Improving the Efficiency of Dysarthria Voice Conversion System Based on Data Augmentation

Wei-Zhong Zheng, Ji-Yan Han, Chen-Yu Chen, Yuh-Jer Chang, and Ying-Hui Lai, *Member, IEEE*

***Abstract*— Dysarthria, a speech disorder often caused by neurological damage, compromises the control of vocal muscles in patients, making their speech unclear and communication troublesome. Recently, voice-driven methods have been proposed to improve the speech intelligibility of patients with dysarthria. However, most methods require a significant representation of both the patient's and target speaker's corpus, which is problematic. This study aims to propose a data augmentation-based voice conversion (VC) system to reduce the recording burden on the speaker. We propose dysarthria voice conversion 3.1 (DVC 3.1) based on a data augmentation approach, including text-to-speech and StarGAN-VC architecture, to synthesize a large target and patient-like corpus to lower the burden of recording. An objective evaluation metric of the Google automatic speech recognition (Google ASR) system and a listening test were used to demonstrate the speech intelligibility benefits of DVC 3.1 under free-talk conditions. The DVC system without data augmentation (DVC 3.0) was used for comparison. Subjective and objective evaluation based on the experimental results indicated that the proposed DVC 3.1 system enhanced the Google ASR of two dysarthria patients by approximately [62.4%, 43.3%] and [55.9%, 57.3%] compared to unprocessed dysarthria speech and the DVC 3.0 system, respectively. Further, the proposed DVC 3.1 increased the speech intelligibility of two dysarthria patients by approximately [54.2%, 22.3%] and [63.4%, 70.1%] compared to unprocessed dysarthria speech and the DVC 3.0 system, respectively. The proposed DVC 3.1 system offers significant potential to improve the speech intelligibility performance of patients with dysarthria and enhance verbal communication quality.**

***Index Terms*— Deep learning, dysarthric patient, phonetic posteriorgram, voice conversion.**

Manuscript received 12 June 2023; revised 30 September 2023 and 28 October 2023; accepted 31 October 2023. Date of publication 8 November 2023; date of current version 30 November 2023. This work was supported by the Ministry of Science and Technology, Taiwan, under Grant NSTC 110-2218-E-A49A-501 and Grant NSTC 111-2221-E-A49-041-MY2. *(Corresponding author: Ying-Hui Lai.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Research Ethics Review Committee TMU-JIRB of Taipei Medical University Hospital under Approval No. N201607030.

Wei-Zhong Zheng, Ji-Yan Han, Chen-Yu Chen, and Yuh-Jer Chang are with the Department of Biomedical Engineering, National Yang Ming Chiao Tung University, Taipei 11221, Taiwan.

Ying-Hui Lai is with the Department of Biomedical Engineering and the Medical Device Innovation & Translation Center, National Yang Ming Chiao Tung University, Taipei 11221, Taiwan (e-mail: yh.lai@nycu.edu.tw).

Digital Object Identifier 10.1109/TNSRE.2023.3331524

I. Introduction

ACCORDING to a study by the American Speech-Language-Hearing Association, dysarthria is primarily caused by stroke, cerebrovascular accidents (CVA), tumors, cerebral palsy (CP), and other diseases [1]. For example, CVA patients often result in flaccid, spastic, and mixed-type dysarthria in patients. However, CP patients may suffer from spastic, ataxic, hyperkinetic, and mixed-type dysarthria. Both conditions result in patients being unable to flexibly control the muscles used for speech production and generating unclear speech, difficulties in consonant pronunciation, shortness of breath, or fatigue while speaking. This, in turn, leads to a decrease in the intelligibility, audibility, naturalness, and communicative efficiency of speech for individuals with speech disorders, making it difficult for listeners (or machines) to understand [2]. Many augmentative and alternative communication (AAC) devices have been developed [1], such as eye-tracking systems [3] and communication boards [4]. These devices use limbs, eyes, etc. to control devices to spell out words and form sentences to help people with dysarthria pursue better communication methods. However, most patients with cranial nerve damage are unable to use such technologies efficiently because of physical disorders or tremor problems. In addition, previous studies have indicated that the speech rate with AAC devices is slower (2–10 words per minute) [5], [6], [7] than that of dysarthria patients who can still speak and voice their opinions. Therefore, these classical AAC systems still have room for improvement.

Voice-driven AAC systems have recently been proposed to improve dysarthric patients' speech intelligibility using the voice conversion (VC) method, which is defined as dysarthria voice conversion (DVC) in this study. The idea of DVC is to convert distorted speech into normal speech using a conversion model to improve intelligibility for listeners [8]. For example, Hosom et al. [9] proposed a VC architecture that clarifies dysarthric speech by adjusting the prosody and formants of dysarthric speech to be more similar to a normal speaker's rhythm. The experiment also proved that the dysarthric speech recognition rate was improved by 19%. Tolba and El_Torgoman [10] proposed a GMM-based VC method with linear coding of prediction coefficients by analyzing the patient's speech envelope, which also successfully improved the speech recognition rates of dysarthric

speech. Fu et al. [11] proposed a joint dictionary learning-based non-negative matrix VC method to improve the speech intelligibility of surgical patients who have had parts of their articulators removed. The method was demonstrated to not only improve the short-term objective intelligibility score (standardized objective intelligibility evaluation index) significantly, but also perform comparably to traditional VC architecture.

Moreover, many recent studies have shown that, compared to traditional algorithms, deep learning-based VC methods can further improve the speech intelligibility of patients with dysarthria. For example, Chen et al. [12] proposed a deep learning-based architecture to convert dysarthric speech using the log-power spectra [13] and Mel frequency cepstral coefficient (MFCC) [14]. The proposed system improved speech intelligibility compared to baseline systems. Yang and Chung [15] utilized a generative adversarial network (GAN) in this context and improved the speech recognition rate by 33.4%. However, there is still room for improvement in terms of the temporal variability and instability of patients' speech. We recently proposed a deep learning-based DVC system (DVC 3.0) [16]. The system addresses the variability of patients' speech characteristics using the speaker-independent property of phonetic posteriorgrams (PPGs) [17], [18] and converts phonemes into normal speech using a gated convolutional neural network model (gated CNN) with long-term memory effects. The listening test revealed that the proposed system exhibited higher speech intelligibility scores with fewer parameters than baseline systems during duplication (i.e., when the dysarthric patient repeated sentences in the training set). However, the accuracy and stability of the DVC 3.0 system rely heavily on the accuracy of the PPGs that are extracted by the acoustic model, which requires a large representative corpus to cover all possible phonemes. Thus, although DVC 3.0 can currently assist patients in performing well on repetitive sentences, a substantial amount of language data recording by both patients and target language speakers is still required to handle unfamiliar sentences. Additionally, the system requires a large amount of patient speech to participate in training, which can be burdensome for patients. Therefore, this will also result in difficulties for users in terms of usability. In this context, the purpose of this research is to develop a training corpus augmentation method for patients and corresponding target speakers; furthermore, the aim of the training data is to enable DVC 3.0 to convert the correct speech accurately even in a free-talk situation (in which any words can be uttered with no restraints), without further recording data from patients.

Many data augmentation and data generation methods have been proposed. For example, Vachhani et al. [19] adjusted the speech rate and rhythm of normal speech to make it similar to that of dysarthric patients to increase the amount of training data. The study results showed that this data augmentation method improved the recognition rate of the proposed system by 4.24%. Shor et al. [20] used a large amount of normal human speech data as the pre-training weight of the training model and then performed training using a small amount of patient speech. This method improved the training effectiveness of dysarthria speech recognition architectures. Jiao et al. [21] proposed an adversarial training model to convert normal speech into dysarthric speech, improving the speech recognition rate by approximately 10%. Although these methods have improved the performances of several classical VC systems, they require a significant amount of normal speech to be recorded, which is burdensome for the speakers. Further, they generally adopt one-to-one (i.e., single normal speaker and single patient speaker) conversion to augment the patient-like training data, which limits the diversity of the augmentation data and constrains the benefits of DVC system training. Jin et al. [22] proposed a method based on VAE-GAN and data augmentation for disordered speech recognition. This method enables the system to encode, produce and differentiate synthesized impaired speech successfully. Soleymanpour et al. [23] presented a novel means of using the DNN-HMM model on synthesized dysarthric speech and derived good results in dysarthric speech recognition. In view of this, this study proposes a new data augmentation method based on a many-to-one approach. More specifically, text-to-speech (TTS) [24] technology is used to synthesize the speech of multiple normal target speakers. Subsequently, the many-to-one VC system is used to convert these corpora into a patient-like corpus using augmentation data, and these augmented data are used to train the DVC system. The main purposes of the present study are as follows: First, we propose a patient-like data augmentation approach for the DVC system, known as DVC 3.1, to improve the speech intelligibility of dysarthric speakers using the TTS system and many-to-one VC model. Second, we assess the similarity of the proposed augmentation units and the speech intelligibility of the converted speech achieved by the entire system. Finally, the performance of the proposed DVC 3.1 on free speech is compared with that of well-known baseline DVC systems.

The remainder of this paper is organized as follows: In Section II, the proposed system, DVC 3.1, is introduced. The methodology and experimental design of this study are also discussed in this section. The experimental results are presented and discussed in Section III. Finally, the conclusions are presented in Section IV.

II. Methods

A. Proposed Architecture

Fig. 1 depicts the proposed DVC 3.1 system, which was obtained by modifying the system proposed in our previous study [16]. DVC 3.1 involves four stages: data augmentation, speaker-dependent automatic speech recognition (SD-ASR) [25] training, conversion model training, and conversion. The detailed descriptions of these four stages are provided in the subsections below.

1) Data Augmentation Stage: During data augmentation, the Tacotron2 TTS technology [26] is used to synthesize normal speech $n(t)$ from a large number of texts. The Tacotron2 architecture is illustrated in Fig. 2. It converts textual data into text vectors via character embedding and uses an attention mechanism [27] to achieve an attention relationship between the Mel spectrum and character embedding. Then,

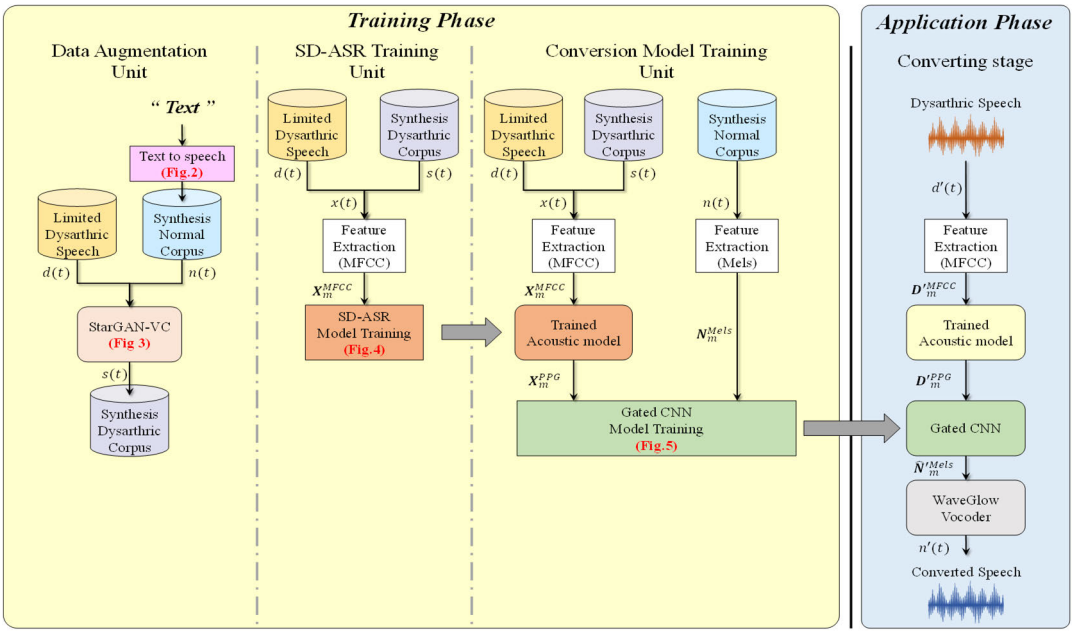

Fig. 1. Proposed DVC 3.1 system. For the training phase, DVC 3.1 separately trains the Data Augmentation Unit, SD-ASR Training Unit, and Conversion Model Training Unit. Note: $n(t)$, $d(t)$, and $s(t)$ represent the waveforms of the synthesized normal speech, dysarthria speech, and synthesized dysarthria speech, respectively.

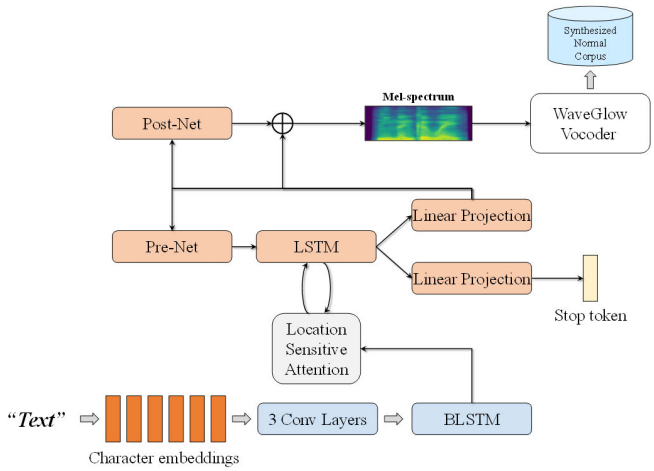

Fig. 2. Tacotron2 [26] TTS method of proposed system in Fig. 1.

the trained model is used to convert the input text into the Mel spectrum. Subsequently, the WaveGlow [28] vocoder is used to generate normal speech from the converted features. For a detailed description of WaveGlow, please refer to [28]. Next, the synthesized normal speech $n(t)$ is used to generate the corresponding paired dysarthric corpus $s(t)$ through the StarGAN-VC [29], [30] model.

Fig. 3 depicts a block diagram of the training process of StarGAN-VC. It adopts the many-to-many speaker VC approach and does not require a paired corpus, which enhances the effectiveness of the data augmentation. More specifically, it uses a single generator $\boldsymbol{G}(\cdot)$ to learn the mapping function for multiple speaker domains and applies the generator for many-to-many data augmentation. To this end, the generator $\boldsymbol{G}(\cdot)$ is designed as a model that can flexibly transform input speech $\boldsymbol{x}_i$ into output speech $\boldsymbol{x}_j$ corresponding to a domain $\boldsymbol{c}_j$ based on random input attributes. The generated speech is denoted by $\boldsymbol{G}(\boldsymbol{x}_i, \boldsymbol{c}_j)$. Note that i, j represent the i^{th} and j^{th} domains, respectively, which can be regarded as the source speaker and conversion target speaker domains, respectively.

To enhance the similarity between the generated data $\boldsymbol{G}(\boldsymbol{x}_i, \boldsymbol{c}_j)$ and real data $\boldsymbol{x}_j$ corresponding to a selected domain $\boldsymbol{c}_j$, a domain classifier $\boldsymbol{C}(\cdot)$ and discriminator $\boldsymbol{D}(\cdot)$, as illustrated in Fig. 3 (b), are adopted. $\boldsymbol{C}(\cdot)$ is used to distinguish input data corresponding to the given domain successfully, whereas $\boldsymbol{D}(\cdot)$ is used to determine whether the input data are real data $\boldsymbol{x}$ or generated data $\boldsymbol{G}(\boldsymbol{x}, \boldsymbol{c})$. Both $\boldsymbol{C}(\cdot)$ and $\boldsymbol{D}(\cdot)$ generate two loss values, namely the domain classification loss ($L_{cls}(\cdot)$) and adversarial loss ($L_{adv}(\cdot)$), based on the error between the recognition result and ground truth, thereby enabling $\boldsymbol{G}(\cdot)$ to generate more realistic data that are conditioned on the target domain. By minimizing $L_{cls}(\boldsymbol{C})$, the domain classifier $\boldsymbol{C}(\cdot)$ correctly classifies the input data into the domains to which they belong. Moreover, the domain classification loss aids the generation of outputs that are more related to the given domain. That is, $\boldsymbol{G}(\cdot)$ minimizes $L_{cls}(\boldsymbol{G})$ to generate data to be classified with respect to the target domain.

The adversarial loss $L_{adv}(\cdot)$ is only applied to $\boldsymbol{G}(\cdot)$ and $\boldsymbol{D}(\cdot)$. During training, $\boldsymbol{D}(\cdot)$ determines whether the input speech is real or generated data by maximizing $L_{adv}(\boldsymbol{D})$. In contrast, $\boldsymbol{G}(\cdot)$ generates $\boldsymbol{G}(\boldsymbol{x}_i, \boldsymbol{c}_j)$, which is as similar to realistic data $\boldsymbol{x}_j$ as possible, to fool the discriminator by minimizing $L_{adv}(\boldsymbol{G})$. Based on the interactions of the two components, $\boldsymbol{G}(\cdot)$ learns to construct a mapping function that can generate realistic speech features based on any input data.

However, training $\boldsymbol{G}(\cdot)$ using the $L_{adv}(\boldsymbol{G})$ and $L_{cls}(\boldsymbol{G})$ losses does not guarantee that $\boldsymbol{G}(\cdot)$ will preserve the correct linguistic information of the input speech while changing only the domain-related parts of the inputs. Thus, the cycle consistency loss $L_{cyc}(\boldsymbol{G})$ and identity mapping loss $L_{id}(\boldsymbol{G})$ are

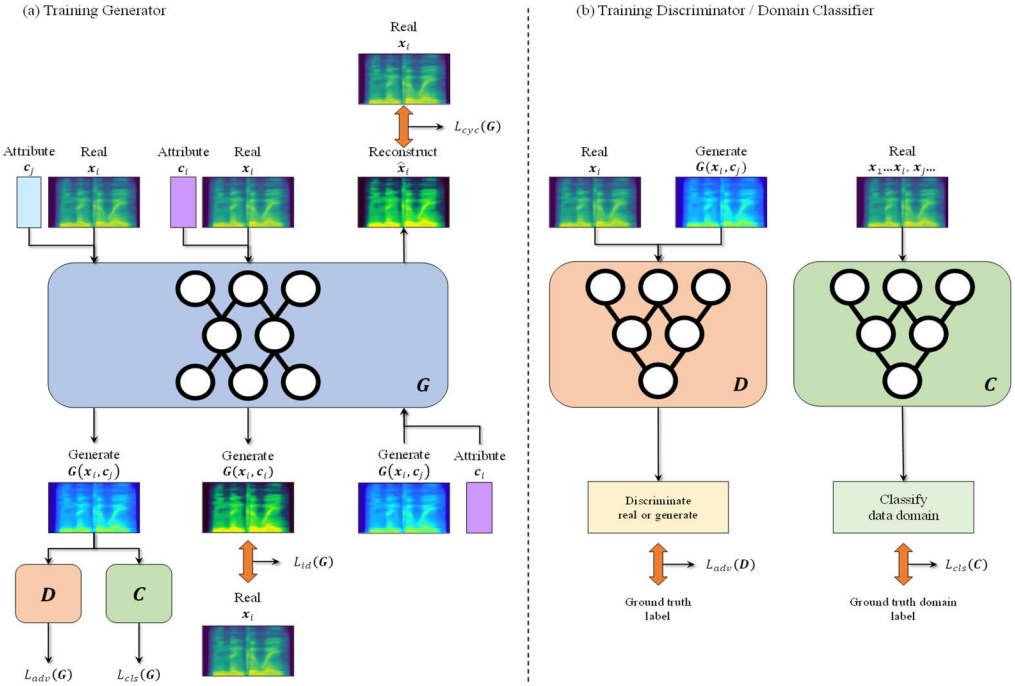

Fig. 3. StarGAN-VC block diagram of proposed system in Fig. 1.

also adopted. The cycle consistency loss encourages $\boldsymbol{G}(\cdot)$ to generate the corpus based on the linguistic information of the input data, where $\boldsymbol{G}(\cdot)$ moves the translated $\boldsymbol{G}(\boldsymbol{x}_i, \boldsymbol{c}_j)$ backwards to the original domain $\boldsymbol{c}_i$ to reconstruct the original input $\boldsymbol{x}_i$. That is, minimizing $L_{cyc}(\boldsymbol{G})$ ensures that $\boldsymbol{G}(\cdot)$ learns to generate speech based on the input content. Furthermore, to prevent $\boldsymbol{G}(\cdot)$ from making unnecessary transformations when the transformation target domain is identical to the source domain, the identity mapping loss $L_{id}(\cdot)$ is employed. In summary, the complete objective function is given by:

$$\begin{aligned} L_G(\boldsymbol{G}) &= L_{adv}(\boldsymbol{G}) + \lambda_{cyc} L_{cyc}(\boldsymbol{G}) + \lambda_{cls} L_{cls}(\boldsymbol{G}) \\ &\quad + \lambda_{id} L_{id}(\boldsymbol{G}) \\ L_D(\boldsymbol{D}) &= L_{adv}(\boldsymbol{D}) \\ L_C(\boldsymbol{C}) &= L_{cls}(\boldsymbol{C}) \end{aligned} \tag{1}$$

Note that $\lambda_{cyc} \geq 0$, $\lambda_{id} \geq 0$, and $\lambda_{cls} \geq 0$ represent the importance of the cycle consistency loss, identity mapping loss, and domain classification loss relative to the adversarial losses, respectively. By completing the optimization of the aforementioned objective function, the generator in StarGAN-VC performs VC in a many-to-many manner and, in combination with TTS (Fig. 2), generates a vast dysarthric-like training corpus.

2) SD-ASR Training Stage: As discussed in our previous study [16], during the SD-ASR training stage, a feature extractor is trained and the acoustic features of dysarthric patients are normalized to linguistic features. Finally, the linguistic features are converted into textual format as output. The SD-ASR training scheme is depicted in detail in Fig. 4.

The speech of one dysarthric speaker and a large amount of synthesized speech of dysarthric speakers are used to train the SD-ASR system. The detailed training approach can be found in [16] and [31]. After completing the SD-ASR training,

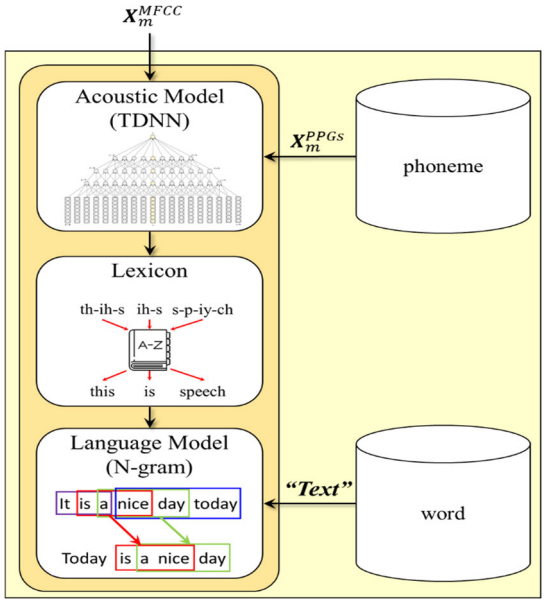

Fig. 4. SD-ASR block diagram of proposed system in Fig. 1.

the well-trained acoustic model in the SD-ASR system is applied as the feature extractor in the DVC system, which uses the well-known time-delay neural network (TDNN) to learn the contextual information between speech to classify highly accurate phonemes. For a more detailed technical description, please refer to [32] and [33].

3) Conversion Model Training Stage and Conversion Stages: During the conversion model training stage, a paired corpus (i.e., paired dysarthric utterances and normal utterances) is used for the gated CNN model training (Fig. 5). The time-domain speech signals of the dysarthric corpus $d(t)$ and synthesized dysarthric corpus $s(t)$ are used as the training source corpus $x(t)$. These utterances are converted

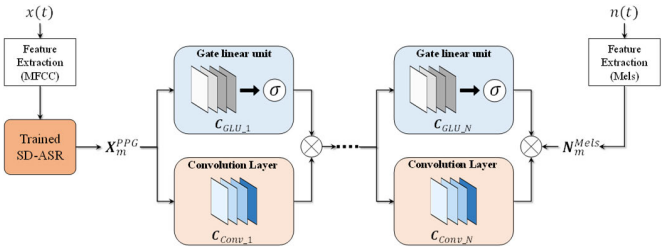

Fig. 5. Gated CNN block diagram of proposed system in Fig. 1.

into 120-dimensional high-resolution MFCCs with dynamic features (40-dimensional original MFCC + 40-dimensional delta features + 40-dimensional delta-delta features), $\boldsymbol{X}_m^{MFCC}$, by the feature extraction (MFCC) unit. The frame size and shift are set to 25 ms and 10 ms, respectively. Subsequently, the trained acoustic model is implemented to extract the 74-dimensional monophone-PPG features, $\boldsymbol{X}_m^{PPGs}$, of the dysarthric speech. Monophone-PPG is a matrix that represents the posterior probability of the phoneme category corresponding to each frame; further details can be found in [16] and [34]. Moreover, the synthesized normal speech $n(t)$ is used as the training target corpus. $n(t)$ is converted by the feature extraction (Mels) unit to obtain the 80-dimensional Mel-spectrum ($\boldsymbol{N}_m^{Mels}$) features. Subsequently, the gated CNN model is further used to learn the mapping function between $\boldsymbol{X}_m^{PPG}$ and $\boldsymbol{N}_m^{Mels}$. The training details of the gated CNN model can be found in our pilot study. [16].

After completing the training phases of the acoustic model and VC model based on the augmented data, the PPG feature $\boldsymbol{D'}_m^{PPG}$ is extracted from the dysarthric speech $d'(t)$ and further converted into the normal Mel spectrum $\boldsymbol{N'}_m^{Mels}$ by the trained gated CNN model. Finally, $\boldsymbol{N'}_m^{Mels}$ is transmitted to the WaveGlow Vocoder and synthesized into clear and intelligible speech. The detailed architecture of WaveGlow can be found in [28].

B. Materials

A total of ten mild-to-severe dysarthria patients and six normal speakers were invited to participate in the recording of the training and testing corpus for this study. Each invited participant recorded 320 utterances twice. Each utterance contained ten characters in Mandarin. All sentences used in the recordings were adopted from the Taiwan Mandarin Hearing in Noise Test [35], which has been used as the listening test in many studies [36], proving that it has sufficient vocabulary and is closely related to daily life. The protocol of the study was approved by the Research Ethics Review Committee TMU-JIRB (N201607030) of Taipei Medical University Hospital and was performed following the principles and policies of the Declaration of Helsinki. Informed written consent for participation was obtained from all participants. The participants were patients with dysarthria between 12 and 80 years of age. The individual biographical data for the dysarthria patients and normal speakers are shown in Tables I and II, respectively.

Six males and four females were invited to record their data in this study. The dysarthric severity of all participants was as follows: three mild, six moderate, and one severe.

TABLE I
BIOGRAPHICAL DATA OF DYSARTHRIA PARTICIPANTS

Participant	Associated disease	Sex	Severity
1	CVA	Male	Moderate
2	CVA	Male	Moderate
3	CVA	Male	Mild
4	CVA	Female	Severe
5	CVA	Female	Mild
6	CVA	Male	Mild
7	CVA	Male	Moderate
8	CVA	Female	Moderate
9	CVA	Male	Moderate
10	CP	Female	Moderate

TABLE II
BIOGRAPHICAL DATA OF NORMAL PARTICIPANTS

Participant	Associated disease	Sex
1	Normal	Male
2	Normal	Male
3	Normal	Female
4	Normal	Female
5	Normal	Male
6	Normal	Female

For subsequent evaluation, one CVA patient with moderate dysarthria (dysarthria participant 9) and one CP patient with moderate dysarthria (dysarthria participant 10) were selected from the ten dysarthria participants as target patients to record 320 additional utterances as a duplicate corpus (1^{st} to 288^{th} utterances) and a free-talk testing corpus (289^{th} to 320^{th} utterances). Meanwhile, one male normal speaker (normal participant 5) and one female normal speaker (normal participant 6) were selected from the six normal participants as corresponding speakers, respectively.

C. Experimental Design

The goal of this study was to propose a dysarthria voice conversion system with a data augmentation architecture to improve the speech intelligibility of dysarthric free speech and reduce the need for a vast dysarthria corpus while training the system. A well-known VC system (namely DVC 3.0) without data augmentation was used as a comparison to demonstrate the benefits of the proposed system. Objective evaluation metrics and a listening test were used to verify the similarity of the synthesized dysarthric corpora and intelligibility of the converted free speech. The training procedure of the model and the evaluation method are described in the following subsections.

1) Model Training: We built separate DVC 3.1 and baseline (DVC 3.0) systems for two target patients to compare the effectiveness of the proposed system.

Table III presents the corpora used by the two aforementioned systems. $\boldsymbol{D}_9^{real}$ and $\boldsymbol{D}_{10}^{real}$ represent the recorded corpus of the two selected moderate dysarthric patients (dysarthric participants 9 and 10, respectively). $\boldsymbol{N}_5^{real}$ and $\boldsymbol{N}_6^{real}$ represent the target pair recorded by two normal speakers (normal

TABLE III
CORPORA USED IN MODEL TRAINING

	DVC 3.0	DVC 3.1
Data Augmentation Unit (TTS)	*No TTS model*	CP model: $\boldsymbol{N}_5^{real}$ CVA model: $\boldsymbol{N}_6^{real}$
Data Augmentation Unit (StarGAN-VC)	*No StarGAN-VC model*	CP model: [$\boldsymbol{D}_9^{real}$, $\boldsymbol{N}_5^{real}$, $\boldsymbol{N}_{1\sim4}^{real}$] CVA model: [$\boldsymbol{D}_{10}^{real}$, $\boldsymbol{N}_6^{real}$, $\boldsymbol{N}_{1\sim4}^{real}$]
SD-ASR Unit (acoustic model)	Model CP: [$\boldsymbol{D}_9^{real}$, $\boldsymbol{N}_5^{real}$, $\boldsymbol{D}_{1\sim8}^{real}$] Model CVA: [$\boldsymbol{D}_{10}^{real}$, $\boldsymbol{N}_6^{real}$, $\boldsymbol{D}_{1\sim8}^{real}$]	CP model: [$\boldsymbol{D}_9^{real}$, $\boldsymbol{D}_9^{syn}$] CVA model: [$\boldsymbol{D}_{10}^{real}$, $\boldsymbol{D}_{10}^{syn}$]
Conversion Unit (gated CNN)	CP model: [$\boldsymbol{D}_9^{real}$, $\boldsymbol{N}_5^{real}$] CVA model: [$\boldsymbol{D}_{10}^{real}$, $\boldsymbol{N}_6^{real}$]	CP model: [$\boldsymbol{D}_9^{real}$, $\boldsymbol{D}_9^{syn}$, $\boldsymbol{N}_5^{syn}$] CVA model: [$\boldsymbol{D}_{10}^{real}$, $\boldsymbol{D}_{10}^{syn}$, $\boldsymbol{N}_6^{syn}$]

$\boldsymbol{D}_{9,10}^{real}$ represents the selected moderate dysarthric patients (dysarthric participants 9 and 10). $\boldsymbol{N}_{5,6}^{real}$ represents the target pair recorded by two normal speakers (normal participants 5 and 6). $\boldsymbol{N}_{1\sim4}^{real}$and $\boldsymbol{D}_{1\sim8}^{real}$ denote the other paired normal corpus (recorded from normal participants 1 to 4) and dysarthric corpus (recorded from dysarthria participants 1 to 8). $\boldsymbol{N}_{5,6}^{syn}$ represents the normal-like speech of normal speakers 5 and 6, generated through TTS. $\boldsymbol{D}_{9,10}^{syn}$ represents the dysarthria-like speech of dysarthric speakers 9 and 10, synthesized using TTS with StarGAN-VC.

TABLE IV
DATA DURATION IN DVC 3.0 AND DVC 3.1 SYSTEMS

	DVC 3.0	DVC 3.1
Real dysarthria corpus duration	Total duration: 20.08 hours per model With speed and perturb augment: 141.08 hours	Total duration: 1.28 hours per model With speed and perturb augment: 8.97 hours
Real dysarthria corpus words	$\boldsymbol{D}_{1\sim8}^{real}$: 320 utterances × 8 participants × 10 words × 2 sets $\boldsymbol{D}_9^{real}$ or $\boldsymbol{D}_{10}^{real}$: 288 utterances × 10 words × 2 sets Total words: 56960 words per model	$\boldsymbol{D}_9^{real}$ or $\boldsymbol{D}_{10}^{real}$: 288 utterances × 10 words × 2 sets Total words: 5760 per model
Synthesized dysarthria corpus duration		Total duration: 1.42 hours With speed and perturb augment: 9.954 hours
Synthesis Dysarthria corpus words		$\boldsymbol{D}_9^{syn}$ or $\boldsymbol{D}_{10}^{syn}$: 320 × 10 words × 2 sets Total: 6400 synthesized words per model

participants 5 and 6, respectively). $\boldsymbol{N}_{1\sim4}^{real}$ and $\boldsymbol{D}_{1\sim8}^{real}$ denote the other paired normal corpus (recorded from normal participants 1 to 4) and dysarthric corpus (recorded from dysarthria participants 1 to 8), respectively. Note that among the aforementioned corpora, only $\boldsymbol{D}_9^{syn}$, $\boldsymbol{N}_5^{syn}$, $\boldsymbol{D}_{10}^{syn}$, and $\boldsymbol{N}_6^{syn}$ were synthesized using the TTS with StarGAN-VC.

The proposed DVC 3.1 requires a well-trained TTS for large-scale normal corpus augmentation. Thus, the recorded corpora $\boldsymbol{N}_5^{real}$ and $\boldsymbol{N}_6^{real}$ of two target normal speakers were used to train the Tacotron2 model of the TTS system. We applied the original model (tacotron2_statedict.pt) released by Tacotron2 as the base model and fine-tuned it for 100,000 epochs. This enabled the TTS to generate normal corpora corresponding to additional input text. Subsequently, the normal speech generated by the TTS ($\boldsymbol{N}_{5,6}^{syn}$) was converted into the corresponding paired patient speech ($\boldsymbol{D}_{9,10}^{real}$) using StarGAN-VC. To take full advantage of the many-to-many conversion ability of StarGAN-VC and enhance the stability of the data augmentation architecture on the generated dysarthric-like corpus, the corpora of four additional normal speakers ($\boldsymbol{N}_{1\sim4}^{real}$) were adopted to train the StarGAN-VC system. The TTS and StarGAN-VC models are trained to enable data augmentation by converting any normal speech corpus into paired synthetic dysarthria-like speech, without the need for alignment. Finally, the synthesized dysarthria-like corpus $\boldsymbol{D}_{9,10}^{syn}$ generated through the Data Augmentation Unit was combined with the real dysarthric speech corpus $\boldsymbol{D}_{9,10}^{real}$ and generated normal speaker speech corpus $\boldsymbol{N}_{5,6}^{syn}$ to serve as the training data for the SD-ASR training and conversion model. The detailed training method is described in Section II-A.2.

In contrast, the baseline system, DVC 3.0, does not use any dysarthric-like data. Instead, it uses a large amount of real patient data ($\boldsymbol{D}_{1\sim10}^{real}$, $\boldsymbol{N}_{5,6}^{real}$) as the training corpus to construct a robust acoustic model. During the training phase of the gated CNN model, only the $\boldsymbol{D}_{9,10}^{real}$ and $\boldsymbol{N}_{5,6}^{real}$ corpora were used as training data. For further details regarding the training of the baseline model, please refer to [16].

Table IV summarizes the total word count and total duration of the dysarthria corpora used by the two systems. DVC 3.0 used a training set from 10 dysarthria speakers to build the acoustic model, totaling approximately 20 hours of dysarthria data. However, the DVC 3.1 architecture only employed two training set from one dysarthria speaker (totaling approximately 1.28 hours) to establish the data augmentation structure and acoustic model.

In the training of the gated CNN model, DVC 3.0 used two sets of real dysarthria data for learning (approximately 1.28 hours), whereas DVC 3.1 employed two set of dysarthria data and two set of synthesized dysarthria data for training (approximately 2.7 hours).

2) Evaluation: We conducted two experiments to demonstrate the benefits of our proposed DVC 3.1 system. The first experiment was used to evaluate the speech recognition performance of the DVC 3.1 system in comparison with the baseline DVC 3.0 system. More specifically, the speech recognition performance was used to investigate the DVC 3.1 performance in human-to-machine and human-to-human communication conditions. For the human-to-machine evaluation, we compared the three different processed sentences, namely dysarthria (i.e., patients' original speech), DVC 3.0, and DVC 3.1, using the Google ASR system. Meanwhile, the average speech recognition rate of these three processed sentences was used to discuss the benefits of DVC 3.1 in the human-to-machine application scenario. We used Google ASR as the human-to-machine benchmark as it is one of the most powerful machine-based ASR systems. For the human-to-human communication evaluation, we invited 13 native Taiwanese Mandarin participants aged between 20 and 25 years (six males and seven females) to perform the listening test. The word correct rate (WCR)

was used to compare the accuracy of each processed sentence. The listening test protocol was approved by a research ethics review conducted by National Yang-Ming University (YM107017E-3). In this experiment, 10 processed utterances were used, which corresponded to unprocessed data (original dysarthria), and two methods were randomly selected for each participant. Each participant was instructed to repeat the sentence (containing 10 words) that they had heard. If the semantic meaning of the repetition was correct, the corresponding pairs of words were assigned one point. The scores corresponding to all three methods were added and the sum corresponding to a specific method was directly proportional to the intelligibility achieved by the method.

Subsequently, we further increased the large synthesized corpus (57600 words, 32 hours per set) generated through Aishell [37] text (the overlapping context was not considered in the test corpus) to retrain two models. The purpose of this evaluation was to verify whether the acoustic models trained through SD-ASR with a massive amount of sentences that patients had never spoken would result in a learning bias. This would lead to the inability to extract the correct dysarthria phonemes, thereby resulting in incorrect conversion in the DVC 3.1 architecture. Therefore, the Aishell synthesized dysarthria-like corpus was only used for the SD-ASR training of DVC 3.1, whereas training for the gated CNN still only included 288 sentences of the target dysarthria corpus and 320 sentences of the synthesized dysarthria-like corpus. As the DVC 3.0 system cannot generate a dysarthria-like corpus, the retraining of SD-ASR used the same Aishell text-generated normal speaker corpus to ensure that the acoustic model had the same phonetic diversity as the retrained DVC 3.1 model. In the training of the gated CNN for DVC 3.0, there were still only two sets of 288 sentences of the target dysarthria corpus. We repeated the above experiments (human-to-machine, human-to-human) to compare the performance between DVC 3.0 and DVC 3.1 in detail.

The second experiment involved evaluating the performance of the synthesized dysarthria-like corpus using our data augmentation approach. Hence, a listening test was used to evaluate the similarity between the synthesized dysarthria-like and real dysarthric speech via subjective testing. A total of 13 subjects (six males and seven females) aged between 20 and 25 years participated in this audiological similarity test in a soundproof room. The testing speech level was calibrated to 65 dB SPL, which followed the standard of the American National Standards Institution (ANSI S3.6) [38]. Before the test, subjects practiced the next test sound to familiarize themselves with the next test process. Thereafter, eight sets of test sentences with the same test situation (normal speaker's speech and synthesized dysarthria-like speech) were played for the subjects to identify whether each sentence was similar to normal speech (or dysarthric speech), and the average score of the eight sentences was calculated as the similarity score.

III. Experimental Results and Discussion

The primary purpose of this study was to propose a data augmentation unit that can synthesize dysarthria-like patients' speech, with the aim of helping users of DVC 3.1

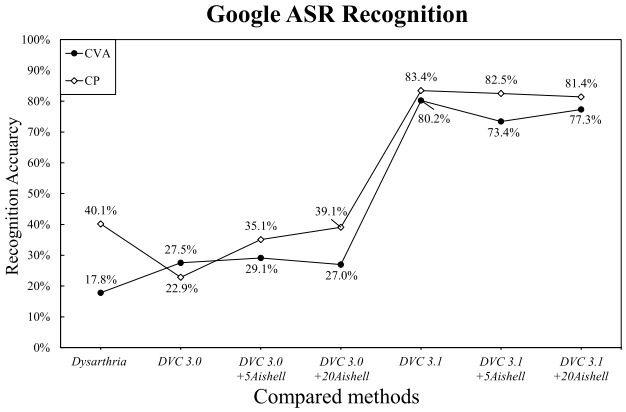

Fig. 6. Average speech recognition rate obtained through Google ASR system in human-to-machine application scenario. The X-axis represents the signal processing approach and the Y-axis represents the speech recognition accuracy obtained using the Google ASR evaluation metric.

to save time in recording while improving the efficiency of the VC model. Therefore, it was necessary to validate whether there was an effective improvement in both the DVC models trained with dysarthria-like speech data and those trained solely on real patient data without dysarthria-like speech. For this purpose, we evaluated the benefits of a synthesized dysarthric-like corpus for the DVC 3.1 system by evaluating the human-to-machine recognition effectiveness using the Google ASR evaluation metric, which was defined as human-to-machine recognition performance in this study. Fig. 6 presents the human-to-machine speech recognition rates of the real dysarthric speech and those processed by the DVC 3.0 and DVC 3.1 systems under free speech testing conditions. In this experiment, seven methods (Dysarthria, DVC 3.0, DVC 3.0+5Aishell, DVC 3.0+20Aishell, DVC 3.1, DVC 3.1+5Aishell, and DVC 3.1+20Aishell) were compared simultaneously. Dysarthria represents the original untreated speech of the patients, whereas DVC 3.0 and DVC 3.1 represent the speech of the patients after being processed by the DVC 3.0 and DVC 3.1 systems, respectively. Additionally, +5Aishell and +20Aishell indicate the inclusion of 5 times and 20 times more of the Aishell corpus for both system training processes, respectively. For dysarthric patients suffering from CVA, the average speech recognition rates of dysarthria, DVC 3.0, and DVC 3.1 were 17.8%, 27.5%, and 80.2% (maximum), respectively. For dysarthric patients suffering from CP, the average speech recognition rates of dysarthria, DVC 3.0, and DVC 3.1 were 40.1%, 22.8%, and 83.4% (maximum), respectively. The results indicated that the Data Augmentation Unit provided benefits for the DVC system in this challenging application scenario. In addition, we further generated 5 times and 20 times more of the Aishell news corpus and retrained the DVC 3.0 and DVC 3.1 systems to verify whether a large amount of synthetic corpus would affect the stability of the system. Note that this corpus did not overlap with any corpus recorded by patients. The results demonstrated that the synthetic corpus generated by StarGAN-VC enhanced the diversity of the dysarthric domain corpus and did not corrupt the VC model or degrade the intelligibility of the converted speech. Furthermore, the speech generated

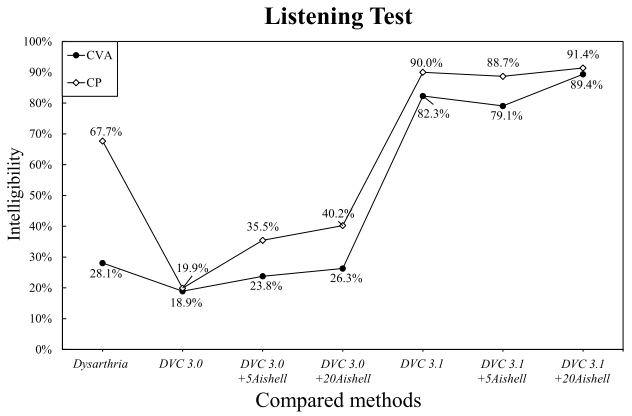

Fig. 7. Average speech recognition rates of the proposed DVC 3.1 in human-to-human application scenario with 13 listeners under free-talk test conditions. The X-axis represents the signal processing approach and the Y-axis represents the speech intelligibility.

via data augmentation was sufficiently similar to the original dysarthric speech to train a more generalized DVC system. More specifically, the data augmentation approach enabled the DVC system to process unknown dysarthric sentences effectively and simulate free speech.

Fig. 7 depicts the speech recognition results of the human-to-human application scenario in this study. For dysarthric patients suffering from CVA, the average speech intelligibility in free speech testing conditions was 28.1% for the original dysarthric speech, 18.9% for the speech processed using DVC 3.0, and 82.3% for the speech processed using DVC 3.1 in the free-talk testing conditions. The results of one dysarthric patient suffering from CP is depicted in Fig. 7. In this case, the average intelligibility was 67.7% for the original dysarthric speech, 19.9% for that processed by DVC 3.0, and 89.3% for that processed by DVC 3.1.

These results indicated that the proposed DVC 3.1 system outperformed the DVC 3.0 system comprehensively; hence, the proposed data augmentation approach aids the conversion of unknown sentences by the DVC system and improves the generalization ability of the conversion model. We also conducted listening tests on models of the Aishell corpora that were 5 and 20 times the size of the original one. The results exhibited similar trends to those of Google ASR, and the intelligibility remained above 79% in DVC 3.1, proving that the data augmentation method remained stable even when a large corpus was added. Owing to the limited speech data from dysarthria patients, the acoustic model of the DVC 3.0 system fails to accurately extract the correct and stable PPG features accurately when faced with unseen patient speech. This also leads to the gated CNN conversion model failing to learn the relationship between phoneme combinations of unseen utterances and their corresponding Mel features. Consequently, DVC 3.0 faces challenges in converting PPG features (dysarthric speech) into stable Mel features (normal speech). However, DVC 3.1 addresses this issue by using the Data Augmentation Unit to synthesize the dysarthria-like corpus. This compensates for the missing phoneme types and combinations in the dysarthria corpus. As a result, the conversion model learns the mapping function between the speech features of patients with dysarthria and those of normal speakers more effectively. It achieves greater conversion benefits, especially in free-talk scenarios.

Based on the above results, we further conducted statistical analysis, table V presents the statistical analysis results of each method in the human-to-human and human-to-machine testing conditions in which multiple linear regression was used. In the experiments, the seven methods and the speech provided by different patients (dysarthria participants 9 and 10) were independent variables, in which patients were regarded as confounders. For the speech provided by the different patients (dysarthria participants 9 and 10) under human-to-human testing conditions, dysarthria had a significantly lower intelligibility of 42.5%, with lower values of [71.0%, 60.8%, 57.1%] for [DVC 3.0, DVC 3.0+ 5Aishell, DVC 3.0+20Aishell] and [4.2%, 6.5%] for [DVC 3.1, DVC 3.1+5Aishell] compared to 96.5% for DVC 3.1+20Aishell, on average. For the speech provided by the different patients (dysarthria participants 9 and 10) under human-to-machine testing conditions, dysarthria had a significantly lower recognition rate of 50.4%, with lower values of [54.1%, 47.2%, 46.3%] for [DVC 3.0, DVC 3.0+5Aishell, DVC 3.0+20Aishell] and [−2.5%, 1.4%] for [DVC 3.1, DVC 3.1+5Aishell] compared to 83.1% for DVC 3.1+20Aishell, on average. More specifically, compared with dysarthria and the DVC 3.0 series, the DVC 3.1+20Aishell method had the highest positive coefficients while the confounder was fixed, which was a significant difference compared to the other methods in the intelligibility listening test and Google ASR evaluation. No significant differences were observed among DVC 3.1, DVC 3.1+5Aishell, and DVC 3.1+20Aishell, which means the large synthesized corpus with no overlap did not affect the stability of the DVC model.

To explore how generated data helps DVC models further, we evaluated the similarity between the generated dysarthria-like patients' speech and real speech from dysarthria patients using a listening test. Fig. 8 depicts the similarity results of the data augmentation method proposed in this study, obtained from 13 participants. On average, 87.59% and 85.65% of the listeners considered the synthesized dysarthria-like speech to be more similar to real dysarthric patients' speech compared to a normal speaker's speech. These results indicate that the Data Augmentation Unit can generate synthetic speech with certain patient-like characteristics. Through the use of synthetic speech data augmentation, the model can improve the speech conversion performance with a smaller amount of real speech data and a larger amount of generated synthetic speech resembling patients' speech. That is, this data augmentation approach of the proposed system can help to alleviate the burden of recording for patients with articulation disorders.

Fig. 9 presents an example of the converted speech obtained via data augmentation using the proposed DVC 3.1 system compared to that obtained using the DVC 3.0 system (without data augmentation) and real dysarthric speech by visualizing the feature $\widehat{N}_m^{Mels}$ converted by the trained gated CNN. As indicated by the red circles and arrows, the Mel spectrograms converted by DVC 3.1 were more similar to the distribution of the target speakers than those obtained using

TABLE V
MULTIPLE LINEAR REGRESSION STATISTICAL ANALYSIS OF EXPERIMENT RESULTS

Variables	Coefficient	Std Err	t	p	95% CI lower	95% CI upper
Human-to-human						
Inference DVC 3.1+20Aishell	0.965					
Method (Dysarthria / DVC 3.1+20Aishell)	-0.425	0.023	-18.542	0.000	-0.470	-0.380
Method (DVC 3.0 / DVC 3.1+20Aishell)	-0.710	0.023	-30.959	0.000	-0.755	-0.665
Method (DVC 3.0+5Aishell / DVC 3.1+20Aishell)	-0.608	0.023	-26.512	0.000	-0.653	-0.563
Method (DVC3.0+20Aishell / DVC3.1+20Aishell)	-0.571	0.023	-24.918	0.000	-0.616	-0.526
Method (DVC 3.1 / DVC 3.1+20Aishell)	-0.042	0.023	-1.846	0.065	-0.087	0.003
Method (DVC 3.1+5Aishell / DVC 3.1+20Aishell)	-0.065	0.023	-2.836	0.006	-0.110	-0.020
Participant (CVA / CP)	-0.122	0.012	-9.974	0.000	-0.146	-0.098
Human-to-machine						
Inference DVC 3.1+20Aishell	0.831					
Method (Dysarthria / DVC 3.1+20Aishell)	-0.504	0.039	-12.949	0.000	-0.580	-0.427
Method (DVC 3.0 / DVC 3.1+20Aishell)	-0.541	0.039	-13.922	0.000	-0.618	-0.465
Method (DVC 3.0+5Aishell / DVC 3.1+20Aishell)	-0.472	0.039	-12.146	0.000	-0.549	-0.396
Method (DVC3.0+20Aishell / DVC3.1+20Aishell)	-0.463	0.039	-11.909	0.000	-0.540	-0.387
Method (DVC 3.1 / DVC 3.1+20Aishell)	0.025	0.039	0.643	0.521	-0.051	0.101
Method (DVC 3.1+5Aishell / DVC 3.1+20Aishell)	-0.014	0.039	-0.353	0.725	-0.090	0.063
Participant (CVA / CP)	-0.074	0.021	-3.574	0.000	-0.115	-0.033

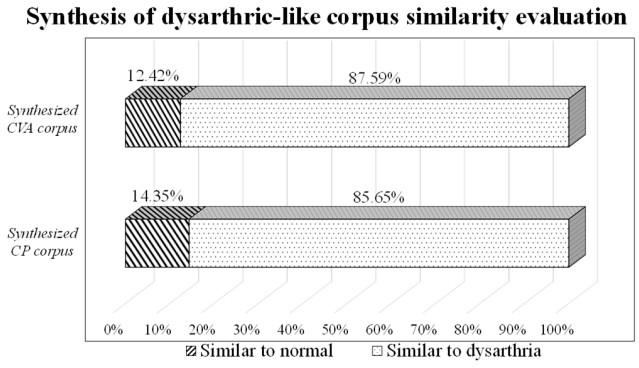

Fig. 8. Similarity listening test of synthetized dysarthric corpus. The X-axis represents the similarity percentage of the total number of utterances, while the Y-axis represents different sets of synthetic dysarthria-like speech.

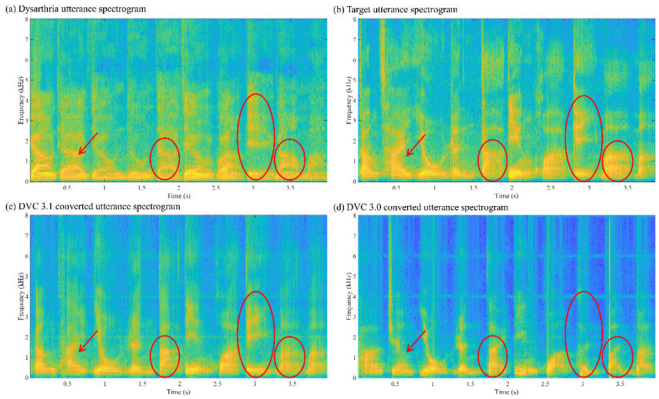

Fig. 9. Visualized spectrogram from three compared approaches under free-talk test conditions. All spectrogram content has the same textual meaning. (a) and (b) are the dysarthric and target speech utterances, whereas (c) and (d) are the utterances converted from DVC 3.1 and DVC 3.0, respectively.

DVC 3.0. This implies that the synthesized dysarthria-like corpus can effectively aid in the generation of more accurate Mel spectrograms using the gated CNN model of DVC 3.1, thereby improving the intelligibility of dysarthric utterances. Conversely, the DVC 3.0 system was observed to yield inaccurate Mel spectrograms frequently, thereby degrading the accuracy during the listening test to lower levels than those obtained based on unprocessed patient speech.

We further conducted Mel cepstral distortion (MCD) comparisons on 32 test sentences using the original dysarthria speech and the two sets of speech processed by the systems. The MCD results for these three sets of speech are presented in Table VI.

It can be observed that the dysarthria speech exhibited an average MCD of 1.1685 compared to the target's normal

TABLE VI
RESULTS OF MEL CEPSTRAL DISTORTION EVALUATION

	Dysarthria	DVC 3.0	DVC 3.1
Model for CVA	1.1293 ± 0.40	2.1994 ± 0.78	$\mathbf{0.8611 \pm 0.39}$
Model for CP	1.2076 ± 0.37	1.0875 ± 0.42	$\mathbf{0.9008 \pm 0.38}$
Average	1.1685 ± 0.38	1.6435 ± 0.60	$\mathbf{0.8809 \pm 0.39}$

speech. The speech converted by the DVC 3.0 system showed an average MCD of 1.6435 compared to the target's normal speech. However, the speech converted by the DVC 3.1 system had an average MCD of only 0.8809 compared to the target's normal speech. These results suggest that in a free-talk scenario, the DVC 3.1 system can produce more precise Mel frequency spectra and provide speech with higher intelligibility compared to the original dysarthria speech and that converted by DVC 3.0.

In summary, we have proposed DVC 3.1 based on data augmentation for dysarthric patients. This framework generates synthetic dysarthria-like speech through the Data Augmentation Unit, thereby replacing the missing phoneme combinations in the patient data and reducing the burden of recording for patients. The experimental results indicate that DVC 3.1 improves the speech intelligibility in free-talk scenarios. More specifically, the objective evaluations using ASR and subjective listening tests revealed that the proposed DVC 3.1 system exhibited an average intelligibility score of 86.0% and an average accuracy of 81.8% on free speech. Further, the results indicated that the data augmentation system implemented in DVC 3.1 yielded a dysarthria-like corpus that was sufficiently similar to real dysarthria data, thereby demonstrating that the augmentation system can synthesize patient-representative corpora to enhance the generalizability of the model in the absence of relevant sentences recorded from actual dysarthria patients. Furthermore, compared to the baseline system, which uses a dysarthria corpus recorded by ten patients for a total of 20.08 hours, the DVC 3.1 system, which uses data augmentation techniques, required only 1.28 hours of data recorded by the target patient. This significantly reduces the total recording time required from patients (a 93% reduction in data usage). Therefore, the DVC 3.1 system offers the potential to enhance speech intelligibility for dysarthric patients in real-world environments in the future.

IV. CONCLUSION

In this study, we investigated the effectiveness of data augmentation in the DVC 3.1 system for dysarthric patients and attempted to reduce the recording burden on patients and normal speakers. In the proposed DVC 3.1 system, the Data Augmentation Unit is used to generate large amounts of paired normal and patient-like corpora to improve the performance. The experimental results showed that the DVC 3.1 system achieved higher speech intelligibility performance for listeners under free-talk testing conditions. Based on the results of this study, we suggest that the DVC 3.1 system can reduce the burden of dysarthric patients to record training data. Moreover, DVC 3.1 can provide suitable speech intelligibility performance for listeners. Thus, the proposed DVC 3.1 is a potentially useful method for improving the speech intelligibility of dysarthria patients in free-talk application scenarios in the future.

This study has some limitations. We mainly focused on developing free-talk dysarthric VC with a data augmentation method and further analyzed the benefits of the system. To establish the baseline DVC 3.0, a significant amount of paired training sets (dysarthric and target speech) were required to help the model to encompass all possible pronunciations of patients and establish the mapping function. However, dysarthria patients tend to slur when speaking and have difficulty in maintaining a consistent volume and speed, which leads to challenges in obtaining paired corpora. Although this study showed that the DVC 3.1 system can provide more accurate communication to people and machines than DVC 3.0 on moderate dysarthric speech in a free-talk scenario, the data amount is still insufficient to claim that DVC 3.1 can maintain better performance than DVC 3.0 given sufficient paired corpora as training data. Unfortunately, during the validity period of our Institutional Review Board approval, the COVID-19 pandemic was rife, which made it difficult to gather and reach the dysarthria patients to record more data. As our study targeted patients with moderate dysarthria, only two patients who were diagnosed with moderate dysarthria were willing to record additional corpora after completing the willingness survey. Hence, we could only provide the results of one CVA and one CP patient with moderate dysarthria in this study. As the next phase of our work, we will invite more mild-to-severe dysarthria patients to record additional paired corpora and retrain both DVC systems to evaluate reliability in a future study.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available at http://sbplab.diskstation.me/TNSRE_demo.html

ACKNOWLEDGMENT

The authors would like to thank APrevent Medical Inc., for providing the training and testing data.

REFERENCES

[1] American Speech-Language-Hearing Association. (Aug. 2, 2021). *Dysarthria in Adults*. [Online]. Available: https://www.asha.org/practice-portal/clinical-topics/dysarthria-in-adults/

[2] P. Enderby, "Disorders of communication: Dysarthria," in *Handbook of Clinical Neurology*. Elsevier, 2013, pp. 273–281, doi: 10.1016/B978-0-444-52901-5.00022-8.

[3] A. Calvo et al., "Eye tracking impact on quality-of-life of ALS patients," in *Proc. Int. Conf. Comput. Handicapped Persons (ICCHP)*, 2008, pp. 70–77.

[4] S. Millar and J. Scott, "What is augmentative and alternative communication? An introduction," *Augmentative Communication in Practice—An Introduction*. Edinburgh, U.K.: Univ. of Edinburgh, 1998, p. 3.

[5] P. G. Blanchet and G. J. Snyder, "Speech rate treatments for individuals with dysarthria: A tutorial," *Perceptual Motor Skills*, vol. 110, no. 3, pp. 965–982, Jun. 2010, doi: 10.2466/PMS.110.3.965-982.

[6] H. C. Shane, S. Blackstone, G. Vanderheiden, M. Williams, and F. DeRuyter, "Using AAC technology to access the world," *Assistive Technol.*, vol. 24, no. 1, pp. 3–13, Mar. 2012, doi: 10.1080/10400435.2011.648716.

[7] A. Waller, "Telling tales: Unlocking the potential of AAC technologies," *Int. J. Lang. Commun. Disorders*, vol. 54, no. 2, pp. 159–169, Mar. 2019, doi: 10.1111/1460-6984.12449.

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

[9] J. P. Hosom, A. B. Kain, T. Mishra, J. P. Van Santen, M. F. Oken, and J. Staehely, "Intelligibility of modifications to dysarthric speech," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2003, p. I, doi: 10.1109/ICASSP.2003.1198933.

[10] H. Tolba and A. S. El-Torgoman, "Towards the improvement of automatic recognition of dysarthric speech," in *Proc. 2nd IEEE Int. Conf. Comput. Sci. Inf. Technol.*, Aug. 2009, pp. 277–281.

[11] 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 Trans. Biomed. Eng.*, vol. 64, no. 11, pp. 2584–2594, Nov. 2017, doi: 10.1109/TBME.2016.2644258.

[12] K.-C. Chen, H.-W. Yeh, J.-Y. Hang, S.-H. Jhang, W.-Z. Zheng, and Y.-H. Lai, "A joint-feature learning-based voice conversion system for dysarthric user based on deep learning technology," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 1838–1841, doi: 10.1109/EMBC.2019.8856560.

[13] J. Du and Q. Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions," in *Proc. Interspeech*, Sep. 2008, pp. 569–572, doi: 10.21437/Interspeech.2008-168.

[14] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," 2010, *arXiv:1003.4083*.

[15] S. Yang and M. Chung, "Improving dysarthric speech intelligibility using cycle-consistent adversarial training," in *Proc. 13th Int. Joint Conf. Biomed. Eng. Syst. Technol.*, 2020, pp. 308–313, doi: 10.5220/0009163003080313.

[16] W.-Z. Zheng, J.-Y. Han, C.-K. Lee, Y.-Y. Lin, S.-H. Chang, and Y.-H. Lai, "Phonetic posteriorgram-based voice conversion system to improve speech intelligibility of dysarthric patients," *Comput. Methods Programs Biomed.*, vol. 215, Mar. 2022, Art. no. 106602, doi: 10.1016/j.cmpb.2021.106602.

[17] K. Kintzley, A. Jansen, and H. Hermansky, "Event selection from phone posteriorgrams using matched filters," in *Proc. 12th Annu. Conf. Int. Speech Commun. Assoc.*, Aug. 2011, pp. 1–4, doi: 10.21437/Interspeech.2011-354.

[18] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Nov. 2009, pp. 421–426, doi: 10.1109/ASRU.2009.5372889.

[19] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition," in *Proc. Interspeech*, Sep. 2018, pp. 471–475, doi: 10.21437/Interspeech.2018-1751.

[20] J. Shor et al., "Personalizing ASR for dysarthric and accented speech with limited data," 2019, *arXiv:1907.13511*.

[21] Y. Jiao, M. Tu, V. Berisha, and J. Liss, "Simulating dysarthric speech for training data augmentation in clinical speech applications," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 6009–6013, doi: 10.1109/ICASSP.2018.8462290.

[22] Z. Jin et al., "Adversarial data augmentation using VAE-GAN for disordered speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[23] M. Soleymanpour, M. T. Johnson, R. Soleymanpour, and J. Berry, "Synthesizing dysarthric speech using multi-speaker tts for dysarthric speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7382–7386.

[24] Y. Wang et al., "Tacotron: Towards end-to-end speech synthesis," 2017, *arXiv:1703.10135*.

[25] S. R. Shahamiri and S. S. B. Salim, "Artificial neural networks as speech recognisers for dysarthric speech: Identifying the best-performing set of MFCC parameters and studying a speaker-independent approach," *Adv. Eng. Informat.*, vol. 28, no. 1, pp. 102–110, Jan. 2014, doi: 10.1016/j.aei.2014.01.001.

[26] J. Shen et al., "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4779–4783, doi: 10.1109/ICASSP.2018.8461368.

[27] K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

[28] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3617–3621, doi: 10.1109/ICASSP.2019.8683143.

[29] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "StarGAN-VC2: Rethinking conditional methods for StarGAN-based voice conversion," in *Proc. Interspeech*, Sep. 2019, pp. 679–683, doi: 10.21437/Interspeech.2019-2236.

[30] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 2100–2104.

[31] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6, doi: 10.1109/ICME.2016.7552917.

[32] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. Interspeech*, Sep. 2015, pp. 3214–3218, doi: 10.21437/Interspeech.2015-647.

[33] M. A. A. Aung and W. P. Pa, "Time delay neural network for Myanmar automatic speech recognition," in *Proc. IEEE Conf. Comput. Appl. (ICCA)*, Feb. 2020, pp. 1–4, doi: 10.1109/ICCA49400.2020.9022808.

[34] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4869–4873, doi: 10.1109/ICASSP.2015.7178896.

[35] L. L. N. Wong, S. D. Soli, S. Liu, N. Han, and M.-W. Huang, "Development of the Mandarin hearing in noise test (MHINT)," *Ear Hearing*, vol. 28, no. 2, pp. 70S–74S, 2007, doi: 10.1097/AUD.0b013e31803154d0.

[36] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 9, pp. 1570–1584, Sep. 2018, doi: 10.1109/TASLP.2018.2821903.

[37] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline," in *Proc. 20th Conf. Oriental Chapter Int. Coordinating Committee Speech Databases Speech I/O Syst. Assessment (O-COCOSDA)*, Nov. 2017, pp. 1–5, doi: 10.1109/ICSDA.2017.8384449.

[38] *Specifications for Audiometers*, Amer. Nat. Standards Inst., Acoust. Soc. Amer., New York, NY, USA, 2018.