Comparing and Analyzing Different Speech Conversion Techniques for Transforming Dysarthric to Normal Speech

Comparing and Analyzing Different Speech Conversion Techniques for Transforming Dysarthric to Normal Speech

Master Thesis

to obtain the degree of Master of Science
in Embedded Systems
at Delft University of Technology,
to be defended publicly on May 29th 2024

by

Jingxian Liu

Faculty of Electrical Engineering, Mathematics Computer Science,
Delft University of Technology, Delft, Netherlands,
born in China.

Thesis committee:

Dr. Odette Scharenborg,	Technische Universiteit Delft
Dr. Qun Song,	Technische Universiteit Delft
Dr. Zhengjun Yue,	Technische Universiteit Delft

Science is a wonderful thing
if one does not have to earn one's living at it.

Albert Einstein

CONTENTS

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 Research Questions . . . 2
- 1.3 Overview . . . 3

2 Background **5**
- 2.1 Dysarthric Speech . . . 6
- 2.2 Dysarthric to Normal Speech Conversion . . . 7
- 2.3 Generative Adversarial Network . . . 7
  - 2.3.1 Generator . . . 7
  - 2.3.2 Discriminator . . . 8
  - 2.3.3 Objective Function . . . 8
  - 2.3.4 Training . . . 8
- 2.4 StarGAN . . . 8
  - 2.4.1 StarGAN . . . 8
  - 2.4.2 StarGAN v2 and StarGAN V2-VC . . . 10
- 2.5 MaskCycleGAN-VC . . . 12
- 2.6 Evaluation Metric . . . 13
  - 2.6.1 PER . . . 13
  - 2.6.2 CER . . . 14

3 Methodolody **15**
- 3.1 Dataset . . . 16
  - 3.1.1 UASpeech . . . 16
  - 3.1.2 TIMIT . . . 17
- 3.2 Signal-processing Techniques . . . 18
- 3.3 Time Stretching (TS) . . . 19
- 3.4 GAN-based VC Models . . . 20
  - 3.4.1 StarGANv2-VC . . . 20
- 3.5 Evaluation . . . 21
  - 3.5.1 Objective Evaluation (Phoneme Error Rate (PER)) . . . 21
  - 3.5.2 Subjective Evaluation (Naturalness and Intelligibility) . . . 21

4 Results **25**
- 4.1 Results of the Objective Evaluation . . . 26
  - 4.1.1 Evaluation System . . . 26
  - 4.1.2 Loudness Increasing . . . 26
  - 4.1.3 Time stretching . . . 27
  - 4.1.4 MaskCycleGAN-VC . . . 28

4.1.5 StarGANv2-VC 28
4.1.6 Overall 28
4.2 Results of the Subjective Evaluation 30
4.2.1 Naturalness Evaluation 30
4.2.2 Intelligibility Evalution 30
4.3 Correlation between objective and subjective results 31

5 Discussion **35**
5.1 Discussion 35
5.2 Limitations and Future Work 37

6 Conclusion **39**

1

INTRODUCTION

1.1. MOTIVATION

Dysarthria [7] is a speech sound disorder triggered by neurological damage to the motor speech system. Speech affected by dysarthria is influenced by restricted movement of the tongues, lips, and jaws. It could be caused by different reasons such as stroke, traumatic brain injury, cerebral palsy, Parkinson's disease, amyotrophic lateral sclerosis (ALS), multiple sclerosis, and so on. Dysarthric speech differs from normal speech in many aspects [26]. Compared to normal speech, dysarthric speech has poorer articulation, slower speech rate, reduced or increased loudness [27], and pitch variation[27] due to muscle weakness or coordination issues. It dramatically influences the ability of people with dysarthric speech to communicate with others and the influence grows with the increase of the severity of dysarthric speech.

Nowadays, automatic Speech Recognition (ASR) is widely used in daily life with high accuracy. More and more smart devices are used in our daily life making our lives more convenient. Integrated into various applications, ASR enhances convenience and efficiency in daily tasks. From telecommunications to smart home automation, ASR enables hands-free communication and voice controls. However, individuals with dysarthria cannot benefit from mainstream ASR due to the low performance of dysarthric speech recognition compared to normal speech [30]. In order to improve dysarthric speech recognition performance, some researchers include dysarthric speech in their training material [29], while others aim to improve the intelligibility of dysarthric speech through the conversion of dysarthric speech to normal speech to ensure that the speech is audible and clear enough to be easily understood by listeners. Our project will focus on the latter one.

Various voice conversion (VC)[19] and signal processing (SP) techniques have been proposed to transform dysarthric speech into normal speech, we refer to these together as "speech conversion techniques". Ideally, these speech conversion techniques adapt the various dysarthric speech characteristics to become more similar to those of normal

speech. One such approach is GAN (Generative Adversarial Network)-based voice conversion (VC) [28, 12, 22, 5, 21, 11]. For VC-based techniques, MaskCycleGAN-VC [12] has been shown to outperform cycleGAN-VC [28], showing approximately 4% absolute better performance for male speakers and 9% for female speakers in dysarthric speech recognition [21]. StarGAN-VC [4] also shows promise [5]. Interestingly, SP techniques such as time-stretching have been shown to have an equal or even better recognition performance than state-of-the-art VC models [21], even though time stretching only focuses on adapting only one dysarthric speech characteristic, i.e., speech rate, to that of normal speech, while VC aims to convert all dysarthric speech characteristics to normal speech characteristics at the same time.

There are usually two ways to evaluate VC results [24]. The objective approach uses an ASR system to assess the error rates of the converted speech. The subjective approach has human listeners judge the converted speech [11], e.g., on their naturalness or intelligibility.

Although research exists that compares different voice conversion techniques [16, 31, 19], it is not clear if these methods work equally well for different severities of dysarthric speech. In this work, we compare different dysarthric-to-normal speech conversion approaches for two levels of severity (low and high severity) by doing phone recognition on the converted speech with an ASR system trained on normal speech. By answering this question, we could potentially further improve dysarthric speech recognition for different severities. Moreover, although it is often assumed that the naturalness of the converted speech is important for ASR performance, this question is actually unanswered. We therefore investigate the naturalness and intelligibility of the resulting speech using human listening experiments and correlate the mean opinion scores (MOS) to the ASR's phone error rates (PER).

1.2. RESEARCH QUESTIONS

This research aims to compare the performance of speech conversion techniques across two levels of dysarthric speech severity, both objectively and subjectively, and to investigate the relationship between objective and subjective evaluations. Specifically, we compare two state-of-the-art VC-based techniques (Masked Cycle-GAN and Star-GAN) and two SP-based techniques (time stretching and loudness, as dysarthric differs in loudness from normal speech [27]) for speech conversion. Our main research questions are as follows:

- **RQ1:** Which speech conversion technique leads to the highest recognition performance for two severities of dysarthric speech?
- **RQ2:** Which speech conversion technique improves the naturalness and intelligibility of dysarthric speech for human listeners?
- **RQ3:** Does increased naturalness lead to better ASR performance for dysarthric speech?
- **RQ4:** Does increased intelligibility lead to better ASR performance for dysarthric speech?

1.3. OVERVIEW

This thesis includes several chapters. Chapter 2 will introduce the necessary background for this work. In Chapter 3, we will explain the methodology used in this thesis, including the dataset, experiment setup, different techniques we are going to compare, and the evaluation methods. Chapter 4 will list all the results we get from the experiments, including the objective results and subjective results. In Chapter 5 and Chapter 6, we will discuss the results and draw a conclusion, respectively.

2

BACKGROUND

In this section, we lay the foundation necessary for comprehending the thesis by introducing essential background information. Section 1 explains the nature of dysarthric speech, providing a detailed overview of what is dysarthric speech. Following that, Section 2 introduces dysarthric to normal conversion, outlining the VC systems employed in transforming dysarthric speech to normal speech. Section 3 is dedicated to explaining the basic knowledge of Generative Adversarial Networks (GANs). Subsequently, Sections 4 and 5 focus on the application of StarGAN—a specific variant of GANs. Section 4 describes the architecture of StarGAN. Section 5 explores the utilization of StarGAN for converting speech with one characteristic into another and the architecture of StarGAN-VC. At last, we present the evaluation metrics adopted in our experiment, which are crucial for assessing the effectiveness of the voice conversion processes.

2.1. DYSARTHRIC SPEECH

Dysarthria[7] is a speech sound disorder resulting from neurological injury of the motor speech system. Dysarthric speech is often characterized by its poorer articulation, slower speech rate, reduced loudness, changes in voice quality (such as breathiness, hoarseness, or nasal speech), and so on because of limited tongue, lip, and jaw movement. The causes of dysarthria are diverse and can include stroke, traumatic brain injury, cerebral palsy, Parkinson's disease, amyotrophic lateral sclerosis (ALS), multiple sclerosis, and Guillain-Barre syndrome, among others. The specific characteristics of dysarthric speech can vary widely among individuals, depending on the underlying cause and the severity of the neurological impairment.

Dysarthria could be categorized according to which part of the nervous system is implicated[6]. Each category has particular characteristics which are associated with underlying neurological conditions. The detailed categorization and characteristics are shown as follows [2, 7]:

- **Flaccid Dysarthria:** This type of dysarthria is associated with damage to the lower motor neurons, leading to weakness, and reduced muscle tone in the speech muscles. Speech may exhibit characteristics such as breathiness, imprecise articulation, reduced loudness, and hypernasality.
- **Spastic Dysarthria:** Spastic dysarthria results from bilateral damage to the upper motor neurons, leading to increased muscle tone, spasticity, and reduced range of motion in speech muscles. Speech may exhibit slow rate, strained vocal quality, reduced stress, and harsh or strained voice.
- **Ataxic Dysarthria:** Ataxic dysarthria is characterized by deficits in coordination and control of speech movements due to damage to the cerebellum or its connections. Speech may exhibit irregular articulatory breakdowns, excessive or irregular variations in pitch and loudness, and difficulties with prosody and rhythm.
- **Hypokinetic Dysarthria:** Hypokinetic dysarthria is predominantly associated with movement disorders, particularly Parkinson's disease, characterized by reduced movement and rigidity. Speech may exhibit rapid rate, reduced loudness (hypophonia), monopitch, monoloudness, and imprecise articulation.
- **Hyperkinetic Dysarthria:** Hyperkinetic dysarthria involves involuntary movements affecting speech production, resulting in variable speech rate, irregular articulatory breakdowns, and dysfluencies. Speech may be characterized by hypernasality, voice tremor, and abnormal prosody.
- **Mixed Dysarthria:** Mixed dysarthria presents features of more than one type of dysarthria, often resulting from complex neurological damage affecting multiple neural pathways. Speech characteristics may vary depending on the combination of underlying deficits.

It could also categorized based on the severity or speech intelligibility[13].

- **High Intelligibility:** Intelligibility between 76% and 100%.

- **Mid Intelligibility:** Intelligibility between 51% and 75%.
- **Low Intelligibility:** Intelligibility between 26% and 50%.
- **Very Low Intelligibility:** Intelligibility between 0% and 25%.

2.2. DYSARTHRIC TO NORMAL SPEECH CONVERSION

Dysarthric to normal speech conversion means transforming the speech influenced by dysarthria into normal and intelligible speech. The following are some voice conversion systems which transform dysarthric to normal speech:

Discover Cross-Domain Relations with Generative Adversarial Networks(DiscoGAN) was first proposed by [14] and then was applied to transforming dysarthric speech into normal speech by [22] along with Mean Square Error (MSE) regularization. In [22], they first extract cepstral features using AHOCODER from both dysarthric speech and corresponding control speech. Then they use Dynamic Time Warping (DTW) [25] to do the time-alignment between dysarthric speech and controlled speech. Then DiscoGAN is used to learn the mapping from dysarthric speech to corresponding normal speech. At last, the speech was reconstructed using AHOCODER again. It was trained and evaluated on the UASpeech corpus and was proven that this voice conversion system outperforms the baseline e Deep Neural Network(DNN)-based system with by 13.16% and 9.64% for male speakers and female speakers respectively. Also, according to the subjective evaluation, the results were shown more natural and intelligible compared to the baseline.

MaskCycleGAN-VC[12] and StarGANv2-VC[15] are proposed to apply to transforming dysarthric to normal speech in [17]. It is proven that StarGANv2-VC outperforms MaskCycleGAN-VC in transforming dysarthric speech into normal speech and was shown as an efficient method to enhance the quality of dysarthric speech. The detailed information about these two architectures will be introduced later.

Besides, there is also Fuzzy ASC-GAN[10] which integrates CycleGANv2-VC and Fuzzy C-means clustering to convert dysarthric speech to normal speech, which shows the potential to help patients with dysarthric voices, achieving an average accuracy of 93.35% in S2T evaluation.

2.3. GENERATIVE ADVERSARIAL NETWORK

The Generative Adversarial Network is a kind of deep learning algorithm which is proposed by [9] in 2014. It learns the distribution of given training samples in order to generate new data samples that are similar to the given ones. It is formed by two neural networks which are called generator and discriminator.

2.3.1. GENERATOR

The generator usually takes a random distribution or a latent vector as an input. As the model is training, it will learn to generate data samples that are similar to the training data samples.

2.3.2. DISCRIMINATOR

The discriminator acts as a classifier which is to distinguish between real data from the training data set and fake data generated from the generator.

During the training process, the generator and the discriminator will be trained together in a competitive manner, which means the generator will try to generate data samples similar to real data to fool the discriminator, but the discriminator will try to distinguish between real data and fake data.

2.3.3. OBJECTIVE FUNCTION

The GAN is trained through a minimax game, where the Generator tries to minimize a function while the Discriminator tries to maximize it. The objective function is shown below:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{2.1}$$

For here, p_{data} represents the distribution of the real data, and $p_z(z)$ represents the random input from the generator.

2.3.4. TRAINING

The training process alternates between the following two steps:

UPDATING THE DISCRIMINATOR

Optimize D to maximize the probability of correctly classifying real and generated data.

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D(x^{(i)}) + \log\left(1 - D(G(z^{(i)}))\right)\right] \tag{2.2}$$

UPDATING THE GENERATOR

Optimize G to minimize the probability that D can correctly classify the generated data as fake. This is equivalent to minimizing the following:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D(G(z^{(i)}))\right) \tag{2.3}$$

2.4. STARGAN

2.4.1. STARGAN

StarGAN[4] is an innovative approach in the field of image-to-image translation, particularly in scenarios where you want to translate an image from one domain to another among multiple domains using only a single model. This is a significant advancement over previous methods that typically required training individual models for each domain translation pair, making the process much more efficient and scalable. The core idea behind StarGAN is to enable a single neural network model to learn mappings among multiple domains. For instance, in facial attribute modification (e.g., changing

hair color, age, or gender), rather than training a separate model for each attribute modification, StarGAN can handle all these transformations using one model. This is achieved through the use of a domain label that specifies the target domain for image translation.

StarGAN is also composed of two neural networks: a generator (G) and a discriminator (D). The generator takes an input sample and a target domain label as input and generates an sample that aims to belong to the target domain. The discriminator aims to discriminate between real and fake samples and classify the data into the corresponding domain class. Figure 2.1 shows the details of the training process.

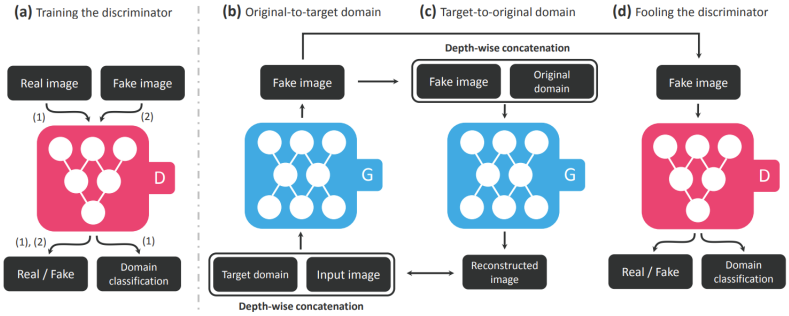


Figure 2.1: Overview of StarGAN from [4]

Adversarial Loss

In order to make the fake data indistinguishable from the real data, the adversarial loss is designed like this:

$$L_{adv} = \mathbb{E}_x[\log D_{src}(x)] + \mathbb{E}_{x,c}[\log(1 - D_{src}(G(x,c)))] \tag{2.4}$$

In this case, $D_{src}(x)$ represents the real data distribution probability, and the $G(x,c)$ represents the sample generated by the generator based on the input sample x and input label c. The generator will try to minimize this objective function, but the discriminator will try to maximize it.

Domain Classification Loss

The domain classification loss is to help the generator produce data that not only resembles real data but is also accurately classified to the target domain. There are two classification losses designed to optimize the generator and the discriminator respectively. For the generator, the classification loss is:

$$L_{cls}^{G} = -\mathbb{E}_{x,c'}[\log D_{cls}(c|G(x,c))] \tag{2.5}$$

For the discriminator, the classification loss is:

$$L_{cls}^{D} = -\mathbb{E}_{x,c}[\log D_{cls}(c'|x)] \tag{2.6}$$

RECONSTRUCTION LOSS

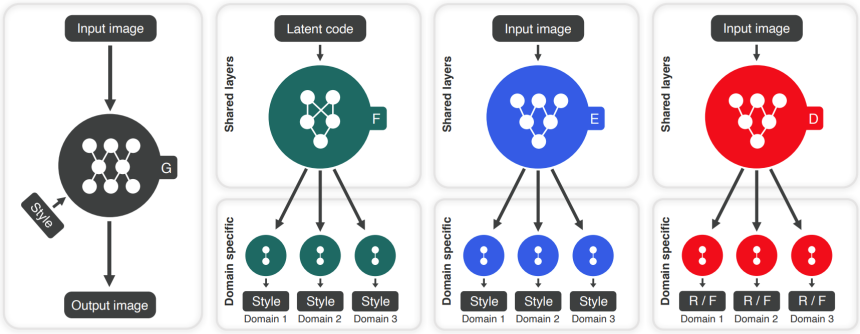
This function is to encourage the generator to produce an output that closely resembles the original sample when reconstructed back to the original domain. This is crucial for maintaining the content of the sample while changing the domain-specific attributes. The loss function is designed as follows:

$$L_{rec} = \mathbb{E}_{x,c,c'}[\|x - G(G(x,c),c')\|_1] \tag{2.7}$$

Here, c represents the target domain, c represents the original domain, and x represents the original input sample. The L1 norm is adopted as reconstruction loss.

2.4.2. STARGAN V2 AND STARGAN V2-VC

StarGAN v2[3] is an advanced deep-learning model designed for image-to-image translation tasks. It is an extension of the original StarGAN architecture, which was developed to perform versatile image translation tasks across multiple domains using a single model. The difference from the original StarGAN is that StarGANv2 uses a style code instead of a fixed label to represent diverse styles of different domains.



(a) Generator (b) Mapping network (c) Style encoder (d) Discriminator

Figure 2.2: Overview of StarGAN v2 from [3]

Here is an overview of the StarGANv2 in Figure 2.2. We can see that there are two components added to the StarGANv2 compared to the StarGAN: the mapping network and the style encoder. The mapping network will generate a style code s based on the input latent vector z and domain label y. The formula is shown below:

$$s = F_y(z) \tag{2.8}$$

In this way, the mapping network will learn the style representation of multiple domains.

The style encoder will extract the style code of the input data x and its domain label y according to the formula below:

$$s = E_y(x) \tag{2.9}$$

In this way, the style encoder will learn the style representation of different reference data.

StarGANv2-VC[15] is an application of the StarGANv2[3] architecture specifically tailored for voice conversion tasks. Voice conversion means to modify the speech characteristics of a source speaker to match those of a target speaker while retaining the linguistic content. StarGANv2-VC[15] extends the capabilities of StarGANv2 to handle voice conversion by employing similar principles but applied to speech signals.

Similar to StarGANv2 which has a generator and discriminator to transfer an sample of one style into another, StarGANv2-VC also has a generator and a discriminator. Each speaker will be treated as a unique domain. The overview of StarGANv2 is shown in Figure 2.3. There are five main components: Generator, F0 network, Style Encoder, Mapping Network, and Discriminator. All the components are similar to the ones in StarGANv2, which we have mentioned in the previous section, except for the F0 network. For a given input mel-spectrogram, the F0 network is to extract its fundamental frequency as input features.

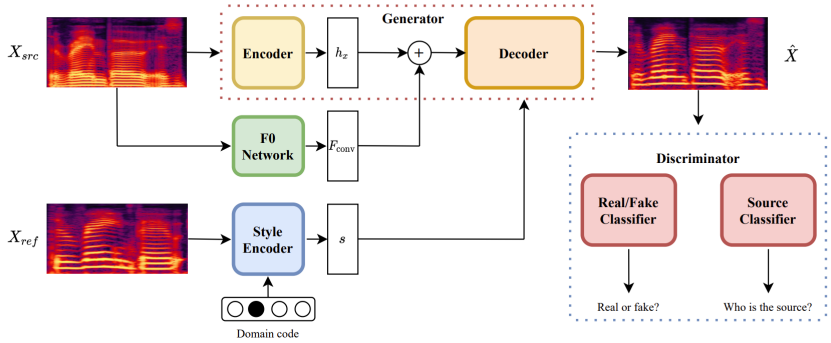


Figure 2.3: Overview of StarGANv2-VC from [15]

Adversarial loss

Given an input mel-spectrogram X and a style code s, the generator will be trained to generate a new mel-spectrogram in the target domain according to adversarial loss. D means discriminator, and G means generator.

$$\begin{aligned} L_{adv} = &\mathbb{E}_{\boldsymbol{X},y_{src}}\left[\log D\left(\boldsymbol{X},y_{src}\right)\right]+ \\ &\mathbb{E}_{\boldsymbol{X},y_{trg},s}\left[\log\left(1-D\left(G(\boldsymbol{X},s),y_{trg}\right)\right)\right] \end{aligned} \tag{2.10}$$

Adversarial source classifier loss

This extra adversarial loss is specifically for the source classifier. $CE(\mathring{u})$ means cross-entropy loss function.

$$L_{\text{advcls}} = \mathbb{E}_{\boldsymbol{X},y_{trg},s}\left[\text{CE}\left(C(G(\boldsymbol{X},s)),y_{\text{trg}}\right)\right] \tag{2.11}$$

STYLE RECONSTRUCTION LOSS

In order to maintain the characteristics of the original speech, StarGANv2-VC uses a style reconstruction loss to make sure the style code can be reconstructed through generated data samples.

$$L_{sty} = \mathbb{E}_{\boldsymbol{X}, y_{trg}, s}\left[\left\| s - S\left(G(\boldsymbol{X}, s), y_{trg}\right)\right\|_1\right] \tag{2.12}$$

STYLE DIVERSIFICATION LOSS

The style diversification loss is used to maximize the difference between generated samples with different styles, including both mean absolute error and also mean absolute error of F0 features.

$$\begin{aligned} L_{ds} = \mathbb{E}_{\boldsymbol{X}, s_1, s_2, y_{trg}}\left[\|G(\boldsymbol{X}, s_1) - G(\boldsymbol{X}, s_2))\|_1\right] + \\ \mathbb{E}_{\boldsymbol{X}, s_1, s_2, y_{trg}}\left[\|F_{\text{conv}}\left(G(\boldsymbol{X}, s_1)\right) - F_{\text{conv}}\left(G(\boldsymbol{X}, s_2)\right))\|_1\right] \end{aligned} \tag{2.13}$$

SUMMARY

To summarize all the objective functions of the generator, the formula is shown below:

$$\begin{aligned} \min_{G,S,M} & L_{adv} + \lambda_{advcls} L_{advcls} + \lambda_{sty} L_{sty} \\ & - \lambda_{ds} L_{ds} + \lambda_{f0} L_{f0} + \lambda_{asr} L_{asr} \\ & + \lambda_{norm} L_{norm} + {}_{cyc}\, L_{cyc} \end{aligned} \tag{2.14}$$

The formula for summarizing all the objective functions of the discriminator is shown below:

$$\min_{C,D} - L_{adv} + \lambda_{cls} L_{cls} \tag{2.15}$$

2.5. MASKCYCLEGAN-VC

MaskCycleGAN-VC[12] is proposed based on the conventional CycleGANv2-VC which overcomes the disadvantage of CycleGANv2-VC that could not capture the characteristics of time-frequency structures. The most important thing about MaskCycleGAN-VC is that it uses a novel auxiliary task called Filling in Frame (FIF). Given an input mel-spectrogram, a temporal mask will be put on the given mel-spectrogram. In this way, the model can learn the time-frequency structures in the process of filling in the frames in the mask area based on the surrounding frames. Also, unlike the CycleGANv3-VC, also a variant of CycleGANv2-VC, MaskCycleGAN-VC will not increase the number of parameters a lot.

As is shown in Figure 2.4, when there is an input x, a mask m which has the same size as x will be created. The black parts of m in the Figure means its value is zero, and the white part means its value is one. And the mask m will be applied to x with element-wise product. In this way, the missing frames will be created since the black region will be zero, while the other parts will keep the same as the original value. Then the concated $\hat{x}$ and m will be put into the Generator to generate y'. The function of m is to tell the converter to fill in which part of the frames. Then another generator will reconstruct the x'' based on generated y' and m' to compare x'' with the original input x. m' means the

mask with all ones and its function is to assume that all the missing frames have already been filled in. In order to make x'' similar to x, the first generator will try to learn how to fill in the missing frames and learn the time-frequency structures during this process.

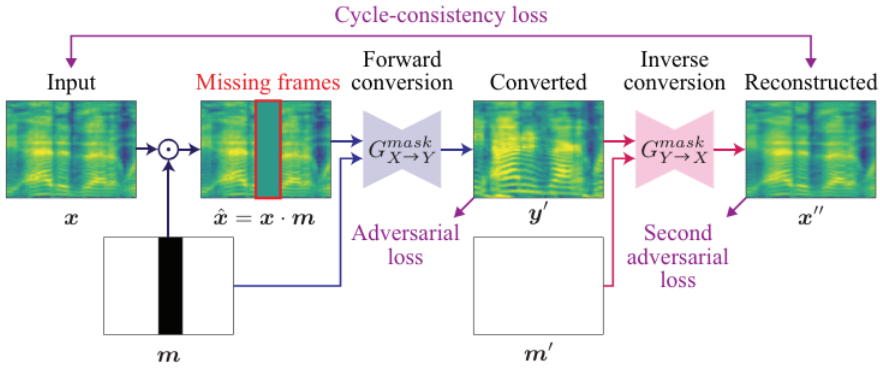


Figure 2.4: Overview of MaskCycleGAN-VC from [12]

2.6. EVALUATION METRIC

2.6.1. PER

Phoneme Error Rate (PER), is a metric used in Automatic Speech Recognition (ASR) systems to evaluate their performance. It measures the accuracy of phoneme-level transcriptions generated by an ASR system compared to a reference transcription. Phonemes are the smallest units of sound that distinguish meaning in a language.

Errors are calculated based on the differences between the aligned phonemes. Errors can be classified into three types:

- Substitution (S): When a phoneme in the ASR output is different from the corresponding phoneme in the reference transcription.
- Deletion (D): When a phoneme in the reference transcription is missing in the ASR output.
- Insertion (I): When a phoneme appears in the ASR output but not in the reference transcription.

The PER will be calculated according to the following formula:

$$PER = \frac{S + D + I}{N} \times 100\% \tag{2.16}$$

Where:

- S: the number of the substitution errors

- D: the number of the deletion errors
- I: the number of the insertion errors

2.6.2. CER

Character Error Rate (CER), is another metric used in ASR systems to evaluate their performance. The difference from PER is that it measures the accuracy of character-level transcriptions compared to reference transcription.

Errors can be also divided into three types the same as PER: Substitution (S), Deletion (D), and Insertion (I). The difference is that CER focuses on the character level.

The CER will also be calculated according to the following formula:

$$CER = \frac{S + D + I}{N} \times 100\% \tag{2.17}$$

3

METHODOLODY

In this section, we outline the methodology adopted to conduct the experiment, encompassing the introduction of the dataset (Section 1), the experimental setup (Section 2), the technological frameworks(Section 3, Section 4, Section5), and the evaluation methods utilized (Section 6).

We will compare two state-of-the-art VC-based techniques (Masked Cycle-GAN and Star-GAN) and two SP-based techniques (time stretching and loudness. Figure 3.1 provides a visual overview of speech conversion methods and evaluation procedures employed in this study. As is shown in Figure3.1, we will first convert dysarthric speech using the methods previously mentioned. And we will do the objective evaluation and subjective evaluation of the converted speech. Finally, we will analyze the correlation between objective results and subjective results. The detailed methodology will be introduced in the following sections.

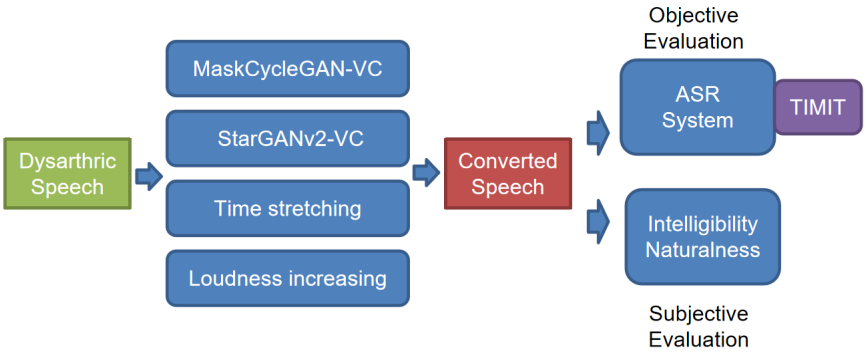


Figure 3.1: Overview of Methodology

3.1. DATASET

In this part, we will introduce the dataset we used in the experiments, including the dataset we use for converting from dysarthric speech to normal speech and also the dataset we use to train our evaluation model.

3.1.1. UASPEECH

UASpeech[13] is a database consisting of dysarthric speech, including recordings from 19 speakers with cerebral palsy. Each speaker contributes 765 isolated words, comprising a combination of 300 distinct uncommon words and repetitions of digits, computer commands, radio alphabet phrases, and common words. The data collection process employs an 8-microphone array alongside a digital video camera to ensure comprehensive recording.

The database is recorded and divided into three distinct blocks, each containing 255 words. Within each block, 155 words are repeated, encompassing digits, radio alphabet letters, computer commands, and common words drawn from the Brown corpus of written English. Additionally, each block includes 100 uncommon words, carefully selected from children's literature using a greedy algorithm to maximize token counts of infrequent biphones.

By providing a repository of dysarthric speech samples, UASpeech serves as an im-

Table 3.1

Speakers	*Intelligibility*
F02	Low
F03	Very low
F04	Mid
F05	High
M05	Mid
M08	High
M09	High
M10	High

portant resource for the advancement of automatic speech recognition technologies for helping individuals with neuromotor disabilities.

UASpeech contains speakers with different types of dysarthria and different levels of severity and allows us to compare different speech conversion techniques for different severities. We followed the selection and dataset split method in [21] with a balanced gender in our training and test sets. Four male speakers (M05, M08, M09, M10) and four female speakers (F02, F03, F04, F05) were used. For controlled speakers, we also chose four male speakers and four female speakers with the same label but they are different speakers from dysarthric ones. The detailed information about the selected speakers is shown in Table 3.1. For our experiment, in order to evaluate the performance on high and low severities, we divided the speech into two groups as follows:

- Low Severity: M05 M08 M09 M10 F04 F05
- High Severity: F02 F03

3.1.2. TIMIT

TIMIT[8] is a corpus specifically designed for research in the development and evaluation of ASR and related fields. The TIMIT dataset encompasses a total of 6300 utterances, with each of the 630 speakers contributing 10 sentences. These speakers represent diverse dialect divisions across the United States, with approximately 70% being male and 30% female. Each speaker's set of 10 sentences amounts to approximately 30 seconds of speech material, resulting in a corpus totaling around 5 hours of speech data. Notably, all speakers are native speakers of American English and have been assessed by a professional speech pathologist to be free of clinical speech pathologies.

Recorded using two microphones, the TIMIT dataset consists of 2-channel recordings. Initially digitized at a sampling rate of 20 kHz, the speech signals were subsequently subjected to digital filtering, debiasing, and downsampling to 16 kHz. This preprocessing ensures the quality and consistency of the recorded speech data, and lays a solid foundation for robust and reliable analyses in automatic speech recognition and related research.

To make our results more comparable to the literature [21], we used TIMIT [8] to train the ASR system for the objective evaluations. Since TIMIT is an English corpus, it allows

us to do the evaluation for converted UASpeech data samples which are also English.

3.2. SIGNAL-PROCESSING TECHNIQUES

The dysarthric speech of TIMIT is converted using the two signal processing approaches and the two VC methods:

LOUDNESS INCREASING (LI)

Dysarthric speech typically exhibits lower loudness compared to normal speech[27]. Therefore, the primary purpose of employing the loudness-increasing method (LI) is to make the loudness of dysarthric speech comparable to the average loudness of normal speech. Here, we use two variations of the loudness increase implementation:

- LI.1: The average loudness of normal speech over all control samples is estimated. Then we set the loudness of each dysarthric speech sample with a lower loudness to the average loudness of normal speech. The detailed algorithm is shown below:

for $Speaker_i$ in dysarthric speakers:

for $Speech_j$ in $dysarthric\ Speaker_i$:

$Loudness_{ij} = 10 * log10\ (rms(Speech_j^2))$

$loudness\_diff = Loudness_{ij} - average\_loudness$

$Converted\_Speech_j = Speech_j * (10 * * (loudness\_diff/20))$

- LI.2: The loudness of dysarthric speech and the corresponding normal speech is estimated for each file separately. Then the loudness of the dysarthric speech was set to the loudness of the normal speech in case it was lower.

for $Speaker_i$ in dysarthric speakers and $CSpeaker_i$ in controlled speakers:

for $Speech_j$ in dysarthric $Speaker_i$ and $CSpeech_j$ in controlled $CSpeaker_i$:

$Loudness_{ij} = 10 * log10(Speech_j^2)$

$loudness\_diff = Loudness_{ij} - CLoudness_{ij}$

$Converted\_Speech_j = Speech_j * (10 * * (loudness\_diff/20))$

LOUDNESS NORMALIZATION (LN)

The loudness normalization method (LN) method aims to match the loudness of every dysarthric speech sample, thus not only those with a lower loudness than normal speech, to be the same as the average loudness of normal speech. Here, we use two variations of the loudness increase implementation:

- LN.1: The average loudness of normal speech over all control samples is estimated. All dysarthric speech files are replaced with the average loudness of normal speech.

- LN.2: We calculate the loudness of dysarthric speech and the corresponding normal speech. Then the loudness of the dysarthric speech sample is set to that of the corresponding normal speech.

3.3. TIME STRETCHING (TS)

The primary objective of time-stretching (TS) is to match the duration of dysarthric speech with that of the corresponding normal speech. The TS factor is determined by the ratio of the duration of dysarthric speech to the duration of normal speech. For instance, if the dysarthric speech lasts 4 seconds and the normal speech is 2 seconds, the TS factor would be 2. We apply TS using the *librosa.effects.time_stretch* function with the estimated factor and ensure that the dysarthric speech matches the duration of its corresponding normal speech.

3.4. GAN-BASED VC MODELS

Each VC model is trained and evaluated on dysarthric speech using a leave-one-speaker-out cross-validation scheme. All the combinations of the training set and test set are shown in Table 3.2.

Training Set	Evaluation Set
M08, M09, M10	M05
M05, M09, M10	M08
M05, M08, M10	M09
M05, M08, M09	M10
F03, F04, F05	F02
F02, F04, F05	F03
F02, F03, F05	F04
F02, F03, F04	F05

Table 3.2: Training set and evaluation set combinations

MaskCycleGAN-VC, an extension of CycleGAN-VC2, aims to transform the speech characteristics of a source speaker into those of a target speaker. Unlike traditional VC methods that often require paired data of source-target speaker pairs, MaskCycleGAN-VC only requires unpaired data from both speakers. MaskCycleGAN-VC is trained using a technique called filling in frames (FIF), which means we will randomly select a part of the mel-spectrogram to mask and try to fill in the missing part when training the model. Compared with CycleGAN-VC2, there is no additional module needed to learn the time-frequency structures, and it has better conversion performance. In our experiments, we use the same experiment settings as in [12]. The VC model was trained for 300 epochs.

3.4.1. STARGANV2-VC

The state-of-the-art StarGANv2-VC model generalizes to a variety of VC tasks, such as many-to-many, cross-lingual, and singing conversion. It uses an auxiliary input c to control the output of generators while CycleGAN only learns a direct mapping between two domains. c denotes an attribute label, represented as a concatenation of one-hot vectors, each of which is filled with 1 at the index of a class in a certain domain and with 0 everywhere else. In our experiment, there are two domains: dysarthric speech and normal speech. We will follow the implementation of [15] and preprocess the data to fit this model. We first join the audio together and then split them into 5-second long files since the common length of speech in UASpeech[13] is about 1 to 2 seconds. The StarGANv2-VC is trained for 150 epochs.

3.5. EVALUATION

3.5.1. OBJECTIVE EVALUATION (PHONEME ERROR RATE (PER))

We use the KALDI Toolkit [20] to build a Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) based ASR system. The ASR is trained with 39 Mel Frequency Cepstral Coefficients (MFCC) and the model is trained until the tri3 stage (i.e., tri-phone based model using Linear Discriminant Analysis (LDA), Maximum Likelihood Linear Transform (MLLT), and Speaker Adaptive Training (SAT)).

The ASR system is used for the objective evaluation of the original dysarthric speech, normal speech, and converted speech. The results are reported in PER.

3.5.2. SUBJECTIVE EVALUATION (NATURALNESS AND INTELLIGIBILITY)

We set up our subjective evaluation experiments as follows. Based on the best-performing methods from our objective evaluation, we used the two best-performing loudness conversion methods, TS, and the best-performing VC method and compared those for naturalness and intelligibility to the original dysarthric speech and normal speech.

EXPERIMENT SETUP

The experiments were run on the Qualtrics Platform [23], and the participants were recruited through Prolific and volunteers. The experiments were carried out separately for naturalness and intelligibility to avoid confusion with the participants on the specific task and to avoid results from potentially influencing each other.

In each experiment, the converted speech is divided into 6 different blocks. We choose a normal speech block, a dysarthric speech block, LI(1 and 2) or LN(1 and 2) blocks, a TS block, and one last block from StarGANv2-VC or MaskCycleGAN-VC. In total, we will have 6 blocks in total. And for each block, we have F02, F03, F04, F05 four female speakers and M05, M08, M09, and M10 four males speakers. In total, we have 8 speakers for each block. For each speaker, we randomly choose 10 samples. So there will be 80 speech samples for each block.

The reason why we set the number in this way is that the samples have an average duration of 1-2 seconds and with a goal to complete the subjective test in one hour, we use 10 samples for each speaker. In total, we chose 480 sample speeches for evaluation. However, as there are only 455 unique words in the UASpeech dataset, there was some overlap between the words of different blocks. To decrease the influence of this kind of overlap, repeated words only occurred in the normal speech block. The normal speech was always the last block of the experiment. All other 5 blocks appeared randomly for each participant and the samples in each block were also presented in random order for each participant.

NATURALNESS EXPERIMENT SETTING

There are 10 participants recruited for the naturalness evaluation experiment. They are recruited via social media. Seven of them are female, two of them are male, and one of them prefer not say about his or her gender. All of the participants are not native speakers of English. The first languages they speak include Chinese, Dutch, Flemish, and Hindi. The ages of the participants range from 25 to 26. The subjective evaluation forms

are distributed via the URL link, so the participants can complete their forms online at wherever they are.

For the naturalness evaluation, mean opinion score (MOS)[**mos**] is used. The participants are provided with a definition of naturalness, i.e., it is defined as the extent to which the speech sample sounds like that of a normal human speaker in terms of e.g., intonation, voice quality, speaking rate, rhythm, and intensity. After listening, participants were asked to rate the sample speech on a five-point scale: 1 for "Bad," 2 for "Poor," 3 for "Fair," 4 for "Good," and 5 for "Excellent." They were made aware that some of the speech samples will sound highly natural while others may sound artificial due to deterioration caused by computer processing. Then the participants are asked to carefully listen to each speech sample and provide a rating of the speech sample's naturalness. The detailed introduction we use in the experiment is shown in Figure 3.2 below.

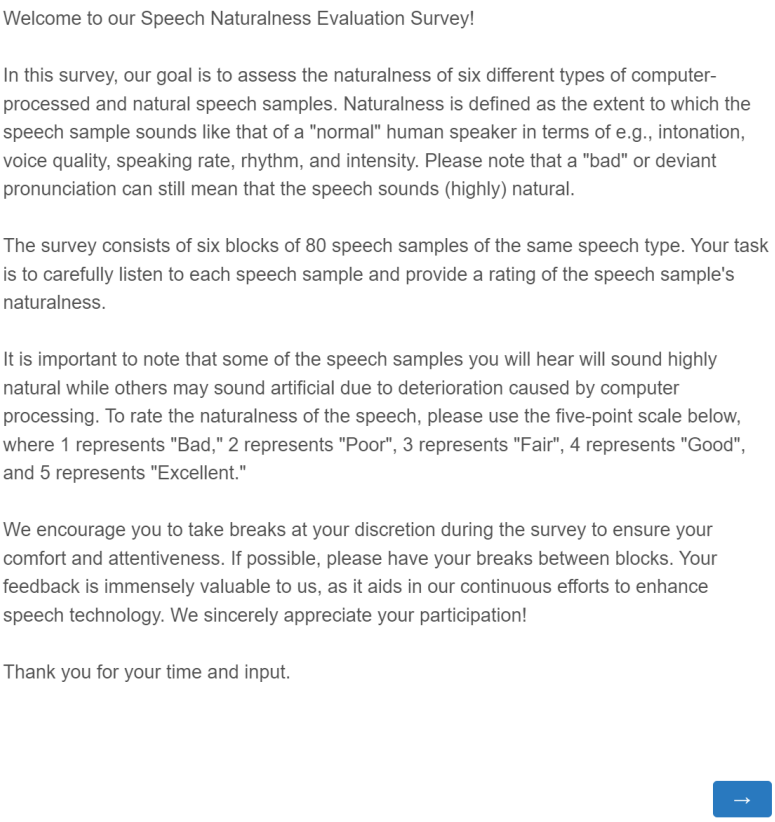


Figure 3.2: Instructions of naturalness experiment

INTELLIGIBILITY EXPERIMENT SETTING

There are 10 participants recruited for the intelligibility evaluation experiment. They are recruited via both social media and the Prolific recruiting Platform. Nine of them are recruited via Prolific with payment and one of them joins as a volunteer. Seven of them are female and three of them are male. Nine of the participants are native speakers of English. The first language of the other speaker is Dutch. The ages of the participants range from 21 to 59. The participants also finish the evaluation form online at wherever they are.

For the intelligibility evaluation, the participants are required to carefully listen to each speech sample and type in the words they heard.

We also inform participants to carefully check the spelling of their answers. We used the character error rate (CER) as our evaluation metric (instead of the PER we use for objective evaluation because the participants are not professional linguists who can convert words into phonemes). In order to avoid the influence of lower-case and upper-case in error rate evaluation, we converted all answers into lower-case. Figure **??**

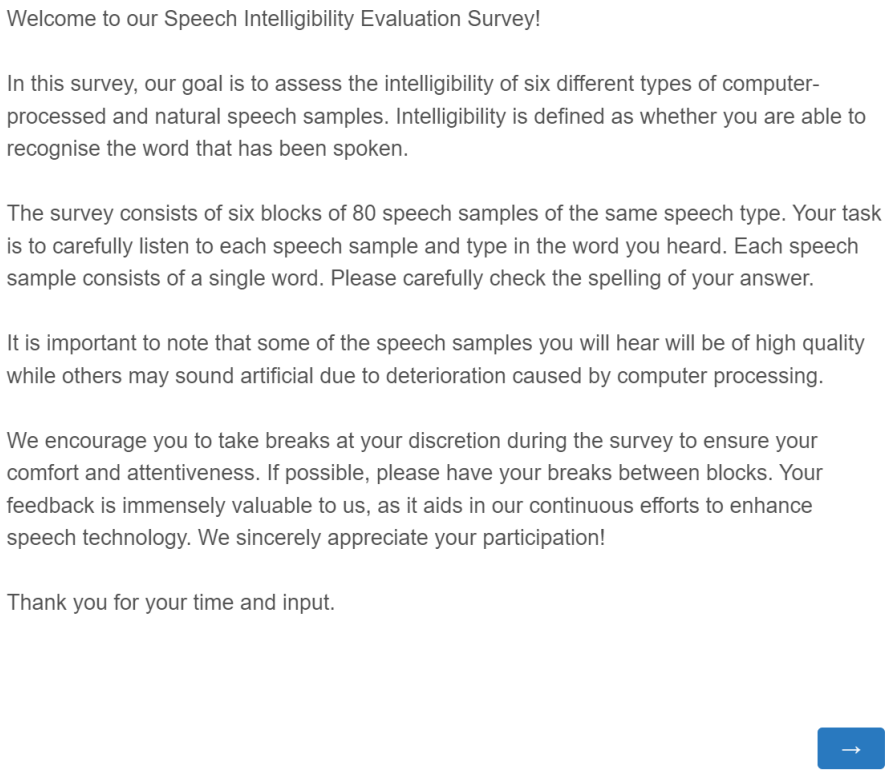


Figure 3.3: Instructions of intelligibility experiment

4

RESULTS

In this section, we will first show the results of the objective evaluation of the four speech conversion techniques, including the baseline results of our evaluation system, the performance of loudness increasing method, time stretching method, MaskCycleGAN-VC, and StarGANv2-VC. This will help us to answer RQ1. Then we will describe the experimental results of subjective results, focusing on intelligibility and naturalness, to answer RQ2. Then, we will show the results of the correlation between objective evaluation and subjective evaluation.

4.1. RESULTS OF THE OBJECTIVE EVALUATION

4.1.1. EVALUATION SYSTEM

Before comparing the four speech conversion techniques, we will first show the result of our evaluation system which will indicate how well our evaluation system performed on the TIMIT dataset. The result on the test set is 21.6%. Our converted speech samples will be evaluated with this evaluation system. For the comparison result by [21], the result is 21.8%. The results are close and make it easy to compare our results to the results of [21].

Table 4.1 shows the ASR performance for normal speech, dysarthric speech, and converted speech through all methods we use. The PER results are initially calculated for each speaker. Subsequently, we categorize these results into high-severity and low-severity groups. For normal speech, we do not have separate high-severity and low-severity groups, so these results remain undivided and will be empty for those categories.

The results table is organized into four blocks:

- Normal speech and dysarthric speech
- Converted speech using signal processing techniques
- Converted speech using voice conversion (VC) models
- Comparison results from [21]

For normal speech, our PER results for all controlled speakers are better than those reported in the experiment by [21]. This improvement is also observed in the average results for both high-severity and low-severity groups.

Regarding dysarthric speech, although the average results for both high-severity and low-severity groups are better than those from [21], some individual speakers do not follow this trend. Specifically, speakers F04, M08, and M09 show worse results, with average decreases of 0.6% for F04, 1.7% for M08, and 0.6% for M09.

This difference between the results of our work and the results reported in [21] in normal and dysarthric speech might be caused by the use of the second version of UASpeech in our experiment which is noise-reduced and gets rid of the influence of the noise signals.

In the following sections, I will present the results of each method separately and conclude with a comparison of all methods.

4.1.2. LOUDNESS INCREASING

LI.1

We can see the results of LI,1 from the second block from Table 4.1. For both the high-severity and low-severity groups, the average values are worse than the original average value. In the high-severity group, all individuals experienced worse results. However, in the low-severity group, we observed improvements for speakers F03, M05, M08, and M09. Specifically, the absolute PER for speaker F03 was reduced by 0.5% compared to the original dysarthric speech. For speaker M05, the absolute PER was reduced by 0.2%, for speaker M08 by 0.2%, and for speaker M09 by 0.4% compared to the original dysarthric speech.

LI.2

From the results of LI.2 shown in the second block of Table 4.1, we observe improvements in the average value for the high-severity group. All individuals in the high-severity group show improvements. Specifically, the absolute phoneme error rate (PER) of speaker F02 was reduced by 0.5% compared to the original dysarthric speech, and the PER of speaker F03 was reduced by 0.9%.

Although LI.2 degrades the average value for the low-severity group, there are still some individual improvements. The absolute PER of speaker F04 was reduced by 1.9%, and for speaker M05, it was reduced by 0.6% compared to the original dysarthric speech. Notably, the PER results for speakers F05 and M10 increased significantly compared to other speakers. The PER for speaker M05 increased by 3.7%, and for speaker M10, it increased by 2.9%.

LN.1

According to Table 4.1, the average performance for both the high-severity and low-severity groups is slightly degraded after applying LN.1. However, there are some individual improvements. In the high-severity group, speaker F03 shows improvement with a reduction in PER of 0.5%. In the low-severity group, the PER for speaker F04 is reduced by 0.2%. Despite this improvement, LN,1 did not significantly enhance ASR performance for most speakers in the low-severity group.

LN.2

In Table 4.1, the ASR performance of converted speech using LN.2 is presented. For the high-severity group, there is an average improvement with a decrease of 0.6% in PER. Specifically, speakers F02 and F03 both show an improvement of 0.6%. However, for the low-severity group, the average result is degraded. Although we observed an improvement for speaker F04 with a decrease of 2.1% in PER, LN.2 did not show enhancement in ASR performance for most speakers in the low-severity group.

4.1.3. TIME STRETCHING

Following the application of the time-stretching technique, improvements are evident for both the high-severity and low-severity groups, as indicated in Table 4.1. In the high-severity group, there is an average improvement of 18.3%. Specifically, speaker F02 shows an improvement of 27%, and for the low-severity group, the improvement is 9.4%.

However, not all individuals in the low-severity group experience improvements. Speakers M08 and M09 exhibit worse performance compared to the original dysarthric speech. This aligns with the findings reported in [21]. The average performance of both the high-severity and low-severity groups in our experiment closely resembles the results reported in [21]. Regarding individual speakers, while M08 in our experiment demonstrates worse results, similar to [21], M09 not only fails to show improvement but also exhibits worse performance compared to the original dysarthric speech. This discrepancy might be attributed to differences in the evaluation system between our experiment and [21]. For the original dysarthric speech, our results indicate a worse performance for speaker M09 compared to the results reported in [21].

Table 4.1: Objective Results (PER) for original dysarthric speech and converted dysarthric speech by various techniques for individual speakers. The best results for each speaker across all techniques are highlighted in bold.

	High Severity			Low Severity						
	F02	**F03**	**Avg-High**	**F04**	**F05**	**M05**	**M08**	**M09**	**M10**	**Avg-Low**
Normal Speech	55.7	60.8	-	72.0	51.9	50.2	48.0	56.0	53.7	-
Dysarthric Speech	109.0	89.3	99.2	80.5	82.5	95.7	63.3	70.6	68.9	76.9
Loudness LI.1	109.2	89.9	99.6	80.0	87.2	95.5	**63.1**	**70.2**	71.8	78.0
Loudness LI.2	108.5	88.4	98.4	78.6	86.2	95.1	63.3	70.8	71.8	77.6
Normalization LN.1	110.3	88.8	99.8	80.3	87.0	98.1	65.2	71.9	71.8	79.05
Normalization LN.2	108.4	88.7	98.6	78.4	86.3	96.8	64.2	70.9	71.8	78.1
TS	**82.0**	**79.9**	**80.9**	**75.2**	**69.3**	**74.1**	68.9	71.8	**65.0**	**70.7**
MaskedCycleGAN	120.1	96.3	108.2	81.9	95.1	99.7	69.5	74.5	78.0	83.1
StarGAN	117.8	88.9	103.4	80.3	89.0	102.0	72.7	74.2	71.8	81.7
Normal Speech by [21]	56.9	61.6	-	74.0	53.1	53.9	48.2	57.6	55.6	-
Dysarthric Speech by [21]	109.0	89.8	99.4	79.9	85.9	94.0	64.1	70.0	68.8	77.1
TS by [21]	81.8	79.3	80.55	75.5	67.9	76.6	68.9	70.4	65.4	70.8
MaskCycleGAN by [21]	116.9	96.3	106.6	78.8	89.9	102.5	73.8	77.1	67.3	81.5

4.1.4. MASKCYCLEGAN-VC

According to the results of MaskCycleGAN-VC in the third block of Table 4.1, we observed degraded results in both the high-severity and low-severity groups. This trend is also evident for individual speakers within each group, with all individuals showing worse results. In [21]'s experiments, the average performance for both the high-severity and low-severity groups is similar to ours. However, it's noteworthy that although the average results for both groups are worsened, some individuals in the low-severity group, such as F04 (improvement of 1.7%) and M10 (improvement of 1.6%), still show improvements.

One possible explanation for this difference is that during the training process of MaskCycleGAN-VC, the missing frames are randomly selected, leading to slight differences in the performance of our model compared to the model in [21].

4.1.5. STARGANV2-VC

Based on the results of StarGANv2-VC, we observe that the average ASR performance is degraded for both the high-severity and low-severity groups. However, looking at individual results, speaker F02 in the high-severity group shows an improvement with a decrease in PER of 0.4% compared to the original dysarthric speech. In the low-severity group, speaker F04 also shows a marginal improvement, with PER decreasing from 80.5% to 80.3%.

In conclusion, StarGANv2-VC did not demonstrate effectiveness in improving ASR performance for most speakers in either the high-severity or low-severity groups.

4.1.6. OVERALL

To compare all the methods we use, we look into the whole Table 4.1. For both the low-severity and high-severity groups, time-stretching (TS) led to substantial improvements in ASR recognition performance compared to the original dysarthric speech, while both

Table 4.2: Subjective results (MOS) for original dysarthric and converted dysarthric speech by various techniques for individual speakers.

	High Severity			Low Severity						
	F02	**F03**	**Avg-High**	**F04**	**F05**	**M05**	**M08**	**M09**	**M10**	**Avg-Low**
Normal Speech	4.01	3.9	-	3.78	4.15	4.07	3.95	3.84	3.93	-
Dysarthric Speech	2.68	2.36	**2.52**	**3.23**	3.75	**2.81**	**3.76**	**3.34**	**3.88**	**3.47**
Loudness LI.1	2.14	2.01	2.08	3.02	**3.84**	**2.81**	3.63	3.11	3.85	3.38
Loudness LI.2	**2.40**	2.16	2.28	2.76	3.80	2.58	3.40	2.97	3.56	3.18
TS	1.90	1.01	1.78	2.62	3.13	2.31	3.43	2.88	3.58	2.99
StarGAN	2.14	**2.65**	2.08	3.02	**3.84**	**2.81**	3.63	3.11	3.85	3.38

Table 4.3: Subjective results (CER) for original dysarthric and converted dysarthric speech by various techniques for individual speakers.

	High Severity			Low Severity						
	F02	**F03**	**Avg-High**	**F04**	**F05**	**M05**	**M08**	**M09**	**M10**	**Avg-Low**
Normal Speech	9.36	15.63	-	24.27	20.20	6.30	12.02	13.47	14.81	-
Dysarthric Speech	64.25	91.94	80.79	**53.06**	**21.69**	**38.48**	**17.81**	53.47	21.64	**36.23**
Loudness LI.1	78.56	**82.39**	81.45	64.27	23.13	52.92	37.43	63.78	**16.34**	44.00
Loudness LI.2	**63.88**	94.07	**77.57**	58.78	28.54	39.31	19.35	**37.40**	32.92	39.69
TS	91.91	98.11	95.02	81.01	42.07	79.60	41.71	54.42	29.29	56.27
StarGAN	91.18	98.38	95.46	87.07	71.63	77.77	77.61	71.37	86.97	79.13

VC methods led to degradations in performance. In line with earlier findings by [21], time-stretching outperformed state-of-the-art VC methods for improving dysarthric speech recognition. Although LI.2 and LN.2 also improve the performance of the high-severity group, the improvements are slight. Also, Of the four loudness methods, two increasing the loudness techniques gave overall slightly better results than normalizing loudness, and thus are used for the subjective evaluations. Of the two VC methods, StarGANv2-VC gave slightly better results than MaskedCycle-GAN and is thus used for subjective evaluation.

4.2. RESULTS OF THE SUBJECTIVE EVALUATION

4.2.1. NATURALNESS EVALUATION

Table 4.2 shows the results of the listening experiment for the naturalness evaluation. It consists of the converted speech using different speech conversion methods, the dysarthric speech, and the normal speech again for the speakers individually and averaged for the two severities. Table 4.2 shows that both normal speech and dysarthric speech achieve higher scores compared to converted speech. Additionally, normal speech scores higher than dysarthric speech, indicating that listeners perceive normal speech as more natural than dysarthric speech, even though neither is distorted by any method.

From the results of LI.1 in Table 4.2, we see that the average performance for both high-severity and low-severity groups is decreased. However, speaker F05 in the low-severity group shows a slight improvement. For LI.2, a similar trend is observed, with average performance decreasing for both severity groups. Nonetheless, individual improvements are noted for speaker F02 in the high-severity group and speaker F05 in the low-severity group. These results suggest that increasing the loudness of dysarthric speech does not generally make it sound more natural for most speakers in either severity group.

The time-stretching technique results in a decrease in the average MOS score compared to the original dysarthric speech in both high-severity and low-severity groups. No individual improvements are observed, indicating that altering the speech rate (speeding up or slowing down) does not enhance the naturalness of dysarthric speech and may even worsen it.

StarGANv2-VC also degrades the average performance for both severity groups, but some slight individual improvements are noted in both high-severity and low-severity groups.

In conclusion, among the converted speech samples, the loudness LI.2 method achieved the highest naturalness scores for high-severity dysarthric speech, while for low-severity dysarthric speech, the highest MOS scores were obtained with the LI.1 loudness method and the VC model.

4.2.2. INTELLIGIBILITY EVALUTION

Table 4.3 shows the intelligibility results in Character Error Rate (CER). The results demonstrate that normal speech consistently achieves the best results, indicating that normal speech is more intelligible than both dysarthric speech and converted dysarthric speech.

For the results of LI.1, we observe an increase in average CER for both high-severity and low-severity groups. However, there are individual improvements: speaker F03 in the high-severity group shows an improvement of 9.55%, and speaker M10 in the low-severity group shows an improvement of 5.3%. For LI.2, the average performance of the high-severity group improves, but the performance for the low-severity group worsens compared to dysarthric speech.

Both the time-stretching (TS) technique and StarGAN worsen the average performance of both severity groups. Additionally, there are no individual improvements for these two methods, indicating that they make the converted speech less intelligible for human listeners.

To conclude, for high-severity speakers, the LI.2 method yields better results than the original dysarthric speech. However, on average, none of the speech conversion methods improved intelligibility for the low-severity speakers. Among all the speech conversion techniques, the LI.2 method delivers the best performance for both high-severity and low-severity groups, while the VC model results in the worst performance for both groups.

4.3. CORRELATION BETWEEN OBJECTIVE AND SUBJECTIVE RESULTS

Figure 4.1 consists of two figures. The first one shows the correlation between naturalness and ASR performance. The x-axis means the MOS Score which measures the naturalness of the speech. The y-axis means the PER(%) which measures the ASR performance of the speech. Different colors of dots represent dysarthric speech and speech using different converting methods. The lines will show the prediction of the trend of the correlation between naturalness and ASR performance. For all methods which are shown in Figure 4.1, higher naturalness scores in the subjective evaluation led to better ASR performance.

The second one shows the correlation between intelligibility and ASR performance. The x-axis means the CER (%) which measures the intelligibility of the speech. The y-axis means the PER(%) which measures the ASR performance of the speech. The dots and the lines function the same as the first figure. For all methods which are shown in Figure 4.1, higher intelligibility in the subjective evaluation led to better ASR performance.

In order to further investigate the correlation between objective and subjective results. We then computed the correlation coefficients and P value between the objective and subjective results for all methods. The correlation coefficient[1] is a number range from -1 to +1. When the correlation coefficient is between -1 and 0, it indicates that there is a negative correlation between the variables; when the correlation coefficient is between 0 and 1, it indicates that there is a positive correlation between the variables; when the correlation coefficient is 0, there is no correlation between the two. When the correlation coefficients get more close to 1 or -1, it indicates a stronger relationship. What is noticeable that the objective results are averaged over all recording samples (as the PER can be estimated for all the available data) while the subjective results are averaged over the tested samples for which the naturalness experiments were completed by the listeners.

Table 4.4 shows the correlation coefficients for dysarthric speech and converted speech using all methods. Based on these results, the correlation coefficients of PER and MOS score for all methods are negative. This indicates that a higher naturalness score is associated with a lower PER, implying a better ASR performance. However, the significance of this correlation varies among methods. The time-stretching (TS) method exhibits the strongest correlation, with a p-value smaller than 0.01, indicating a statistically significant relationship. In contrast, the correlations for LI.1 and LI.2 are not strong or significant.

As shown in Table 4.4, the correlation coefficients of PER and CER for all methods are positive. This suggests that a lower CER is associated with a lower PER, indicating better

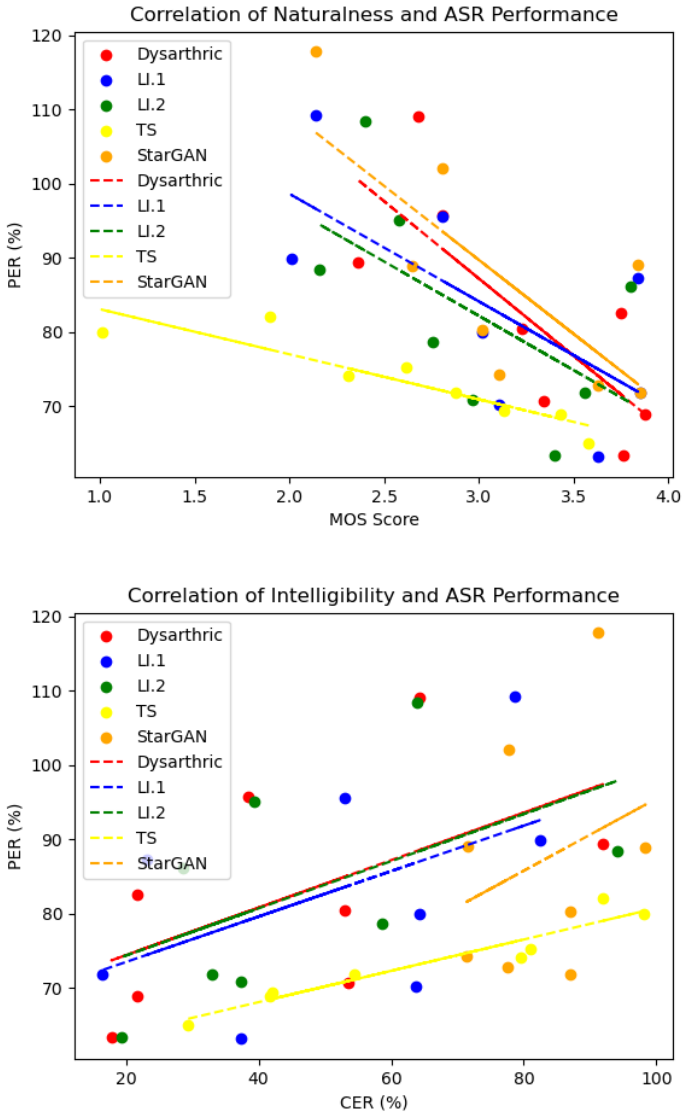


Figure 4.1: Correlation between MOS and PER (top) and CER and PER (bottom) for different methods

Table 4.4: Correlaitonship between Intelligibility Results and ASR Performance for different methods

	PER & MOS		PER & CER	
	Correlation Coefficient	P Value	Correlation Coefficient	P Value
Dysarthric Speech	-0.77	0.024	0.54	0.16
LI.1	-0.68	0.060	0.50	0.21
LI.2	-0.58	0.12	0.52	0.18
TS	-0.91	0.0018	0.96	0.0002
StarGAN	-0.75	0.030	0.29	0.48

ASR performance. Again, the significance of this correlation varies. The TS method has the strongest correlation, with a p-value smaller than 0.01, showing that this relationship is statistically significant. For the LI.1, LI.2, and StarGANv2-VC models, the correlations are neither strong nor significant.

5

DISCUSSION

5.1. DISCUSSION

In our experiment, we choose two signal processing techniques (Loudness Increasing/Normalization and Time Stretching) and two voice conversion models (MaskCycleGAN-vc and StarGANv2-VC) for transforming dysarthric speech into normal speech. We trained one ASR system to do the objective evaluation. We choose two evaluation metrics for subjective evaluation, which are naturalness and intelligibility. The participants were recruited to carry out the subjective evaluation experiments. Finally, we gather all the results to answer the following questions:

- **RQ1:** Which speech conversion technique leads to the highest recognition performance for two severities of dysarthric speech?
- **RQ2:** Which speech conversion technique improves the naturalness and intelligibility of dysarthric speech for human listeners?
- **RQ3:** Does increased naturalness lead to better ASR performance for dysarthric speech?
- **RQ4:** Does increased intelligibility lead to better ASR performance for dysarthric speech?

Table 5.1 shows the all the results we get from objective evaluation and subjective evaluation, making it clear to discuss in this part.

For RQ1, we conducted experiments to compare four different speech conversion techniques. As shown in Table 5.1, time stretching leads to the highest recognition performance in both high-severity and low-severity groups and has a significant advantage compared to the results of other techniques which aligns with the finding in [21]. This shows the effectiveness of improving the ASR performance of dysarthric speech for both the high-severity group and the low-severity group. For loudness modification methods, the results are really close to each other and worse than the results of time stretching.

Both StarGANV2-VC and MaskCycleGAN-VC get worse results than the signal processing techniques in both high-severity and low-severity groups. Specifically, MaskCycleGAN-VC gets the worst result among all techniques. One possible explanation is the speed rate might be a key difference between normal speech and dysarthric speech that leads to the difference in ASR performance of them. The loudness difference between normal and dysarthric speech is not significant and also will not make a big difference in ASR performance. Moreover, increasing the loudness can introduce additional noise to the original speech to make the results worse. Voice conversion models aim to convert all characteristics of normal speech to match those of dysarthric speech. However, some characteristics may not be important for automatic speech recognition, which could explain why these models are not effective in improving dysarthric speech recognition as Time stretching. Additionally, in our experiment, StarGAN outperforms MaskCycleGAN, which might be due to the differences in the vocoders used. StarGAN uses the Parallel WaveGAN vocoder, while MaskCycleGAN uses the MelGAN vocoder.

For RQ2, we observed from the naturalness evaluation results in Table 5.1 that LI.2 (Loudness Increasing 2) gets the best result for the high-severity group and LI.1 (Loudness Increasing 1) and StarGANV2-VC get the best result for the low-severity group. This shows that the distortion of dysarthric speech using loudness increasing is relatively low for both high and low-severity groups. Although there is some noise introduced by loudness increasing, it has a relatively low influence on the perception of human beings. For both high-severity and low-severity groups, the Time Stretching method all get the worst performance. That might be caused by the speed up or slow down of the original speech is easy to recognize for human beings. For Intelligibility evaluation results in Table 4.3, LI.2 (Loudness Increasing 2) achieves the highest performance for both high-severity and low-severity groups. It even gets a better intelligibility performance than original dysarthric speech in the high-severity group and gets a close result with original dysarthric speech in the low-severity group. This indicates that people could avoid the influence of introduced noise brought by the loudness increasing, which align with the finding in [18].

For RQ3, we calculate the correlation coefficients between objective evaluation results and naturalness scores which are shown in Table 4.4. For all the methods, the result shows that higher naturalness scores will lead to better ASR performance. However, for different methods, the significance of correlation varies. For dysarthric speech, time stretching and StarGAN, this kind of correlation is strong and significant. The strong correlation means aiming to improve the naturalness of speech for these methods will lead to better ASR performance. But for loudness increasing, it is not the case. One possible reason is that although increasing loudness will get a relatively high MOS score for human beings, it has little influence on improving the ASR performance. What is found in [18] is that it is easy for human beings to avoid the influence of noise.

For RQ4, we compute the correlation coefficients between objective evaluation results and intelligibility scores, as presented in Table 4.4. For all the methods, the result shows that higher intelligibility scores correlate with better ASR performance. the strength of this correlation varies across different methods. Only for time stretching, the correlation is significantly positive. This shows that we could improve the ASR performance of the converted speech for time-stretching by making it sound more intelligible

Table 5.1: All Results (PER, MOS, CER) for normal speech, original dysarthric speech, and converted dysarthric speech by various techniques for individual speakers. The best results for each speaker across all techniques are highlighted in bold.

	High Severity			**Low Severity**						
	F02	**F03**	**Avg-High**	**F04**	**F05**	**M05**	**M08**	**M09**	**M10**	**Avg-Low**
Normal Speech (PER)	55.7	60.8	-	72.0	51.9	50.2	48.0	56.0	53.7	-
Normal Speech (MOS)	4.01	3.9	-	3.78	4.15	4.07	3.95	3.84	3.93	-
Normal Speech (CER)	9.36	15.63	-	24.27	20.20	6.30	12.02	13.47	14.81	-
Dysarthric Speech (PER)	109.0	89.3	99.2	80.5	82.5	95.7	63.3	70.6	68.9	76.9
Dysarthric Speech (MOS)	109.0	89.3	99.2	80.5	82.5	95.7	63.3	70.6	68.9	76.9
Dysarthric Speech (CER)	64.25	91.94	80.79	**53.06**	**21.69**	**38.48**	**17.81**	53.47	21.64	**36.23**
Loudness LI.1 (PER)	109.2	89.9	99.6	80.0	87.2	95.5	**63.1**	**70.2**	71.8	78.0
Loudness LI.1 (MOS)	2.14	2.01	2.08	3.02	**3.84**	**2.81**	3.63	3.11	3.85	3.38
Loudness LI.1 (CER)	78.56	**82.39**	81.45	64.27	23.13	52.92	37.43	63.78	**16.34**	44.00
Loudness LI.2 (PER)	108.5	88.4	98.4	78.6	86.2	95.1	63.3	70.8	71.8	77.6
Loudness LI.2 (MOS)	**2.40**	2.16	2.28	2.76	3.80	2.58	3.40	2.97	3.56	3.18
Loudness LI.2 (CER)	**63.88**	94.07	**77.57**	58.78	28.54	39.31	19.35	**37.40**	32.92	39.69
TS (PER)	**82.0**	**79.9**	**80.9**	**75.2**	**69.3**	**74.1**	68.9	71.8	**65.0**	**70.7**
TS (MOS)	1.90	1.01	1.78	2.62	3.13	2.31	3.43	2.88	3.58	2.99
TS (CER)	91.91	98.11	95.02	81.01	42.07	79.60	41.71	54.42	29.29	56.27
StarGAN (PER)	117.8	88.9	103.4	80.3	89.0	102.0	72.7	74.2	71.8	81.7
StarGAN (MOS)	2.14	**2.65**	2.08	3.02	**3.84**	**2.81**	3.63	3.11	3.85	3.38
StarGAN (CER)	91.18	98.38	95.46	87.07	71.63	77.77	77.61	71.37	86.97	79.13

for humans. However, for other speech conversion techniques, there is no significant correlation between ASR performance and human intelligibility. This disparity underscores differing perceptual mechanisms between human listeners and ASR systems. Human beings are more robust to speech recognition[18].

5.2. LIMITATIONS AND FUTURE WORK

A limitation of our work is that for subjective evaluation, we could not implement a larger experimental setup. For future work, we can address this issue by selecting more samples for each speaker, adopting additional methods (including all those used for objective evaluation), and recruiting more participants.

In addition, we divided dysarthria into two groups: high severity and low severity. However, the characteristics of dysarthria are more subtle and detailed. Future research can focus on exploring these specific characteristics in greater depth.

6

CONCLUSION

In our study, we conducted a comparative analysis of various signal processing and voice conversion techniques aimed at transforming dysarthric speech into normal speech. The evaluation encompassed objective evaluation and assessments of naturalness and intelligibility. Our findings revealed that among the techniques investigated, time-stretching exhibited superior performance in the objective evaluation experiment, outperforming state-of-the-art voice conversion techniques. For all methods, we observed that increased naturalness and increased intelligibility led to improved ASR performance. However, this kind of correlation is significant for some methods, but not for others. For example, we observed a significant positive correlation between naturalness and ASR performance and between intelligibility and ASR performance for the time-stretching method. This indicates if we could improve the naturalness and intelligibility of the converted speech, it would lead to better ASR performance for this method. Taken together, these results show that aiming to make dysarthric speech more natural sounding and more intelligible has a effect on ASR performance for some methods. For other methods without this correlation, future research should focus on improving (the acoustic modeling of) specific aspects of dysarthric speech for improved dysarthric speech recognition.

BIBLIOGRAPHY

[1] Herve Abdi. "Multiple correlation coefficient". In: *Encyclopedia of measurement and statistics* 648.651 (2007), p. 19.

[2] Hermann Ackermann, Ingo Hertrich, and Wolfram Ziegler. "Dysarthria". In: *The handbook of language and speech disorders* (2010), pp. 362–390.

[3] Yunjey Choi et al. "Stargan v2: Diverse image synthesis for multiple domains". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 8188–8197.

[4] Yunjey Choi et al. *StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation*. 2018. arXiv: 1711.09020 [cs.CV].

[5] Minghang Chu et al. "E-DGAN: An Encoder-Decoder Generative Adversarial Network Based Method for Pathological to Normal Voice Conversion". In: *IEEE Journal of Biomedical and Health Informatics* (2023), pp. 1–14. DOI: 10.1109/JBHI.2023.3239551.

[6] Frederic L Darley, Arnold E Aronson, and Joe R Brown. "Differential diagnostic patterns of dysarthria". In: *Journal of speech and hearing research* 12.2 (1969), pp. 246–269.

[7] Pam Enderby. *Neurological Rehabilitation: Chapter 22. Disorders of communication: dysarthria*. Vol. 110. Elsevier Inc. Chapters, 2013.

[8] John S Garofolo et al. "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1". In: *NASA STI/Recon technical report n* 93 (1993), p. 27403.

[9] Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014).

[10] Yi-Wei Huang et al. "Fuzzy ASC-GAN: A Fuzzy C-means Audio Similarity CycleGAN on Articulation Disorder Voice Conversion". In: *2023 International Conference on Fuzzy Theory and Its Applications (iFUZZY)*. IEEE. 2023, pp. 1–6.

[11] Marc Illa et al. "Pathological voice adaptation with autoencoder-based voice conversion". In: *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*. 2021, pp. 19–24. DOI: 10.21437/SSW.2021-4.

[12] Takuhiro Kaneko et al. "MaskCycleGAN-VC: Learning Non-parallel Voice Conversion with Filling in Frames". In: *CoRR* abs/2102.12841 (2021). arXiv: 2102.12841. URL: https://arxiv.org/abs/2102.12841.

[13] Heejin Kim et al. "Dysarthric speech database for universal access research". In: *Proc. Interspeech 2008*. 2008, pp. 1741–1744. DOI: 10.21437/Interspeech.2008-480.

[14] Taeksoo Kim et al. “Learning to discover cross-domain relations with generative adversarial networks”. In: *International conference on machine learning*. PMLR. 2017, pp. 1857–1865.

[15] Yinghao Aaron Li, Ali Zare, and Nima Mesgarani. “StarGANv2-VC: A Diverse, Unsupervised, Non-parallel Framework for Natural-Sounding Voice Conversion”. In: *CoRR* abs/2107.10394 (2021). arXiv: 2107.10394. URL: https://arxiv.org/abs/2107.10394.

[16] Jaime Lorenzo-Trueba et al. *The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods*. 2018. DOI: 10.48550/ARXIV.1804.04262. URL: https://arxiv.org/abs/1804.04262.

[17] Hadil Mehrez, Mounira Chaiani, and Sid Ahmed Selouani. “Using StarGANv2 Voice Conversion to Enhance the Quality of Dysarthric Speech”. In: *2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. 2024, pp. 738–744. DOI: 10.1109/ICAIIC60209.2024.10463241.

[18] Kinfe Tadesse Mengistu and Frank Rudzicz. “Comparing humans and automatic speech recognition systems in recognizing dysarthric speech”. In: *Advances in Artificial Intelligence: 24th Canadian Conference on Artificial Intelligence, Canadian AI 2011, St. John's, Canada, May 25-27, 2011. Proceedings 24*. Springer. 2011, pp. 291–300.

[19] Seyed Hamidreza Mohammadi and Alexander Kain. “An overview of voice conversion systems”. In: *Speech Communication* 88 (2017), pp. 65–82.

[20] Daniel Povey et al. “The Kaldi speech recognition toolkit”. In: *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society. 2011.

[21] Luke Prananta et al. *The Effectiveness of Time Stretching for Enhancing Dysarthric Speech for Improved Dysarthric Speech Recognition*. 2022. arXiv: 2201.04908 [cs.SD].

[22] Mirali Purohit et al. “Intelligibility improvement of dysarthric speech using mmse discogan”. In: *2020 International Conference on Signal Processing and Communications (SPCOM)*. IEEE. 2020, pp. 1–5.

[23] Qualtrics. *Qualtrics*. 2005. URL: https://www.qualtrics.com/.

[24] Berrak Sisman et al. “An overview of voice conversion and its challenges: From statistical modeling to deep learning”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2020), pp. 132–157.

[25] Jane Smith, Firstname2 Lastname2, and Firstname3 Lastname3. “A really good paper about Dynamic Time Warping”. In: *Proc. INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*. Incheon, Korea, Sep. 2022, pp. 100–104.

[26] Deborah G Theodoros, Bruce E Murdoch, and Helen J Chenery. “Perceptual speech characteristics of dysarthric speakers following severe closed head injury”. In: *Brain injury* 8.2 (1994), pp. 101–124.

[27] Deborah G Theodoros, Bruce E Murdoch, and Helen J Chenery. “Perceptual speech characteristics of dysarthric speakers following severe closed head injury”. In: *Brain injury* 8.2 (1994), pp. 101–124.

[28] Seung Hee Yang and Minhwa Chung. “Improving dysarthric speech intelligibility using cycle-consistent adversarial training”. In: *arXiv preprint arXiv:2001.04260* (2020).

[29] Emre Yilmaz et al. “Multi-stage DNN training for automatic recognition of dysarthric speech”. In: (2017).

[30] Victoria Young and Alex Mihailidis. “Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review”. In: *Assistive Technology* 22.2 (2010), pp. 99–112.

[31] Yi Zhao et al. *Voice Conversion Challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion*. 2020. DOI: 10.48550/ARXIV.2008.12527. URL: https://arxiv.org/abs/2008.12527.