and Parallel WaveGAN (PWG)[6] for the vocoder.

In addition, LLE-VC [13] was implemented to convert dysarthric speech to each of the TMSV speakers in our reference dataset. For the parallel LLE-VC, we extracted features from the sixth layer of WavLM (i.e. WavLM-6th) for weight estimation. Furthermore, to test the effect of using different types of content features for neighbor search, the LLE-VC was performed and evaluated with

[5] https://github.com/espnet/espnet/tree/master/egs/arctic/vc1
[6] https://github.com/kan-bayashi/ParallelWaveGAN

content features extracted from the last layer of pretrained Chinese-HuBERT Large [5], WavLM Large [1], or Whisper Large v3 [19]. Then, dynamic time warping (DTW) [21] was applied to the source and target matching set to align the content features during conversion. For waveform synthesis, the WavLM-6th features reconstructed and output by LLE-VC were fed into a HiFi-GAN vocoder [12] trained on Sinica-COSPRO and LibriSpeech corpora [18].

3.3 Stage 2: Speaker Identity Restoration

Although VTN-VC or LLE-VC increases the intelligibility of dysarthric speech, the identity of the dysarthric speaker is not preserved during conversion. The methods convert information of not only suprasegmental features but also speaker identity. Therefore, our objective for the second stage was to restore speaker similarity through various VC approaches while maintaining improved intelligibility from Stage 1. To achieve this, we used each set of converted speech from Stage 1 as the source speech for six different VC approaches (Seed-VC [15], FreeVC [14], SPARC [2], MKL-VC [16], Vevo [31], and VQ-VAE [17]) and evaluated the effect of each conversion on intelligibility and speaker similarity. The target for all Stage 2 conversions was the dysarthric speech dataset. Due to the success of these approaches as voice changers, we hypothesize that the second-stage VC can restore the identity of the dysarthric speaker without significantly affecting the high intelligibility of the converted speech from Stage 1. Moreover, since Seed-VC, FreeVC, SPARC, MKL-VC, and Vevo are novel approaches compared to VQ-VAE, we expect these models to perform better than the VQ-VAE basline. In our experiments, we used specific settings for each method.

Seed-VC Settings. We used Seed-VC v1.0[7] with Whisper-small as the content encoder. The output results were converted to 16kHz for objective evaluation. The number of diffusion steps was set to 40 as recommended for best quality.

FreeVC Settings. We used the FreeVC[8] checkpoint pretrained with VCTK dataset [29] to perform our experiments. Other settings were kept as default.

SPARC Settings. Since our datasets were in Mandarin, we used the SPARC[9] checkpoint trained on multilingual data (labeled as "multi"). Other settings were kept as default.

MKL-VC Settings. We used the default checkpoints and settings for all conversions with MKL-VC[10].

[7] https://github.com/Plachtaa/seed-vc
[8] https://github.com/OlaWod/FreeVC
[9] https://github.com/Berkeley-Speech-Group/Speech-Articulatory-Coding
[10] https://github.com/alobashev/mkl-vc

Vevo Settings. For all Vevo experiments, we used only the flow-matching transformer from the Vevo framework[11] to perform timbre conversion exclusively. The checkpoints and settings were kept as default.

VQ-VAE Settings. To reproduce the experiment in [7], we used crank[12], an open-source autoencoder-based VC model, as the baseline for our experiments [11]. The VQ-VAE model was trained with Sinica-COSPRO [23], recordings from the 17 TMSV speakers [4], and the dysarthric patient's voice for 400000 steps. One of the TMSV speakers (SP11) was not included in the experiment due to misspoken sentences. The output audio was synthesized with a PWG neural vocoder trained on the Sinica-COSPRO dataset.

4 Experimental Results

In this study, we conducted objective evaluation in both stages. For Stage 1, we incorporated mel cepstral distortion (MCD), which indicates the spectral distortion between the generated and the target speech. The mel cepstral coefficient from order 0 were removed. Furthermore, we measured the pitch accuracy of converted speech and similarity between the F0 contours by calculating the F0 root mean square error (F0 RMSE) and the F0 correlation coefficient (F0 CORR), respectively, for Stage 1 results. To measure speech intelligibility in both stages, we calculated the CER and SER of the converted speech with automatic speech recognition (ASR) results from Whisper Large v3 [19] and a self-trained Transformer-based Mandarin syllable recognition system, respectively. To show the duration difference between the speech, we also calculated the average absolute duration difference between the converted and target utterances (DDUR). Last but not least, speaker similarity scores were calculated for all converted speech using the metrics tool in the Amphion toolkit[13] [30]. The similarity score is based on the cosine similarity of speaker embeddings calculated by a resemblyzer[14], with the dysarthric speaker's voice as the target speech. We provide audio samples on this demo site[15].

4.1 Investigate DSR via VTN and LLE with Different Content Features

In Stage 1, we converted dysarthric speech to normal speech of 2 male and 2 female target speakers. For each speaker pair, LLE-VC was conducted with WavLM, Chinese-HuBERT, or Whisper as the content feature. Objective and subjective evaluation metrics discussed above were calculated for the converted results. Then, for both VTN-VC and LLE-VC results, we computed the average for each metric across the speaker pairs.

[11] https://github.com/open-mmlab/Amphion/tree/main/models/vc/vevo
[12] https://github.com/k2kobayashi/crank
[13] Amphion GitHub: https://github.com/open-mmlab/Amphion
[14] Resemblyzer GitHub: https://github.com/resemble-ai/Resemblyzer
[15] Demo site: https://devchang918.github.io/two-stage-VC-DSR/

Table 1. Evaluation of DSR Results via VTN-VC, LLE-VC with Different Content Features (WavLM (WL), Chinese-Hubert (CH), and Whisper (WH)), and Stage 2 Methods.

System	MCD↓	F0RMSE↓	F0CORR ↑	DDUR↓	CER (%)↓	SER (%)↓	Similarity↑
Before VC	11.21	60.69	0.03	2.96	93.24	82.50	1.0
Seed-VC	9.19	48.47	0.19	2.95	83.19	87.78	0.56
FreeVC	9.22	49.06	0.10	2.95	108.38	99.10	0.56
SPARC	8.19	53.06	0.07	2.94	92.50	90.90	0.60
MKL-VC	9.65	51.43	0.03	2.95	96.06	94.23	**0.62**
Vevo	10.01	46.94	0.21	2.94	82.00	89.33	0.59
VQ-VAE	9.13	48.93	0.14	2.97	115.75	166.23	0.59
VTN	8.01	39.96	0.43	**0.39**	75.63	**63.03**	0.52
LLE-WL	**6.94**	42.19	0.38	2.95	70.13	80.25	0.58
LLE-CH	**6.94**	43.47	0.35	2.95	**62.13**	77.20	0.58
LLE-WH	6.95	**39.51**	**0.48**	2.97	68.88	83.78	0.59

In Table 1, we present the evaluation of DSR with six Stage 2 VC approaches, VTN-VC, and LLE-VC via different content features. When performing DSR with only the Stage 2 methods, speech intelligibility does not increase since the CER and SER of the converted speech are close to or higher than before conversion. Furthermore, the improvement in MCD and F0 metrics is smaller than with LLE-VC or VTN-VC. Thus, we need LLE-VC or VTN-VC in Stage 1 to improve dysarthric speech intelligibility, while using Stage 2 models only to preserve speaker similarity during post-processing.

Although MCD decreases after both VTN-VC and LLE-VC, there is a larger decline in MCD for LLE-VC (11.21 to 6.9) than VTN-VC (11.21 to 8.01), suggesting greater reduction in spectral distortion with LLE-VC. Moreover, the decrease in F0 RMSE and the increase in F0 CORR for VTN-VC and LLE-VC with different content features are comparable in magnitude, indicating similar improvements in both pitch accuracy and F0 contour similarity. In addition, since VTN-VC alters the length of the speech while LLE-VC does not, VTN-VC results have a DDUR closer to 0 whereas the DDUR for LLE-VC results is almost the same as before conversion (about 2.96).

In general, the metrics in the table indicate that both VTN-VC and LLE-VC using any selected type of content feature improve speech intelligibility. For LLE-VC, using Chinese-HuBERT as the content feature achieves the lowest CER (62.13%) and SER (77.20%), outperforming both Whisper and WavLM. This suggests that using Chinese-HuBERT as the content feature provides more robust linguistic representations for dysarthric speech. However, LLE-VC with Whisper as content feature yields the highest SER (83.78%), indicating greater difficulty for DSR. On the other hand, VTN-VC results show a much lower SER (63.03%) but higher CER (75.63%) compared to LLE-VC results. In terms of speaker similarity, the drop in the similarity score after VTN-VC and LLE-VC shows that speaker identity was not well-preserved. Furthermore, there are no notable differences in the similarity scores for LLE-VC groups (0.58-0.59), suggesting that the choice of content feature in Stage 1 does not influence the

Table 2. Results of two-stage VC for DSR

Stage 1 Model	Stage 2 Model	DDUR↓	CER (%)↓	SER (%)↓	Similarity↑
VTN	Seed-VC	0.39	**77.00**	**66.58**	**0.89**
	FreeVC	0.40	80.06	69.78	0.65
	SPARC	0.40	83.88	73.78	0.75
	MKL-VC	0.40	77.19	68.08	0.75
	Vevo	0.45	79.38	67.40	**0.89**
	VQ-VAE	**0.38**	89.13	82.03	0.78
LLE	Seed-VC	2.93	73.50	**77.10**	**0.93**
	FreeVC	2.93	99.44	94.35	0.72
	SPARC	**2.90**	86.31	94.88	0.78
	MKL-VC	2.93	**70.94**	101.53	0.78
	Vevo	2.91	80.25	80.05	**0.93**
	VQ-VAE	2.95	99.31	98.23	0.77

system's ability in preserving speaker identity. In contrast, the speaker similarity score of VTN-VC results (0.52) is slightly lower than that of LLE-VC results, indicating that more information about speaker identity is lost with VTN-VC than LLE-VC.

Since we did not observe large differences in F0 metrics and speaker similarity with LLE-VC, we chose which converted results to use in Stage 2 based on spectral distortion and speech intelligibility. Therefore, we selected, along with VTN-VC results, the LLE-VC results with Chinese-HuBERT as the content feature due to the largest decrease in MCD, CER, and SER. Although these models effectively increase intelligibility, these results point out that improvements in speaker similarity require further refinement beyond Stage 1.

4.2 Surveying VC Approaches for Speaker Similarity Preservation

In Stage 2, we assessed the performance of Seed-VC, FreeVC, SPARC, MKL-VC, Vevo, and VQ-VAE in preserving speaker similarity. For all target speakers, the output of VTN-VC or LLE-VC with Chinese-HuBERT as content features served as input for each of these approaches, and the results were evaluated with the same metrics from Stage 1 except MCD and F0 metrics. The baseline used for Stage 2 is VTN+VQ-VAE with the same settings as the previous paper [7].

For each VC approach, we report the averages for the metrics across results from all speaker pairs to observe the effectiveness of each model in restoring speaker identity without reducing intelligibility. We also explored how the choice of the Stage 1 model affects the metrics after post-processing. In terms of intelligibility, Seed-VC conversion generally influences CER and SER the least for both LLE-VC and VTN-VC results. For LLE-VC results, although the lowest CER is achieved with MKL-VC (70.94%), implementing Seed-VC in Stage 2 leads to a substantially lower SER (77.10%) compared to other methods, but with CER (73.50%) only slightly higher than that of MKL-VC. Among all second-stage conversions, the improvement in intelligibility obtained from Stage 1 is most retained by applying Seed-VC on VTN-VC output, which only raised the CER by 1.37% and SER by 3.55%. When using LLE-VC in stage 1, Stage 2 methods in

addition to Seed-VC substantially increase the CER or SER, struggling to preserve intelligibility when applied to LLE-converted speech. In terms of speaker similarity, Seed-VC and Vevo substantially outperform other approaches with the largest boost in similarity scores after conversion. Compared to the similarity scores of VTN-VC and LLE-VC results (VTN-VC: 0.52, LLE-VC with Chinese-HuBERT: 0.58), Seed-VC and Vevo restore the similarity scores of VTN-VC and LLE-VC to 0.89 and 0.93, respectively, while other models only increase similarity scores to 0.72-0.78. Overall, the results highlight the effectiveness of Seed-VC in restoring the dysarthric speaker's identity without overly increasing CER and SER. Although VC with Vevo enhances speaker similarity to the same extent as Seed-VC, it leads to a larger increase in CER and SER. Other approaches including FreeVC, SPARC, MKL-VC, and VQ-VAE offer only moderate increase in speaker similarity while sacrificing more intelligibility.

Compared to the VTN+VQ-VAE baseline (CER: 82.03), both VTN+Seed-VC and LLE+Seed-VC frameworks have significantly lower CER and SER. Furthermore, VQ-VAE conversion after VTN-VC increases CER by 13.5% and SER by 19.0%, which is higher than using methods with Seed-VC for post-processing. In terms of speaker similarity, the similarity scores of VTN and LLE converted speech increase by 0.37 and 0.35 after Seed-VC conversion, respectively, outperforming the 0.26 increase for VTN-VC results after VQ-VAE conversion. Thus, all these results suggest that Seed-VC performs the best in restoring speaker similarity while maintaining competitive intelligibility from VTN-VC or LLE-VC relative to other VC approaches and the baseline method.

4.3 Spectrogram Analysis

To obtain a better understanding of the effect of conversions on speech intelligibility and quality, we plot the spectrograms of the same sentence chosen from the source speech, target speech, and the conversion results of each important model. Spectrogram analysis can point out specific words in dysarthric speech that have low intelligibility, the degree of improvement in the intelligibility of those words, and the overall quality of each speech.

In Fig. 2, we observed that dysarthric speech is longer than the reference normal speech, reflecting dysarthric speaker's difficulty in articulation. Since VTN-VC makes adjustments to temporal characteristics of the input speech while LLE-VC does not, the length of the VTN-VC output is more similar to that of the normal speech, whereas the LLE-VC output is closer in length to the dysarthric speech. Furthermore, we observe that the low-mid frequency formants for the 6th and 7th words in the dysarthric speech spectrogram is not clear compared to that of the normal speech, indicating that the intelligibility of these two words is low in dysarthric speech. From the spectrograms of the converted speech, we observe the increase in resemblance of the low-mid frequency formants for the 6th and 7th words with the normal speech after VTN-VC or LLE-VC with Chinese-HuBERT, implying improvements in intelligibility. In terms of speech quality, the VTN-converted speech has higher quality than the LLE-converted

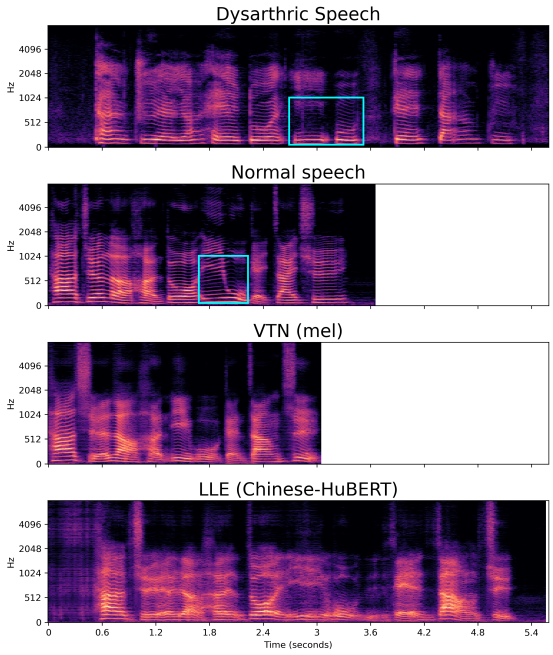


Fig. 2. Mel-spectrograms of dysarthric, normal, VTN-converted (with Mel-spectrogram), and LLE-converted speech (with Chinese-HuBERT as the content feature). The squared region indicates the low-mid frequency formants of the 6th and 7th words in the dysarthric and normal speech, where the formants from the dysarthric speech is less clear than that from the normal speech. Dysarthric and LLE converted speech has longer length, whereas the duration of VTN-converted speech is shorter and closer to the target speech due to seq2seq alignment.

speech since the spectrogram of LLE-converted speech with Chinese-HuBERT as the content feature more blurry than that of the VTN-converted speech.

In Fig. 3, we show the spectrogram of converted speech from Stage 2. Despite the clear low-mid frequency formants for VTN converted speech in Fig. 2, the spectrogram of converted speech from VTN+VQ-VAE exhibits diminished clarity in the formants. In contrast, the low-mid frequency formants in the spectrograms of VTN+Seed-VC and LLE+Seed-VC are less affected, with the spectrogram of VTN+Seed-VC resembling that of the normals speech to a greater extent. This indicates that Seed-VC outmatches the baseline VQ-VAE in maintaining the improved intelligibility from VTN-VC or LLE-VC. The spectrogram of the Seed-VC results appears cleaner but not as bright as those of the normal and other converted speech. Since the volume of the Seed-VC output does not differ from other converted audio, this might be because Seed-VC recognizes more features from VTN and LLE converted speech as noise and removes them during denoising. In terms of audio quality, VTN+Seed-VC output has

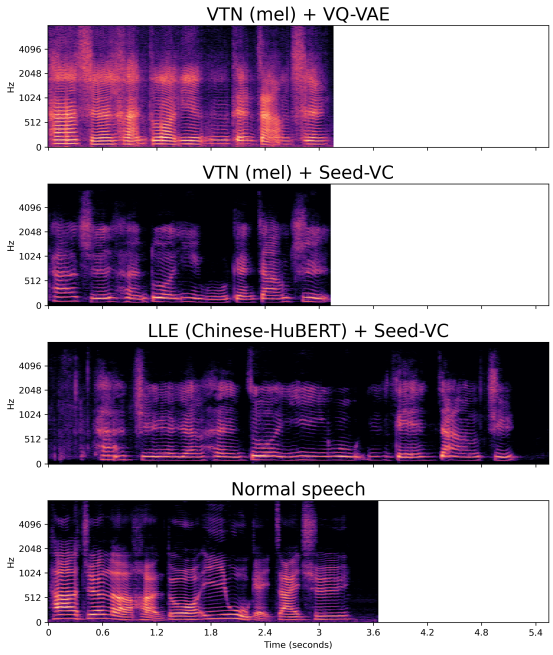


Fig. 3. Mel-spectrograms of second-stage conversion results, which includes VQ-VAE after VTN-VC, Seed-VC after VTN-VC with Mel-spectrogram, and Seed-VC after LLE-VC with Chinese-HuBERT as the content feature. The Mel-spectrogram of the normal speech is included for reference.

quality higher than speech produced by LLE+Seed-VC and the VTN+VQ-VAE baseline, as evidenced by considerably less blurriness.

5 Conclusion

Overall, both VTN+Seed-VC and LLE (Chinese-HuBERT)+Seed-VC significantly improve dysarthric speech intelligibility while preserving speaker similarity, with VTN+Seed-VC retaining the improved CER to a greater extent. In Stage 1, LLE-VC with Chinese-HuBERT as the content feature increases speech intelligibility more than LLE-VC using other content features, as shown in lower CER and SER values. VTN-VC results, however, have the lowest SER overall. In Stage 2, applying Seed-VC after VTN-VC or LLE-VC significantly increases speaker similarity and achieves a high similarity score of 0.89 and 0.93, respectively.

One limitation of this study is that the training data for each Stage 2 approach differ, so the amount of Mandarin training data might vary. This might introduce bias since the experiments were conducted via Mandarin speech. For future work, it is worth investigating methods that can further improve dysarthric

speech intelligibility. In our study, we observe that LLE-VC leads to a greater reduction in CER whereas VTN-VC yields better results for MCD, F0 metrics, and SER. Thus, implementing a framework that takes advantage of both methods may be more beneficial than employing each method alone.

References

1. Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., Wei, F.: 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)
2. Cho, C.J., Wu, P., Prabhune, T.S., Agarwal, D., Anumanchipalli, G.K.: Coding speech through vocal tract kinematics. IEEE Journal of Selected Topics in Signal Processing **18**(8), 1427–1440 (2024)
3. Chu, M., Yang, M., Xu, C., Ma, Y., Wang, J., Fan, Z., Tao, Z., Wu, D.: E-DGAN: An encoder-decoder generative adversarial network based method for pathological to normal voice conversion. IEEE Journal of Biomedical and Health Informatics **27**(5), 2489–2500 (2023)
4. Chuang, S.Y., Wang, H.M., Tsao, Y.: Improved lite audio-visual speech enhancement. IEEE/ACM Transactions on Audio, Speech, and Language Processing **30**, 1345–1359 (2022)
5. Guo, P., Liu, S.: Chinese speech pretrain (2022), https://github.com/TencentGameMate/chinese_speech_pretrain
6. Huang, W.C., Hayashi, T., Wu, Y.C., Kameoka, H., Toda, T.: Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining. In: Proc. Interspeech. pp. 4676–4680 (10 2020)
7. Huang, W.C., Kobayashi, K., Peng, Y.H., Liu, C.F., Tsao, Y., Wang, H.m., Toda, T.: A preliminary study of a two-stage paradigm for preserving speaker identity in dysarthric voice conversion. In: Proc. ICASSP. pp. 1329–1333 (2021)
8. Huang, W.C., Wu, Y.C., Hayashi, T.: Any-to-one sequence-to-sequence voice conversion using self-supervised discrete speech representations. In: Proc. ICASSP. pp. 5944–5948 (2021)
9. 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)
10. Kent, R.D.: Research on speech motor control and its disorders: A review and prospective. Journal of Communication Disorders **33**(5), 391–428 (2000)
11. Kobayashi, K., Huang, W.C., Wu, Y.C., Tobing, P.L., Hayashi, T., Toda, T.: Crank: An open-source software for nonparallel voice conversion based on vector-quantized variational autoencoder. In: Proc. ICASSP. pp. 5934–5938 (2021)
12. Kong, J., Kim, J., Bae, J.: HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis. In: Proc. NIPS. vol. 33, pp. 17022–17033 (2020)
13. Li, F.R., Hwang, H.T., Yen, M.C., Lo, M.T., Tsao, Y., Wang, H.M.: Improving exemplar-based electrolaryngeal speech voice conversion via robust content representations. In: Proc. APSIPA (2025)
14. Li, J., Tu, W., Xiao, L.: FreeVC: Towards high-quality text-free one-shot voice conversion. In: Proc. ICASSP. pp. 1–5 (2023)

15. Liu, S.: Zero-shot voice conversion with diffusion transformers (2024), https://arxiv.org/abs/2411.09943
16. Lobashev, A., Yermekova, A., Larchenko, M.: Training-free voice conversion with factorized optimal transport (2025), https://arxiv.org/abs/2506.09709
17. van den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning. In: Proc. NIPS. p. 6309–6318 (2017)
18. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: An ASR corpus based on public domain audio books. In: Proc. ICASSP. pp. 5206–5210 (2015)
19. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: Proc. ICML (2023)
20. Rudzicz, F.: Adjusting dysarthric speech signals to be more intelligible. Computer Speech Language **27**, 1163–1177 (09 2013)
21. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing **26**(1), 43–49 (1978)
22. Sisman, B., Yamagishi, J., King, S., Li, H.: An overview of voice conversion and its challenges: From statistical modeling to deep learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 132–157 (2021)
23. Tseng, C.Y., Cheng, Y.C., Chang, C.H.: Sinica COSPRO and toolkit – corpora and platform of Mandarin Chinese fluent speech. https://api.semanticscholar.org/CorpusID:204792394 (2005)
24. 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: Proc. ICASSP. pp. 7744–7748 (2020)
25. Wang, H.M., Chen, B., Kuo, J.W., Cheng, S.S.: MATBN: A Mandarin chinese broadcast news corpus. International Journal of Computational Linguistics & Chinese Language Processing **10**(2), 219–236 (Jun 2005), https://aclanthology.org/O05-3004
26. Watanabe, S., Boyer, F., Chang, X., Guo, P., Hayashi, T., Higuchi, Y., Hori, T., Huang, W.C., Inaguma, H., Kamo, N., Karita, S., Li, C., Shi, J., Subramanian, A.S., Zhang, W.: The 2020 ESPnet update: New features, broadened applications, performance improvements, and future plans. In: Proc. DSLW. pp. 1–6 (2021)
27. Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Enrique Yalta Soplin, N., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A., Ochiai, T.: ESPnet: End-to-end speech processing toolkit. In: Proc. INTERSPEECH. pp. 2207–2211 (2018)
28. Wong, L., Soli, S., Liu, S., Han, N., Huang, M.W.: Development of the Mandarin hearing in noise test (mhint). Ear and hearing **28**, 70S–74S (05 2007)
29. Yamagishi, J., Veaux, C., MacDonald, K.: CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (2019). https://doi.org/10.7488/ds/2645
30. Zhang, X., Xue, L., Gu, Y., Wang, Y., Li, J., He, H., Wang, C., Liu, S., Chen, X., Zhang, J., Fang, Z., Chen, H., Tang, T.Y., Zou, L., Wang, M., Han, J., Chen, K., Li, H., Wu, Z.: Amphion: an open-source audio, music, and speech generation toolkit. In: Proc. SLT. pp. 879–884 (2024)
31. Zhang, X., Zhang, X., Peng, K., Tang, Z., Manohar, V., Liu, Y., Hwang, J., Li, D., Wang, Y., Chan, J., Huang, Y., Wu, Z., Ma, M.: Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement (2025), https://arxiv.org/abs/2502.07243