Using StarGANv2 Voice Conversion to Enhance the Quality of Dysarthric Speech

Hadil Mehrez
Physical Engineering
and Instrumentation,
National Institute of Applied
Science and Technology
Tunis, Tunisia
hadil.mehrez@umoncton.ca

Mounira Chaiani
Research Laboratory in
Human-System Interaction,
University of Moncton
Shippagan Campus
New Brunswick, Canada
mounira.chaiani@umoncton.ca

Sid Ahmed Selouani
Research Laboratory in
Human-System Interaction,
University of Moncton
Shippagan Campus
New Brunswick, Canada
sid-ahmed.selouani@umoncton.ca

***Abstract*—In this paper, we propose to use StarGAN, a powerful Generative Adversarial Network (GAN), to improve the quality of dysarthric speech. Through extensive experiments, we demonstrate the effectiveness of StarGANv2-VC in converting dysarthric speech and significantly improving its intelligibility and naturalness. In addition, this research contributes to the field by conducting a comparative study between StarGANv2-VC and MaskCycleGAN-VC, another well-established GAN architecture, recently used in dysarthric speech conversion tasks. The results show that StarGANv2-VC performs the best, making it a promising solution for improving the speech quality of people suffering from dysarthria.**

***Index Terms*—Dysarthric speech, voice conversion, generative adversarial networks**

I. Introduction

DYSARTHRIA encompasses several speech issues caused by brain or nerve damage, affecting the control of muscles necessary for speech production. This muscular imbalance originates from factors such as paralysis or weakening of the muscles. Unlike other speech disorders, dysarthria is not related to difficulties in comprehension or movement planning. It can affect aspects such as volume, tone, and speech speed, making speech unclear, hard to understand, or irregular [1]. Different types of dysarthria exist, each with its distinct characteristics impacting speech differently, as mentioned in [2]. Flaccid dysarthria exhibits hyperreflexia and muscle flaccidity, resulting in hypernasality and imprecise consonants. Spastic dysarthria involves imprecise consonants, monotonous pitch, and reduced stress in speech. Ataxic dysarthria affects timing and movement direction, causing imprecise consonants and distorted vowel sounds. Hypokinetic dysarthria, associated with parkinsonism, includes symptoms like tremors and limited movements, resulting in monotonous speech and imprecise consonants. Hyperkinetic dysarthria, seen in dystonia and chorea, presents with imprecise consonants, variable speed rate, and distorted vowels, influenced by the nature of hyperkinesia in each condition. Understanding these traits is crucial for interpreting the results, especially in dysarthric voice conversion studies.

Having explored dysarthria as a speech disorder, our focus shifts to strategies for improving voice recognition among dysarthric individuals afflicted by dysarthria. This challenge is commonly tackled through two main methods. The first method performs a data augmentation [3], involving the enhancement of recognition capabilities by training Automatic Speech Recognition (ASR) models with synthetic data replicating dysarthric speech; the second method focuses on the enhancement of dysarthric speech itself. The latter approach, known as Voice Conversion (VC) [4], focuses on transforming a voice to sound like another person's voice without altering the linguistic content. In the context of converting dysarthric speech to normal speech, it aims to modify the dysarthric speech so that it becomes more intelligible, while preserving the words originally spoken.

A VC task is categorized as either parallel or nonparallel. Most voice conversion systems use parallel data, where the model is trained to convert audio between two speakers using the same set of prompts for both speakers. However, when dealing with speakers with dysarthria, finding high-quality healthy speech data from them is often challenging. This difficulty arises because people with dysarthria usually do not record themselves frequently, leading to a lack of clear speech samples [5]. Consequently, the voice conversion model needs to be trained on two different speakers: the dysarthric speaker and a different, healthy speaker. This situation induces and increases the need for a non-parallel voice conversion system.

This is where GANs come in opening up exciting possibilities for creating realistic artificial voices. Essentially, a GAN comprises two concurrently trained models: a generator and a discriminator. The generator is tasked with generating novel content, while the discriminator aims to differentiate between original data and that provided by the generator. Recent research has shown the effectiveness of MaskCycleGAN-VC when compared to other models, particularly when incorporating techniques like time stretching to enhance dysarthric speech quality [4]. Additionally, another approach proposed in [6] introduces a data augmentation-based VC system, DVC 3.1, designed to alleviate the recording burden on speakers.

This system utilizes text-to-speech and the StarGAN-VC architecture to synthesize a large target and patient-like corpus, aiming to reduce the challenges associated with recordings.

This paper presents an approach based on StarGANv2-VC [7] to refine dysarthric speech and therefore to enhance dysarthric speech recognition. The proposed approach is compared with the one using MaskCycleGAN-VC [8], a well-known voice conversion method, to identify the most effective strategy for enhancing the intelligibility of dysarthric speech. Through rigorous evaluation and comparison, our research aims to contribute to the development of advanced speech enhancement techniques, specifically tailored to address the communication challenge posed by dysarthric speech. Our findings have the potential to significantly impact assistive technology by helping individuals with speech disorders in their everyday communication.

II. Methodology

A. Datasets and Experimental Setup

In this study, we used the UASpeech [9] and Nemours [10] datasets.

The UASpeech dataset comprises 15 dysarthric speakers aged between 18 and 58, including 4 females and 11 males, along with 13 healthy speakers of the same age group. The dysarthric speakers present a diverse range of characteristics related to their specific speech disorders, revealing variations in speech intelligibility with severity levels spanning from very low to high percentages. Diagnosed types of dysarthria include spastic, athetoid, and mixed, highlighting diverse motor control issues affecting speech production. Notably, athetoid dysarthria is synonymous with dyskinetic cerebral palsy, characterized by gradual and unregulated movements, placing it within the hyperkinetic dysarthria category [5]. The broad age range of the speakers adds further complexity to their dysarthric profiles. Ongoing assessments of speech intelligibility underscore the dynamic nature of their communication abilities. Recordings were made using a 7-channel microphone with a sampling frequency of 16KHz, resulting in 7 recordings per prompt, along with a digital video camera. Each speaker in the database recorded a set of 765 isolated words, which serve as the speech materials for our experiments and analyses. Thus, there are a total of 5355 recordings per speaker. Speakers read three blocks of words: B1, B2, and B3, each containing 255 words. In these blocks:

- 155 words are repeated, including 10 digits (D0, D1, ..., D9), 26 radio-alphabet letters (LA, LB, LC, ..., LZ), 19 computer commands (C1, C2, ..., C19), and 100 common words (CW1, CW2, ..., CW100).
- 100 unique words that differ between the blocks (UW1, UW2, ..., UW100).

Blocks B1 and B3 are used for model training, while block B2 is reserved for testing purposes.

The Nemours database comprises 11 male speakers, each presenting unique speech characteristics assessed in terms of tongue, laryngeal, and respiratory functions, as well as conversational intelligibility. These evaluations collectively contribute to a thorough comprehension of the diverse dysarthria profiles observed within the speaker group. They accentuate variations in proficiency across various aspects of speech production, enhancing our understanding of the nuanced features inherent in each speaker's dysarthric expression. In terms of the recording setup, each speaker recorded 74 nonsensical short phrases and two continuous speech paragraphs, with a sampling frequency of 16 KHz. Each phrase in the database follows a specific structure: "The X is Ying the Z". X and Z (X $\neq$ Z) were selected randomly and without replacement from a set of 74 monosyllabic nouns, while Y was chosen without replacement from a set of 37 disyllabic verbs. This process resulted in 37 phrases, from which another 37 phrases were generated by swapping the tokens X and Z. Thus, across the entire set of 74 phrases, each noun and verb were pronounced twice by each speaker. Furthermore, all utterances are uttered by a healthy speaker JP. For the training, 70% of vocals were used leaving the remaining 30% for evaluation purposes.

For the VC process, we chose three speakers from the Nemours and UASpeech databases. BB, with moderate intelligibility, and KS, exhibiting very low intelligibility, were trained with the healthy speaker JP in the Nemours dataset. Additionally, we selected M05, who possesses an average level of intelligibility, trained along with the healthy speaker CM01 from the UASpeech database. These selections were carefully made to provide a diverse representation of the dataset's common characteristics.

B. Experimental Design

In this section, we outline the detailed steps we took to achieve voice conversion using the StarGANv2 model [7]. Our goal was to adapt the pre-trained StarGANv2 VC[1] model, originally designed for normal voice processing, to be capable of converting dysarthric voices. Concurrently, we also conducted conversions using the pre-trained MaskCycleGAN-VC model to assess the performance of our StarGANv2-VC through comparison.

Throughout our experiments, we worked with a sampling frequency of 24 KHz for the StarGANv2-VC model which operates alongside the Parallel WaveGAN [11] vocoder. For MaskCycleGAN-VC, the MelGAN [12] vocoder, trained at 22050 Hz, was used. The training process lasted for 150 epochs for each model, enabling them to converge toward better performance.

C. GAN Architectures

*1) **MaskCycleGAN-VC**:* MaskCycleGAN-VC is a non-parallel voice conversion technique that employs a MelGAN vocoder and incorporates a two-step adversarial loss mechanism to counteract the excessive smoothing effects induced by the cycle-consistency loss. It transforms acoustic features from a source domain x $\in$ X to a target domain y $\in$ Y through the

[1]https://github.com/yl4579/StarGANv2-VC

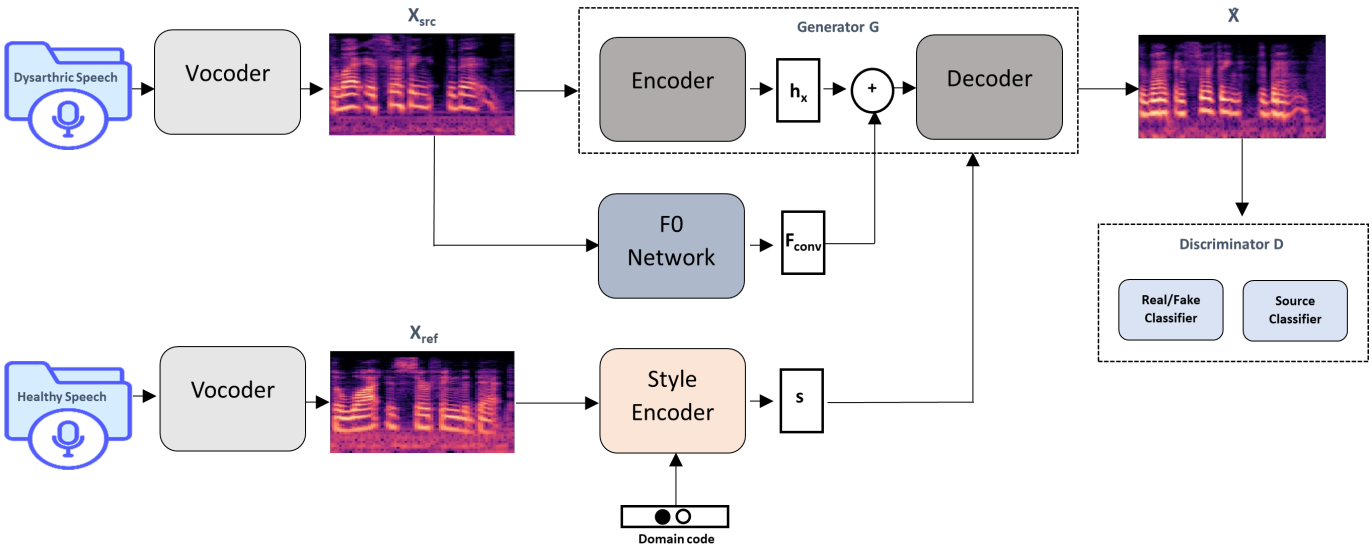

Fig. 1: StarGANv2 architecture for dysarthric speech voice conversion

use of a neural network F as a forward-generator (X → Y), and G as the backward-generator (Y → X). Additionally, this model incorporates two extra discriminators, D'$_X$ and D'$_Y$, specifically for a secondary adversarial loss concerning bi-directionally converted features.

$$\mathcal{L}_{\text{GAN2}}(G,F,D',X) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log(0 - D'(x))] + \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log(1 - D'(F(G(x))))] \quad (1)$$

However, in 2021, MaskCycleGAN-VC underwent significant improvements, extending its capabilities to include mel-spectrogram conversion. It effectively introduces a frame in-painting task. This self-supervised learning method enabled the model to master temporal structures without the need for excessive parameters. Furthermore, MaskCycleGAN-VC introduced a novel data augmentation technique called fill-in-the-frame data augmentation (FIF DA). Through FIF, a temporal mask was applied to the input mel-spectrogram, stimulating the converter to fill in missing frames based on surrounding frames. These advancements resulted in superior performance, confirming MaskCycleGAN-VC's effectiveness over its previous versions [8].

We implemented MaskCycleGAN-VC[2] using the provided implementation with identical parameters [5]. Training segments, each consisting of only 64 frames, are randomly selected from the training samples. The MaskCycleGAN-VC model is initially trained for 25 epochs, although we opted to extend the training duration to a maximum of 150 epochs.

*2) **StarGANv2-VC**:* StarGANv2-VC is a non-supervised, non-parallel many-to-many voice conversion method that uses a GAN.

The architecture of StarGANv2-VC, as shown in Figure 1, is composed of the following components:

[2]https://github.com/GANtastic3/MaskCycleGAN-VC

Generator: The generator G processes an input audio recording presented as a mel-spectrogram X_{src} along with a randomly chosen style vector s, signifying particular traits to be infused into the voice to generate a new mel-spectrogram G(X,s) by employing an adversarial loss function L_{adv}. This function encourages the generator G to generate realistic mel-spectrograms by converting a sample X from the source domain y_{src} to a sample $\hat{X}$ in the target domain y_{trg}, attempting to deceive a discriminator D that is trained to differentiate between real and generated mel-spectrograms. Here, D(.; y) represents the output of the real/fake classifier for the domain y in the set Y.

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{X;y_{\text{src}}}[\log D(X;y_{\text{src}})] + \mathbb{E}_{X;y_{\text{trg}};s}[\log(1 - D(G(X;s);y_{\text{trg}}))] \quad (2)$$

An additional adversarial loss function denoted as L_{advcls} with a source classifier C uses the cross-entropy loss function, denoted as CE(.). It encourages the model to generate samples that both deceive the discriminator in the target domain and exhibit unique features from the source domain, as identified by the source classifier C.

$$\mathcal{L}_{\text{advcls}} = \mathbb{E}_{X;y_{\text{trg}};s}[\text{CE}(C(G(X;s));y_{\text{trg}})] \quad (3)$$

F0 network: The F0 network is a pre-trained Joint Detection and Classification (JDC) network [13] designed to extract the fundamental frequency from a given mel-spectrogram input. However, The F0 consistency loss function L_{f0} ensures that the generated results are consistent with the normalized F0 curve obtained from the F0 network.

ASR model: The ASR model is a pre-trained system designed to transcribe speech into text. This model is essential for the training and evaluation processes, allowing the system to convert spoken language into written form. To ensure that the converted speech retains the identical linguistic content as

the source, we apply a speech consistency loss function L_{asr} using convolutional features extracted from a pre-trained joint CTC-attention VGG-BLSTM network [14].

Mapping network: The mapping network M generates a style vector. This process ensures diverse style representations across different domains by utilizing random latent codes sampled from a Gaussian distribution.

Style encoder: The style encoder S transforms an ordinary voice into various styles. To ensure that the style code h_{sty} can be accurately reconstructed from the generated samples, a style reconstruction loss function L_{sty} is implemented. However, to force the generator to generate different samples with different style codes, a style diversification loss function $L_{sd}$is needed.

$$L_{\text{sty}} = \mathbb{E}_{X,y_{\text{trg}},s}\left[\|s - S(G(X,s),y_{\text{trg}})\|_1\right] \tag{4}$$

Discriminator: The discriminator D has shared layers that recognize common features between real and fake samples across all domains. It uses a domain-specific binary classifier C to verify samples in each domain. However, the single-layer domain classifier might miss important domain-specific details. To address this issue, an additional classifier that recognizes original domain features in converted samples was added. This information helps the generator to capture unique domain traits, improving sample accuracy.

Additionally, to maintain the speech/silence intervals in the generated samples, StarGANv2 VC uses a norm consistency loss function L_{norm}. It employs also a cycle consistency loss function $_{cyc}$ to preserve all other features of the input.

Otherwise, the full objective functions for the generator can be summarized as follows where λ represents the hyper-parameter:

$$\begin{aligned}\min_{G,S,M} L_{\text{adv}} + \lambda_{\text{advcls}}L_{\text{advcls}} + \lambda_{\text{sty}}L_{\text{sty}} - \lambda_{\text{ds}}L_{\text{ds}} \\ +\lambda_{\text{f0}}L_{\text{f0}} + \lambda_{\text{asr}}L_{\text{asr}} + \lambda_{\text{norm}}L_{\text{norm}} + \lambda_{\text{cyc}}L_{\text{cyc}}\end{aligned} \tag{5}$$

The full objective for the discriminators is articulated as follows, where λ_{cls} denotes the hyperparameter for the source classifier loss L_{cls}, as detailed below:

$$\min_{C,D} - L_{\text{adv}} + \lambda_{\text{cls}}L_{\text{cls}} \tag{6}$$

$$L_{\text{cls}} = \mathbb{E}_{X;y_{\text{src}};s}[CE(C(G(X,s)),y_{\text{src}})] \tag{7}$$

Moreover, the StarGANv2-VC model is entirely convolutional. When coupled with a high-speed vocoder like Parallel WaveGAN, it can execute voice conversions in real-time, meaning it operates at a speed comparable to natural speech [7].

D. Evaluation

To assess how voice conversion impacts dysarthric speech quality, we chose to investigate its effects on Automatic Speech Recognition (ASR) systems. This involves converting speech into text and then analyzing the results using important metrics like Word Error Rate (WER) and Character Error Rate (CER).

*1) **Automatic Speech Recognition (ASR)**:* After the conversion of dysarthric speech, the converted audio files need to pass through an ASR system to evaluate the effects of this conversion. These outcomes will be compared with the reference model that showcases the ASR results of the original, non-converted files. To carry out this evaluation, two state-of-the-art ASR systems, namely Wav2Vec 2.0 [15] and OpenAI [16], have been chosen.

Wav2Vec 2.0 by Facebook AI employs self-supervised learning with non-transcribed audio data for speech recognition. It transforms audio into phonemes to enhance recognition with contextual information. On the other hand, OpenAI's Whisper uses weak supervised learning, and pre-training models to predict words directly from audio. Whisper refines its predictions by considering the overall context of the sentence. However, it shows that training on a broad supervised dataset and prioritizing zero-shot transfer can greatly enhance the strength of a speech recognition system [16].

In summary, both approaches use context to enhance speech recognition, but they do so in their unique ways. Wav2Vec 2.0 achieves this by transforming audio into phonemes for better context, while Whisper examines the entire sentence to ensure words fit appropriately within the global context.

*2) **Word Error Rate (WER) and Caracter Error Rate (CER)**:* To evaluate the effectiveness of voice conversion methods on ASR systems, it's important to use specific measures like WER and CER. These measures help us to understand how accurately the transformation is done by counting errors in words and characters.

The WER determines the difference between two texts in terms of words. The distance between two text strings is calculated by the Levenshtein distance [17] which represents the minimum number of simple operations (insertions, deletions, or substitutions of a single character) needed to transform one word into another.

For instance, the Levenshtein distance between "kitten" and "sitting" is 3. This means it takes 3 operations to change one into the other, and there's no way to do it with fewer than 3 modifications: kitten $\rightarrow$ sitten (substituting "s" for "k"), sitten $\rightarrow$ sittin (substituting "i" for "e") and sittin $\rightarrow$ sitting (adding "g" at the end).

To calculate the WER, various operations are involved: S denotes the count of substitutions, D represents the count of deletions, I stands for the count of insertions, and N signifies the total number of words in the reference text. The calculation of WER is done according to the following formula:

$$WER = \frac{S + D + I}{N} \tag{8}$$

In simple terms, this expression encapsulates the measure of accuracy in terms of modifications needed to align the words in the generated text with those in the reference text. A WER of 0% is achieved when all predictions perfectly match the reference text. However, the number of steps needed can be larger than the number of words in the reference, resulting in a WER greater than 100%.

TABLE I: WER and CER in percentage for voice conversion using StarGANv2 and MaskCycleGAN on dysarthric speakers BB, KS, and M05

Speaker	StarGANv2-VC			MaskCycleGAN-VC		
		WER(%)	CER(%)		WER(%)	CER(%)
BB_JP	ParallelWave GAN	61,36	39,8	MelGAN	58,33	34,06
	25_epoch	75	49.07	25_epoch	84.85	61.21
	50_epoch	**57.58**	**37.27**	50_epoch	87.12	57.84
	75_epoch	65.15	39.97	75_epoch	73.48	45.87
	150_epoch	90.91	60.54	150_epoch	61.36	38.95
KS_JP	ParallelWave GAN	100	82,27	MelGAN	102,27	84,62
	25_epoch	100	80.17	25_epoch	152.27	132.11
	50_epoch	**95.45**	**74.79**	50_epoch	141.67	123.08
	75_epoch	104.55	80.5	75_epoch	137.88	130.1
	150_epoch	104.55	79.5	150_epoch	133.33	116.72
M05_CM01	ParallelWave GAN	202,18	166,85	MelGAN	208,74	123,45
	25_epochs	294.51	258.69	25_epochs	270.76	178.55
	50_epochs	283.87	262.36	50_epochs	243.64	171.75
	75_epochs	**190.03**	178.26	75_epochs	230.25	156.45
	150_epochs	262,24	249,42	150_epochs	226,11	150,6

The CER takes character-level errors into account and provides a more detailed evaluation. The calculation of CER is based on the same function as WER. However, instead of using a list of words as input for the edit distance function, a single character string is used. This approach allows for a better capture of character-level errors and provides a more nuanced evaluation of the model's performance.

III. Results and Discussion

In this study, we thoroughly analyzed how the StarGANv2-VC model transforms dysarthric speech, highlighting specific performance metrics for each speaker.

To determine the impact of voice conversion on dysarthric audio quality and the Wav2vec 2.0 system, we used a reference model. The reference model consists of the dysarthric original audio files processed by the appropriate vocoder, MelGAN for MaskCycleGAN-VC, and ParallelWave GAN for StarGANv2-VC.

The results of voice conversion for dysarthric speech using StarGANv2 and MaskCycleGAN are presented in Table I where values in bold represent the best results compared to the reference model.

For all speakers, StarGANv2-VC demonstrates a positive impact on voice quality. This observation is noted by the lower WER compared to the reference model. On the other side, MaskCycleGAN-VC degrades the quality of dysarthric speech, resulting in a much higher WER compared to that of the reference model.

The distinct dysarthria profiles of speakers BB and KS, shed light on the differences observed in the voice conversion results using StarGANv2-VC and MaskCycleGAN-VC. Speaker BB, with strong articulatory proficiency and sporadic articulation, exhibits relatively lower WER and CER across both voice conversion models. In contrast, speaker KS, characterized by hyperkinetic dysarthria and challenges in tongue and laryngeal functions, presents higher WER and CER percentages in both models. The disparities in dysarthric speech characteristics, such as the ability to maintain control over speech-related motor functions, likely contribute to the varied performance in voice conversion between the two speakers. These findings emphasize the impact of individualized dysarthria traits on the effectiveness of voice conversion techniques. However, M05 shows notably higher WER and CER compared to speakers BB and KS across both StarGANv2-VC and MaskCycleGAN-VC. The increased WER and CER percentages for M05 suggest that the voice conversion models face greater challenges in accurately converting the dysarthric speech of M05, potentially due to the specific characteristics associated with spastic dysarthria, such as variations in muscle tone and control affecting speech production. This discrepancy can be also attributed to the nature of the database. BB is part of the Nemours database, where Wav2Vec 2.0 is tasked with predicting 22 phrases. In contrast, M05 belongs to the UASpeech database, where predictions are made for 1785 words.

Furthermore, the CER is always as low as the WER, this is because the Wav2Vec 2.0 system predicts a lot of words almost correctly. For example, it predicted the word "backspace" as "bekspace", which is a large error according to the WER, but a reasonably small error according to the CER.

It is important to note that the performance of StarGANv2-VC varies depending on the individual characteristics of each speaker. Speakers from Nemours and UASpeech exhibit unique dysarthria traits, which have influenced the outcomes.

Moreover, the results are even more significant when comparing the performance of StarGANv2-VC with that of its competitor, MaskCycleGAN-VC. Indeed, the WER and CER rates conclusively demonstrate that StarGANv2-VC stands out as a more effective solution for dysarthric voice conversion. This observation suggests that StarGANv2-VC can significantly improve dysarthric voice conversion compared to other approaches such as MaskCycleGAN-VC.

Additionally, the importance of these results is emphasized by the spectrograms presented in Figures 2, 3, and 4. Figure 2 illustrates the original pathological voice of speaker BB, while Figures 3 and 4 show the voice converted using MaskCycleGAN-VC and StarGANv2-VC, respectively. The spectrogram reveals notable enhancements, such as a clearer distinction of the four formants and a significant reduction in

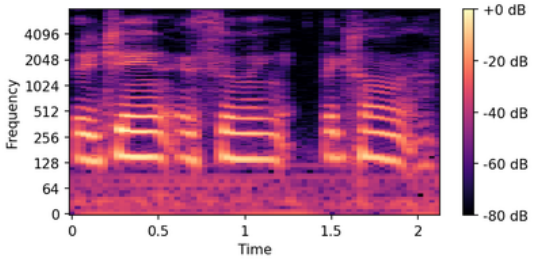

Fig. 2: Original pathological voice spectrogram of speaker BB

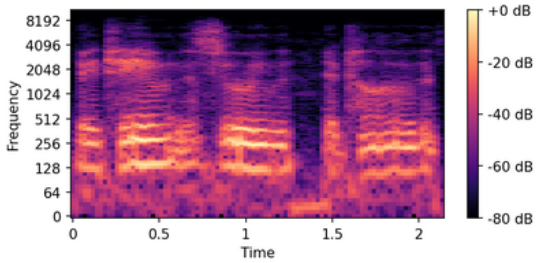

Fig. 3: MaskCycleGAN-VC spectrogram of speaker BB

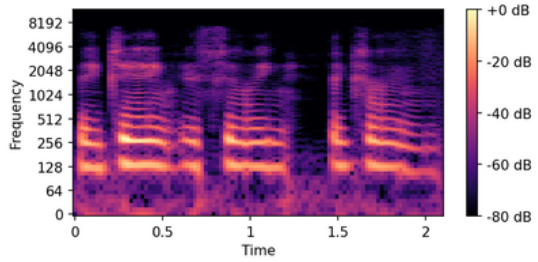

Fig. 4: StarGANv2-VC spectrogram of speaker BB

signal noise. A direct comparison with MaskCycleGAN-VC highlights that StarGANv2-VC achieves optimal clarity and quality improvement results.

To identify which ASR system performs best in conjunction with StarGANv2-VC for dysarthric speech conversion, a comparison was conducted between OpenAI's ASR and Wav2Vec 2.0 ASR, on dysarthric speakers BB, KS, and M05, as shown in Table II. This evaluation was based on calculating the WER and CER in percentage, considering the prediction outcomes of StarGANv2 VC with 50 epochs for both BB_JP and KS_JP, and 75 epochs for M05_CM01.

After examining the outcomes of each ASR system, it is clear that OpenAI demonstrates superior performance for the speaker BB. There is a noticeable difference in the lower WER and CER rates when compared to those generated by Wav2Vec 2.0. However, it appears that Wav2Vec 2.0 performs better overall, as evidenced by its performance with speakers KS and M05. Moreover, the OpenAI ASR system performs better in predicting sentences rather than individual words. This is

TABLE II: A comparative study between Wav2Vec 2.0 and OpenAI ASR with StarGANv2-VC on dysarthric speakers

	Wav2Vec 2.0		**OpenAi**	
Speaker	**WER(%)**	**CER(%)**	**WER(%)**	**CER(%)**
BB_JP	57,58	37,27	**42,42**	**30,69**
KS_JP	**95,45**	**74,79**	100	98,16
M05_CM01	**190,03**	**178,26**	335,35	258,03

demonstrated by the increasing WER and CER rates for the speaker M05 from UASpeech.

It is crucial to emphasize that integrating the StarGANv2-VC model with the Parallel WaveGAN vocoder and the Wav2Vec 2.0 ASR system resulted in significantly better outcomes compared to combining MaskCycleGAN with the MelGAN vocoder and Wav2Vec 2.0 ASR system. Notably, MaskCycleGAN demonstrates remarkable performance only when time stretching is added [4]. During our experiments, we also attempted to apply time stretching to StarGANv2's output files. Unfortunately, this approach led to a significant deterioration in vocal quality when converted by the Parallel WaveGAN vocoder. This underscores that this vocoder is not suitable for satisfactory time stretching.

These results cannot be generalized to the broader population of individuals with dysarthria. Even when individuals are categorized as having the same type of dysarthria, there exists considerable variability in their speech characteristics. Other speakers may exhibit different levels of improvement, ranging from worse to better outcomes.

IV. Conclusion

In this paper, we proposed a novel dysarthric voice conversion method using StarGANv2-VC, which was originally designed for normal voice conversion. The obtained results demonstrate the efficacy of the StarGANv2-VC model in enhancing the quality of dysarthric speech, particularly for speakers with moderate to severe dysarthria. When compared to the MaskCycleGAN model, StarGANv2 showed superior conversion performance.

As a limitation, our attempts to apply time stretching to the output files of StarGANv2 revealed a significant deterioration in vocal quality when processed by the Parallel WaveGAN vocoder. This demonstrates the inadequacy of the current vocoder for satisfactory time-stretching applications. In future work, addressing this limitation will be a focal point, involving the exploration of additional adjustments to the StarGANv2 model and investigating alternative vocoders that can better accommodate and enhance the effectiveness of time stretching in voice conversion. As a notable point of comparison, it is worth mentioning that MaskCycleGAN, in contrast to StarGANv2, demonstrated success in implementing time stretching. Specifically, when time stretching was applied to the output files of MaskCycleGAN and processed with the MelGAN vocoder, satisfactory results were achieved. This highlights a capability of MaskCycleGAN that contrasts with the limitations encountered in the context of StarGANv2 and Parallel WaveGAN vocoder. However, it is

essential to acknowledge a limitation within the architecture of MaskCycleGAN. Unlike StarGANv2, which incorporates ASR in its architecture, MaskCycleGAN lacks this feature. The absence of an ASR component in MaskCycleGAN may pose constraints when leveraging linguistic information during voice conversion.

REFERENCES

[1] Laureano Moro-Velazquez, JaeJin Cho, Shinji Watanabe, Mark A Hasegawa-Johnson, Odette Scharenborg, Heejin Kim, and Najim Dehak. Study of the performance of automatic speech recognition systems in speakers with parkinson's disease. In *Interspeech*, volume 9, pages 3875–3879, 2019.

[2] Frederic L Darley, Arnold E Aronson, and Joe R Brown. Differential diagnostic patterns of dysarthria. *Journal of speech and hearing research*, 12(2):246–269, 1969.

[3] Enno Hermann and Mathew Magimai.-Doss. Few-shot Dysarthric Speech Recognition with Text-to-Speech Data Augmentation. In *Proc. INTERSPEECH 2023*, pages 156–160, 2023.

[4] Luke Prananta, Bence Halpern, Siyuan Feng, and Odette Scharenborg. The Effectiveness of Time Stretching for Enhancing Dysarthric Speech for Improved Dysarthric Speech Recognition. In *Proc. Interspeech 2022*, pages 36–40, 2022.

[5] Marjolein Spijkerman. Using voice conversion and time-stretching to enhance the quality of dysarthric speech for automatic speech recognition, 2022.

[6] Wei-Zhong Zheng, Ji-Yan Han, Chen-Yu Chen, Yuh-Jer Chang, and Ying-Hui Lai. Improving the efficiency of dysarthria voice conversion system based on data augmentation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, pages 1–1, 2023.

[7] Yinghao Aaron Li, Ali Zare, and Nima Mesgarani. StarGANv2-VC: A Diverse, Unsupervised, Non-Parallel Framework for Natural-Sounding Voice Conversion. In *Proc. Interspeech 2021*, pages 1349–1353, 2021.

[8] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. Maskcyclegan-vc: Learning non-parallel voice conversion with filling in frames. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5919–5923, 2021.

[9] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas S. Huang, Kenneth Watkin, and Simone Frame. Dysarthric speech database for universal access research. In *Proc. Interspeech 2008*, pages 1741–1744, 2008.

[10] Xavier Menendez-Pidal, James B Polikoff, Shirley M Peters, Jennie E Leonzio, and H Timothy Bunnell. The nemours database of dysarthric speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1962–1965. IEEE, 1996.

[11] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6199–6203, 2020.

[12] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019.

[13] Sangeun Kum and Juhan Nam. Joint detection and classification of singing voice melody using convolutional recurrent neural networks. *Applied Sciences*, 9(7):1324, 2019.

[14] Suyoun Kim, Takaaki Hori, and Shinji Watanabe. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4835–4839. IEEE, 2017.

[15] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[16] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[17] Rishin Haldar and Debajyoti Mukhopadhyay. Levenshtein distance technique in dictionary lookup methods: An improved approach. *arXiv preprint arXiv:1101.1232*, 2011.