IMPROVING AUTOMATIC SPEECH RECOGNITION FOR DYSARTHRIC SPEECH

IMPROVING AUTOMATIC SPEECH RECOGNITION FOR DYSARTHRIC SPEECH

by

Luke PRANANTA

to obtain the degree of Master of Science at the Delft University of Technology, to be defended publicly on a September 10th, 2021.

Thesis committee:

Dr. O. Scharenborg	
	Technische Universiteit Delft
Dr. C. Oertel,	Technische Universiteit Delft
B. M. Halpern,	Universiteit van Amsterdam

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

Contents

PREFACE

The last five years have been tough to say the least. My world was shattered when my father passed away almost five years ago now. I never felt more lost and directionless, yet at the same time I was more determined than ever to make something of my life. After the things I've been through, I felt more morally obligated to use my skills for the purpose of helping others. This is what attracted me to the topic at hand, which was introduced to me by Professor Odette Scharenborg. Her vision for speech technology to also include the people who are less fortunate inspired me to take this problem on. Yet I had no background in speech technology. I took a deep dive into an (academic) world that often felt overwhelming and very complex. In the meantime, the world around me collapsed as COVID spread to every corner of the world.

During these tough times, the steady guidance of Professor Scharenborg and the supervision of Bence Halpern became more invaluable than ever. I'm very grateful to both of them for showing me the way through. As I wrap up this thesis, it marks the end of another chapter of my life. I'm not sure what to make of it, I only hope that the work I've done might provide benefits for the people who need it.

Luke Prananta
Delft, June 2021

1

INTRODUCTION

1.1. MOTIVATION

Dysarthria is an encapsulating term for various motor speech disorders in which the muscles that produce speech are weakened or damaged. It is often the by-effect of degenerative diseases such as Parkinson's disease or amyotrophic lateral sclerosis (ALS), but can also be caused by traumatic brain injuries or strokes. The reduced motor capabilities of certain speech muscles results in speech that is slurred and less intelligible. Symptoms include, but are not limited to, abnormal speech rate, hypernasality and dysphonia. The effects become more pronounced as the severity increases [1].

Dysarthria can greatly reduce a person's quality of life and independence [1]. Patients suffering from dysarthria have often more difficulty communicating with other people. Meanwhile, the use of automatic speech recognition (ASR) systems in our daily lives is becoming more normalized. Virtual assistants operating on home devices or smartphones are becoming more and more common. Assistive technologies, devices designed to aid patients with disabilities, are starting to follow this trend as well of incorporating automated speech recognition [2]. These systems could be helpful for dysarthric speech patients that require additional aid.

The problem lies in the ability of current ASR systems to correctly interpret dysarthric speech. In recent studies, the performance of state-of-the-art ASR systems trained on healthy speech were measured against dysarthric speech samples. It became clear that the performance of ASR systems on dysarthric speech is still lacking compared to typical speech [3], [4].

Solving this issue is not straightforward. Training ASR systems with dysarthric speech data is difficult, due to the scarcity of existing dysarthric speech data, and the difficulty of procuring additional data [5]. Approval for recording sessions are hard to obtain and can furthermore be difficult for the dysarthric speech patients to endure. Current state-of-the-art ASR systems are often based on neural network-based acoustic models, thus they require a large quantity of data to converge correctly. Training such models on dysarthric speech data alone is often insufficient for dysarthric speech recognition [5].

To overcome this issue, various solutions have been proposed. Some solutions propose certain training schemes such as multi-step adaptation using both healthy speech data and dysarthric speech data to make ASR systems more robust to variations of speech [5]. Other solutions propose data augmentation techniques, such as creating synthetic dysarthric speech data by modifying healthy speech data [6]. A more recent approach aims to make the dysarthric speech signal directly more intelligible using adversarial learning techniques such as CycleGAN [7]. This is an approach that this thesis will explore further.

Cycle-consistent generative adversarial network (CycleGAN) is a deep learning model for mapping input data from the original domain to a target domain, and it was introduced for unpaired image-to-image translation [8]. Recent studies [7], [9] have proposed using CycleGAN-based models for converting dysarthric speech to healthy speech to improve intelligibility for ASR systems. These studies have validated the performance of dysarthric-to-healthy converted speech in terms of word error rate (WER) and phoneme error rate (PER). The results are promising and indicate that the converted speech is more intelligible for ASR systems than the original dysarthric speech.

The speech rate discrepancy between dysarthric speech and healthy speech is a point of attention. Certain speech enhancement techniques can be applied to possibly improve the performance of this method. One such technique is to apply time alignment using dynamic time warping (DTW) during training, in addition to parallel data processing [9]. Another pre-processing technique to consider is audio time stretching. Time stretching speeds up or slows down the audio without changing the pitch. Other more general speech enhancement techniques such as (static) noise reduction will be considered as well.

1.2. RESEARCH QUESTIONS

This thesis will approach the dysarthric speech recognition problem from a dysarthric-to-normal speech conversion angle, with the aim to 'enhance' dysarthric speech to be more intelligible. State-of-the-art solutions using CycleGAN-based models will be investigated for dysarthric-to-normal speech conversion which utilizes speech encoding and deep adversarial learning. Possible improvements for this approach will be investigated as well, which can improve the intelligibility of a dysarthric speech signal further and consequently improve ASR performance. The main research question will thus be:

- Can CycleGAN-based speech conversion be used to make dysarthric speech more intelligible for ASR systems?

We break this question down into smaller research questions:

- **RQ1**: Does the use of CycleGAN-based speech conversion improve ASR performance of converted dysarthric speech in terms of phoneme and word error rate?

- **RQ2**: Can the effectiveness of CycleGAN-based speech conversion further be improved with parallel training and additional modification such as time alignment using dynamic time warping?

- **RQ3**: Does additional audio pre-processing such as denoising, time stretching and loudness normalization improve the dysarthric speech signal such that it increases ASR performance?

We will conduct experiments to find answers for our research questions. The basis of our experimental design is a reproduction of [9]. In their work they trained a DiscoGAN model using dysarthric speech data from the UASpeech corpus with the purpose of converting dysarthric speech into typical speech. Afterwards, they measure the ASR performance in terms of Phoneme Error Rate (PER). Our experiments will be similar in nature, but we will instead use CycleGAN-based models to perform the speech conversion.

1.3. OUTLINE

This thesis is divided into several chapters, starting with necessary background knowledge in Chapter 2. Next, we discuss our methodology in Chapter 3. In Chapter 4 we describe the experiments that have been conducted and the results of these experiments. Afterwards, we will discuss the results of the experiments in Chapter 5. The thesis will conclude its findings in Chapter 6.

2

Background

In this chapter we provide the requisite background knowledge for this thesis. We start by explaining dysarthria in Section 2.1. We follow up with the knowledge needed to understand CycleGAN by first introducing the concepts of deep (adversarial) learning in Section 2.2 and Generative Adversarial Networks (GANs) in Section 2.3. We then introduce CycleGAN in Section 2.4 and in Section 2.5 we dive deeper into the specific models used in our experiments, namely CycleGAN-VC and MaskCycleGAN-VC. Finally, in Section 2.6 we present some relevant work and problems which are related to the problem at hand.

2.1. DYSARTHRIA

Dysarthria refers to a group of motor speech disorders where muscular control over certain speech muscles become lessened or completely absent. Dysarthria is caused by neuromuscular damage which affect the central or peripheral nervous system. It can affect all aspects of speech, such as respiration, phonation, articulation, nasal resonance, prosody and fluency. The loss of muscular control leads to abnormal speech and a reduction in speech intelligibility. Diseases that can cause dysarthria include, but are not limited to, Parkinson's disease, amyotrophic lateral sclerosis (ALS), stroke and cerebral palsy [1], [10].

A commonly used classification system classifies dysarthria according to which part of the nervous system is implicated [1]. The part of the nervous system that has been damaged in turn affects the symptoms and how certain aspects of speech are affected. The major types of dysarthria with their characteristics are listed below [1], [10]:

- **Flaccid** — Associated with lower motor neurons lesions. Phonation suffers from reduced loudness; voice is notably lower pitched. Articulation suffers from imprecise consonants and hypernasality. Prosody suffers from slow speech rate and monotonous speech.

- **Spastic** — Associated with upper motor neurons lesions. Phonation suffers notably from strained or harsh voice quality; vocal loudness is reduced. Articulation suffers from imprecise consonants and hypernasality. Prosody suffers from notably slower speech rate and monotonous speech.

- **Hypokinetic** — Commonly associated with Parkinson's disease. Phonation suffers from breathy or harsh voice quality; vocal loudness is reduced. Articulation suffers from imprecise consonants. Prosody suffers from monotonous speech, but speech rate is normal or accelerated.

- **Ataxic** — Associated with cerebellar dysfunctions. Phonation suffers from fluctuating voice quality; vocal pitch and loudness fluctuate. Articulation suffers from imprecise and "explosive" sound production. Prosody suffers from slightly slower speech rate.

There is furthermore a common way to categorize dysarthric speakers based on the severity or speech intelligibility. A categorization was proposed by [11], where dysarthric speakers can be divided into four groups:

- **Low severity** — Speech intelligibility rate of 76% or higher.

- **Mid severity** — Speech intelligibility rate between 51% and 75%.

- **High severity** — Speech intelligibility rate between 26% and 50%.

- **Very High severity** — Speech intelligibility rate of 25% or lower.

2.2. DEEP AND ADVERSARIAL LEARNING

Deep learning is a branch of machine learning which uses neural networks, a network which consists of multiple layers of interconnected perceptrons, for representation learning. Adversarial machine learning is a machine learning technique where models are fed false or fake data in order to fool the model. While this technique is commonly used in an attempt to break a model, it can also be used to train other generative models which aim to fool a model. We will elaborate on this in follow-up sections.

2.3. GENERATIVE ADVERSARIAL NETWORKS

Generative Adversarial Networks (GANs) is a machine learning framework proposed by [12]. Two neural networks, the generator G and the discriminator D, challenge each other to a minimax game. In essence, the generator aims to minimize its own loss by maximizing the loss of the discriminator. Thus, adversarial learning is used here to train a generator by trying to fool the discriminator.

During training, the generator learns to estimate samples from the true data distribution, while the discriminator estimates the probability that a sample came from the training data or G. The generator G aims to minimize its loss by maximizing the probability of fooling the discriminator, while D aims to maximize the probability of correct classification. This framework allows for unsupervised learning tasks.

When both G and D are differentiable multi-layer perceptrons (MLPs), they can be trained simultaneously using backpropagation. The minimax loss is formulated as follows:

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

In this loss function, $\boldsymbol{x}$ is drawn from the true data distribution p_{data} and noise $\boldsymbol{z}$ is sampled from a noise prior $p_{\boldsymbol{z}}(\boldsymbol{z})$. The generator G takes $\boldsymbol{z}$ as input and produces a fake sample, which the discriminator D will attempt to classify as fake or real. In a separate term, the discriminator also takes $\boldsymbol{x}$ in an attempt to correctly classify it.

The minimax loss encapsulates the objectives of both the generator and the discriminator. The generator aims to minimize the expected value of $\log(1-D(G(\boldsymbol{z})))$, the inverse log probability of fake samples, by maximizing $D(G(\boldsymbol{z}))$.

The discriminator aims to maximize the expected value of $\log(D(\boldsymbol{x}))+\log(1-D(G(\boldsymbol{z})))$, by both maximizing $\log(D(\boldsymbol{x}))$, the log probability of real samples, and $\log(1-D(G(\boldsymbol{z})))$, the inverse log probability of fake samples. Note that this is essentially a binary cross-entropy (BCE) loss.

In an ideal situation, the implicitly defined probability distribution p_g of generator G converges to the true data distribution p_{data}, with global optimum $p_g = p_{\text{data}}$ [12]. In practice, this criterion can not be reached; the generator G is only able to represent a limited subset of all possible distributions. This also means that by default, GANs do not have well defined stopping criteria. It requires human supervision to validate its results.

While the BCE loss is a natural choice for a two-class decision problem, it was found that it leads to vanishing gradients during the training procedure. An improved loss function was formulated by [13] when they introduced least squares GAN (LSGAN). This loss function uses a least squares loss function for the discriminator objective instead of a

BCE. They propose the use of least square loss functions to address this problem. The use of LSGANs became widespread in practical implementations as it stabilizes training. The modified loss function is defined as follows:

$$\min_G \max_D V_{\text{LSGAN}}(D) = \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[(D(\boldsymbol{x}))^2] + \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[(D(G(\boldsymbol{z})-1))^2] \tag{2.2}$$

The objective of the generator is to minimize the expected value of $(D(G(\boldsymbol{z}))-1)^2$, the loss for classifying fake data. The discriminator must be fooled by maximizing $D(G(\boldsymbol{z}))$. The discriminator aims to maximize the expected value of $(D(\boldsymbol{x}))^2$ and $(D(G(\boldsymbol{z})-1))^2$. The term $(D(\boldsymbol{x}))^2$ represents the loss for classifying real data and $(D(G(\boldsymbol{z})-1))^2$ the loss for classifying fake data.

2.4. CycleGAN

The idea of CycleGAN has its roots in the so-called style transfer problem, or image-to-image translation problem. An example of such problem is transforming images of paintings into lifelike photographs, or transforming a summer landscape into a winter landscape. Solutions for this problem were restrained by the need for paired image data. Such data is both hard to procure and impractical, thus the need for unpaired image-to-image translation solutions arose.

Cycle-consistent adversarial networks were introduced for the purpose of unpaired image-to-image translation [8] and became a staple solution. It is an extension of GAN in which two generators and two discriminators are used, one generator and one discriminator for each mapping direction. The generator takes images from one domain and synthesizes images for the other domain. The discriminator will compute a probability for the generated images belonging to the other domain.

Assume we have a domain X and a domain Y. We define the generators G to map samples from X to Y and F to map samples from Y to X. This framework also produces two discriminators D_X and D_Y. D_X aims to distinguish between training samples from X and samples produced by G. Similarly, D_Y aims to distinguish between training samples from domain Y and samples produced by F. This process is visualized in Figure 2.1.

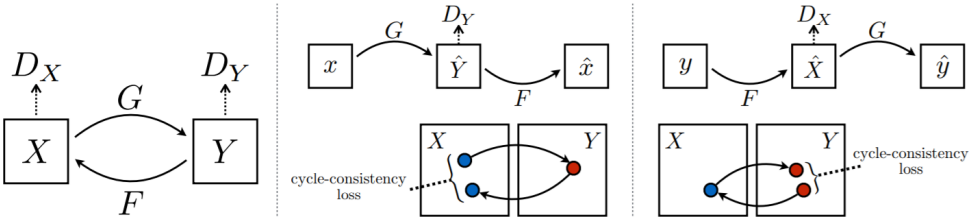

Figure 2.1: A diagram by [8] which visualizes the CycleGAN framework.

We define the GAN loss, which is also known as the adversarial loss, similarly as Equation 2.1:

$$\mathcal{L}_{\text{GAN}}(G, D, X, Y) = \mathbb{E}_{\boldsymbol{y}\sim p_{\text{data}}(\boldsymbol{y})}[\log(D(\boldsymbol{y}))] + \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log(1 - D(G(\boldsymbol{x})))]. \tag{2.3}$$

Note that in Equation 2.1, noise $\boldsymbol{z}$ is sampled and used as input for G. In the adversarial loss function, $\boldsymbol{x}$ is drawn from input domain X and used as input for G, and $\boldsymbol{y}$ is drawn from output domain Y. Also note that in practical implementations, the LSGAN loss defined in Equation 2.2 is often used over the minimax GAN loss.

In addition to the adversarial losses introduced by GANs, they introduce a cycle-consistency loss for both mapping directions to encourage one-to-one mapping. The intuition is to measure the similarity between a sample and the same sample mapped to another domain and back to the original domain. Thus the cycle-consistency loss aims to minimize the difference between a sample $\boldsymbol{x} \in X$ and $F(G(\boldsymbol{x}))$, and the difference between a sample $\boldsymbol{y} \in Y$ and $G(F(\boldsymbol{y}))$. The cycle-consistency loss is defined as follows:

$$\mathscr{L}_{\text{cycle}}(G,F) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\|F(G(\boldsymbol{x})) - \boldsymbol{x}\|_1] + \mathbb{E}_{\boldsymbol{y}\sim p_{\text{data}}(\boldsymbol{y})}[\|G(F(\boldsymbol{y})) - \boldsymbol{y}\|_1]. \tag{2.4}$$

In the final loss function, an additional scaling factor λ_{cycle} is used to increase or decrease the importance of the cycle-consistency loss. The complete CycleGAN loss is then defined as follows:

$$\mathscr{L}_{\text{CycleGAN}}(G,F,D_X,D_Y,X,Y) = \mathscr{L}_{\text{GAN}}(G,D_Y,X,Y) + \mathscr{L}_{\text{GAN}}(F,D_X,Y,X) + \lambda_{\text{cycle}}\mathscr{L}_{\text{cycle}}(G,F) \tag{2.5}$$

The architecture of the generator consists of multiple convolutional layers and residual blocks. The discriminator is implemented with a PatchGAN, a convolutional neural network (CNN) that estimates its decision on local image patches [14].

2.5. CycleGAN-VC

CycleGAN has found application in the speech domain to address a range of problems. It has most notably been applied for Voice Conversion (VC). The goal of voice conversion is to convert the speech from a source speaker to that of a target speaker while keeping the linguistic information intact. This method of voice conversion was first introduced by [15], who named their framework CycleGAN-VC. It is a non-parallel voice conversion method, meaning that it can be trained with unpaired or non-parallel speech data. With non-parallel speech data, speech data of both domains are not paired by their linguistic content. CycleGAN-VC converts speech by converting mel-cepstral coefficients (MCEPs). The model has since then been iterated over several times with novel improvements. This lead to CycleGAN-VC2 [16], CycleGAN-VC3 [17], and most recently, MaskCycleGAN-VC [18].

In terms of architecture, CycleGAN-VC is configured to use CNNs with gated linear units (GLUs) to better represent the sequential structure of speech. CycleGAN uses CNNs with ReLU activation functions by default which are more suited for images. The discriminator uses a more traditional CNN with a fully connected (FC) layer as the last layer.

While the cycle-consistency loss constraints the structure of the mapping, on its own it does not suffice for preserving linguistic information [15]. Therefore an identity-mapping loss is used to better preserve linguistic information. This loss was recommended for the original CycleGAN as well to preserve color composition between the input and output images. The identity-mapping loss is added to the complete CycleGAN loss defined in

Equation 2.5. Like the cycle-consistency loss, it is multiplied with a scaling factor λ_{id} to decrease or increase the importance of this loss. The identity-mapping loss is defined as follows using the symbols defined in the previous subsection:

$$\mathscr{L}_{\text{id}}(G,F) = \mathbb{E}_{\boldsymbol{y}\sim p_{\text{data}}(\boldsymbol{y})}[\|G(\boldsymbol{y})-\boldsymbol{y}\|_1] + \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\|F(\boldsymbol{x})-\boldsymbol{x}\|_1]. \tag{2.6}$$

For CycleGAN-VC2, [16] proposed a range of improvements to the original CycleGAN-VC model. The most notable change is the introduction of two-step adversarial loss. This loss is added to address the over-smoothing caused by the cycle-consistency loss. They propose the use of additional discriminators D'_X and D'_Y for a second adversarial loss for bidirectionally converted features. The loss is defined as follows:

$$\mathscr{L}_{\text{GAN}_2}(G,F,D',X) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log(D'(\boldsymbol{x}))] + \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log(1-D'(F(G(\boldsymbol{x}))))]. \tag{2.7}$$

The architecture was also modified to use a PatchGAN discriminator and the generator architecture has been modified to use a mix of 2D and 1D convolutional layers.

CycleGAN-VC and CycleGAN-VC2 is limited in that it cannot be used to convert mel-spectrograms, as it is unable to learn the time-frequency structures when given these features. The harmonic structure of the output speech is compromised due to this limitation. This led to the development of CycleGAN-VC3 and MaskCycleGAN-VC, which are both extensions of CycleGAN-VC2. CycleGAN-VC3 uses an additional time-frequency adaptive normalization (TFAN) module at the expense of an increased number of parameters [17]. Increasing the number of parameters increases the computational load and makes overfitting more likely. MaskCycleGAN-VC instead proposes a more novel solution without having to increase the number of parameters significantly.

MaskCycleGAN-VC is the latest iteration of CycleGAN-VC proposed by [18]. They propose a novel task which they call fill in frames (FIF), which is performed during training. The idea of this task is to mask a part of the input mel-spectrogram, so that the model can learn to synthesize what is missing using the surrounding (non-masked) sections. This in turn allows the model to learn the time-frequency structures such that output can be generated with proper harmonic structure.

The FIF procedure is illustrated in Figure 2.2. Given a source mel-spectrogram $\boldsymbol{x} \in \mathbb{R}^{b\times t}$ with t the number of frames, and b the number of filter banks, an equally sized binary mask $\boldsymbol{m} \in \mathbb{R}^{b\times t}$ is created. The mask contains a randomly selected temporal region which is set to zero. The mask is applied element-wise on $\boldsymbol{x}$ to produce $\hat{\boldsymbol{x}}$. The generator G is given a channel-wise concatenation of $\hat{\boldsymbol{x}}$ and $\boldsymbol{m}$ to produce $\boldsymbol{y}'$.

$$\boldsymbol{y}' = G(\text{concat}(\hat{\boldsymbol{x}}, \boldsymbol{m})) \tag{2.8}$$

The conditional information given by $\boldsymbol{m}$ allows the generator G to fill in the blanks. To measure the cyclic loss with the input $\boldsymbol{x}$, $\boldsymbol{y}'$ must first be converted back to the original domain using the generator F. The generator F is given a channel-wise concatenation of $\boldsymbol{y}'$ and an all-ones mask $\boldsymbol{m}'$ of equal dimensions. This produces $\boldsymbol{x}''$:

$$\boldsymbol{x}'' = F(\text{concat}(\boldsymbol{y}', \boldsymbol{m}')) \tag{2.9}$$

The cyclic loss can now be defined with an additional cycle-consistency-like loss:

$$\mathscr{L}_{\text{mcycle}}(G,F) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x}),\boldsymbol{m}\sim p_{\text{data}}(\boldsymbol{m})}[\|\boldsymbol{x}''-\boldsymbol{x}\|_1]. \tag{2.10}$$

This loss encourages G to fill in frames using information from surrounding frames. It is a self-supervised process to learn the time-frequency structure in a mel-spectrogram. Unlike CycleGAN-VC3 it does not increase the number of model parameters significantly.

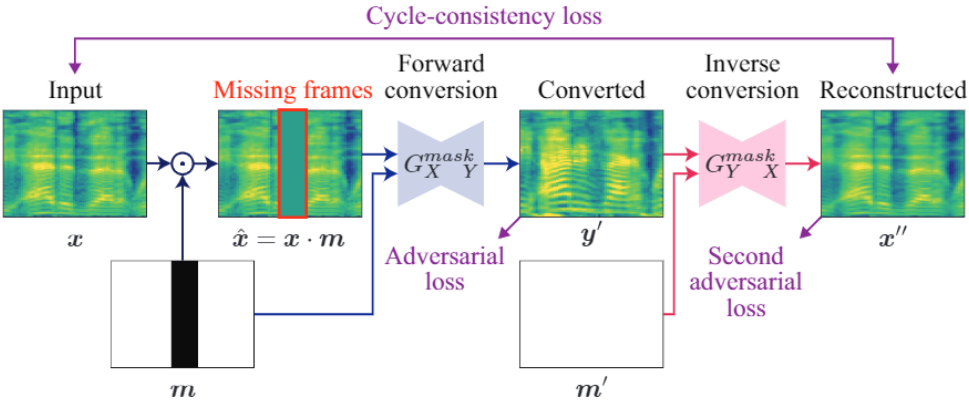

Figure 2.2: A diagram by [18] which visualizes the FIF procedure.

2.6. RELEVANT WORK

We will now look at recent work which uses CycleGAN(-VC) or similar models to make dysarthric or atypical speech more intelligible for ASR systems. These works directly influenced our interest for this particular research direction.

2.6.1. DYSARTHRIC-TO-HEALTHY SPEECH CONVERSION

A dysarthric-to-healthy speech mapping solution using CycleGAN was proposed by [7] to improve intelligibility of dysarthric speech. They use the CycleGAN model to map mel-spectrograms of dysarthric speech to healthy speech. They interpreted the problem as a style transfer problem, similarly to CycleGANs original problem.

Their CycleGAN implementation follows the same architecture as the original CycleGAN implementation by [8]. The CycleGAN model is trained with samples from a Korean-speech database known as the Quality of Life Technology (QoLT) database, which consists of 100 dysarthric speakers of various severities and 30 healthy control speakers. A selection of 187 utterances are available for each speaker.

The waveforms of dysarthric and healthy speakers were first converted into mel-spectrograms with an unknown vocoder. Afterwards the model was trained with parallel training data. One hundred dysarthric utterances and 200 control utterances from the QoLT database were converted using the model. The ASR performance in Word Error Rate (WER) were measured with an open source speech recognition engine from Google. The results of these experiments were promising, as a WER of 33.3% was measured for converted dysarthric speech, while a WER of 67.7% was measured for dysarthric speech, an improvement of 33.3%.

Another work by [9] proposed the use of Discover GAN (DiscoGAN) with Mean Square Error (MSE) regularization for improving the intelligibility of dysarthric speech. DiscoGAN is similar to CycleGAN, differing mostly in the losses used. Instead of using a single cycle-consistency loss, it uses two reconstruction losses for both mapping directions [19]. They propose to train a MMSE DiscoGAN model for mapping MCEPs from dysarthric to healthy speech. They also apply a time-alignment technique on the parallel training data. This is done using Dynamic Time Warping (DTW).

In their experiments, they train and validate the model using the UASpeech corpus. The UASpeech corpus is a parallel dysarthric speech dataset [11]. A subset of the speakers were selected for training and testing. The MCEPs were first extracted using AHOCODER from the dysarthric speech and control speech training sample. DTW was applied to align the features and afterwards the model was trained to map the features from dysarthric to healthy speech. They also trained a baseline mapping method using a deep neural network (DNN) based architecture for later comparison.

An ASR was set up to measure the performance in Phoneme Error Rate (PER). The converted speech of the DNN baseline model were measured and compared with the converted speech produced by MMSE DiscoGAN as well as the baseline dysarthric speech. The DiscoGAN model achieved on average a 22.59% relative decrease in PER for female speakers, and a 6.25% relative decrease in PER for male speakers with respect to the baseline dysarthric speech. The DiscoGAN model outperforms the DNN baseline model with on average a 9.64% and 13.16% improvement in PER for female and male speakers respectively. They conclude that a GAN-based method is efficient for dysarthric-to-healthy speech conversion and preferable over DNN-based methods.

2.6.2. OTHER ATYPICAL SPEECH CONVERSION

We highlight some additional studies done in the domain of atypical-to-normal speech conversion. We first look at whisper-to-normal speech solutions. These tackle the problem of converting soft spoken or whispered speech into normal speech. A CycleGAN-based approach was proposed by [20]; in their work they compare the CycleGAN-based method with a previously established DiscoGAN baseline. With objective measurements, they found that the results are comparable to the baseline, and superior in terms of F0 prediction. In subjective evaluations, 55.75% preferred the CycleGAN-based method over the baseline, and they concluded that the CycleGAN-based method yields more natural-sounding converted speech.

Emotional voice conversion aims to change the emotional state of a given input speech, while preserving the linguistic content and speaker identity. An example of this problem is changing the speech of a female speaker from sad to happy. A CycleGAN-based solution was proposed by [21]. CycleGAN is beneficial over previous methods as it allows non-parallel voice conversion. Previous methods required parallel speech data between different emotional states, which is impractical. Another key aspect of their work is effective F0 conversion using CycleGAN. This is done by first modelling the F0 in different temporal scales using wavelet transform and then converting these scales using a separate CycleGAN model.

3

Methodology

In this chapter we present our methods that allows us to conduct experiments for our research questions. In section 3.1, we first elaborate on the need for dysarthric speech data for our experiments, and the requirements that the speech data has to fulfill. We present the dysarthric speech dataset that we have chosen, and how we select a subset of the data for training and testing.

In section 3.2 we present the CycleGAN implementations that we have chosen for our experiments, according to the requirements of our research questions. Afterwards we discuss the need for feature extraction and data pre-processing in section 3.3.

Evaluation is an important aspect of our experiments. The root of the issue is that modern ASRs are not able to understand dysarthric speech. We aim to enhance the dysarthric speech so that it becomes more intelligible for ASRs. We therefore require ASRs that can emulate modern ASRs which are trained to recognize healthy speech. This will help us to measure the performance of the CycleGAN-based method objectively using measures such as word error rate and phoneme error rate. We delve deeper into this topic in section 3.4.

3.1. DATASET

We require dysarthric speech data to conduct our experiments. The speech dataset must also be a parallel dataset, a dataset where dysarthric speech data is paired with healthy speech data that has equivalent linguistic content. Parallel speech data is useful for the evaluation our experiments. The ASR performance of different types of speech can only be fairly compared if the linguistic content is equivalent. Parallel speech data will also be necessary for **RQ2**, where we conduct experiments with models trained on parallel speech data. As mentioned before, dysarthric speech data is scarce. Our choice of a publicly available dysarthric speech dataset been restricted to two choices, the UASpeech corpus [11] and the TORGO database [22].

The Universal Access Speech Technology Corpus, also known as UASpeech, was introduced by [11]. This corpus contains dysarthric speech data of four female and 13 male dysarthric speakers. Each dysarthric speaker is paired with a healthy control speaker; the control speaker provides the healthy equivalent of the dysarthric speech data. The severity of the speakers range from high (6% speech intelligibility) to low (95% speech intelligibility), and the utterances are carefully chosen to be phonetically balanced.

The TORGO database [22] is an English-speech dysarthric speech corpus. This corpus contains recordings of non-words, short words and both restricted and unrestricted sentences. Restricted sentences were carefully chosen by the authors according to certain requirements. Unrestricted sentences are not limited by any specifications or requirements. A total of three female dysarthric speakers and five male dysarthric speakers are available, and each dysarthric speaker is paired with a healthy control speaker. However, the speech data are not paired according to their linguistic content.

A Dutch Dysarthric Speech Database was recently introduced [23]. At the early stages of this project we considered applying our solutions on Dutch dysarthric speech. Having access to the Dutch Dysarthric Speech Database would have greatly assisted us in this endeavour. However, due to GDPR laws, we were unable to get access to the dataset and thus unable to pursue this path any further.

There are pros and cons for both UASpeech and TORGO; both datasets organize their recordings as easily accessible WAV files, which are arranged by speaker. The main benefit of TORGO is that it also offers recordings of sentence-level utterances, while UASpeech only provides recordings of single word utterances. However, UASpeech is a parallel speech dataset and is clearly organized as such; in contrast, the TORGO database is by and large organized in a non-parallel manner, meaning that paired speech data is missing or not labeled as such. Each dysarthric speaker in the UASpeech corpus has a fixed number of pre-labeled dysarthric speech recordings (bar data corruption) and the control speakers provide an equal number of parallel healthy speech recordings.

In TORGO, each speaker provides a variable number of utterances. The recordings are labeled by number, and a prompt file with the transcription is provided for each recording. While each dysarthric speaker is paired with a control speaker, the recordings are not paired according to its utterance; the recordings are labeled randomly, and no correct pairing could be established between equally labeled recordings from a dysarthric-control speaker pair. There is also an imbalance in the number of recordings between the available speakers, as some speakers have more recording sessions and more recordings available compared to other speakers.

We use the UASpeech corpus for our experiments to reproduce the experiments conducted by [9], and we do not conduct additional experiments with the TORGO database. UASpeech allows us to compare evaluations between different types of speech, and also allows us to perform parallel data experiments. This would not be possible with TORGO due to its non-parallel file organization. Thus, we only focus on conducting experiments with the UASpeech corpus.

3.1.1. DATA SELECTION

We select a subset of the speakers available in UASpeech to create a gender balanced training set. For this we follow [9] by selecting four male speakers (M05, M08, M09 and M10) and four female speakers (F02, F03, F04 and F05). An overview of the selected speakers and their speech intelligibility and dysarthria diagnosis is provided in Table 3.1. A total of 765 utterances were recorded per speaker using a 7-channel microphone array.

Table 3.1: An overview of the selected speakers.

Speaker	Intelligibility	Diagnosis
M05	58%	Spastic
M08	93%	Spastic
M09	86%	Spastic
M10	93%	Unknown
F02	29%	Spastic
F03	6%	Spastic
F04	62%	Athetoid
F05	95%	Spastic

Of the 765 utterances, 300 are unique utterances of uncommon words. The remaining 465 utterances can be divided in three repetition batches. Each batch consists of utterances of 10 digits, the English alphabet, 19 computer commands and 100 common words, a total of 155 utterances. Thus a total of 455 unique utterances are available per speaker, which we will be used for training and testing. Of the seven available microphones, we follow [9] and select only recordings of microphone number 3.

While a specific reason for the speaker and microphone selection was not provided by [9], we can infer their reasoning after our own inspection and validation pass of the corpus. We assume that due to the gender imbalance, the authors decided to compensate this by only selecting four male speakers.

Missing and corrupted files might also be a reason why certain male speakers were omitted from the selection. Despite the good overall quality of the dataset, we deduce that the data for certain speakers are incomplete. In particular, recording files may be missing for certain microphones and some files have been corrupted, providing no speech at all. We have validated that the chosen male and female speakers in combination with the selected microphone do not suffer from the previously mentioned issues.

3.1.2. TRAINING SCHEME

Similar to [9], we will use a leave-one-out cross-validation scheme when training and evaluating a model. This essentially means that we will train a separate model for every speaker combination. With four speakers per gender, we have four combinations per gender. If we take the male speakers as an example, one model will be trained with speakers M05, M08 and M09, and evaluated with speaker M10. This means that every model will be trained with 1365 utterances from three different speakers, and evaluated with 455 utterances from the remaining speaker. Table 3.2 summarizes the test and training speakers per model.

Table 3.2: An overview of every speaker combination for training and evaluation.

Test Speaker	Training Speakers
M05	M08, M09, M10
M08	M05, M09, M10
M09	M05, M08, M10
M10	M05, M08, M09
F02	F03, F04, F05
F03	F02, F04, F05
F04	F02, F03, F05
F05	F02, F03, F04

Training the conversion model requires source speech and target speech. We input speech from the three dysarthric speakers (e.g. M05, M08 and M09) as source speech. As mentioned before, each dysarthric speaker is paired with a healthy control speaker. These control speakers are denoted by prepending the letter C to the speaker id (e.g. CM05, CM08, CM09 etc.). We input speech from the control speakers (e.g. CM05, CM08 and CM09) as the target speech. After training the model, we convert the dysarthric speech of the test speaker (e.g. M10) using the trained conversion model. This example is visualized in Figure 3.1.

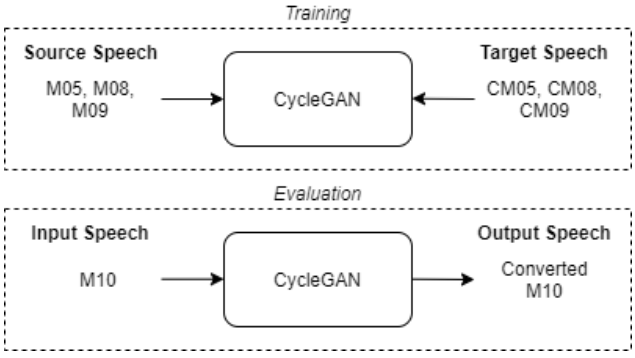

Figure 3.1: A diagram that visualizes the training procedure with speakers M05, M08, M09 for training, and M10 for evaluation.

3.2. MODEL AND ARCHITECTURE

Our research question focuses on the use of CycleGAN models for the purpose of dysarthric-to-normal speech conversion. To find appropriate implementations for **RQ1**, we have looked at CycleGAN implementations used for voice conversion; these implementations are already adapted to handle speech and speech audio features. We have chosen two implementations for our experiments.

The first model implementation is based on a PyTorch implementation of CycleGAN-VC by [24]. CycleGAN-VC is the first iteration of a series of CycleGAN-VC models by [15]. It represents a baseline CycleGAN model without any improvements or upgrades to conduct our experiments with. This implementation also allows us to conduct experiments specific for **RQ2**, where we apply modifications such as training with parallel data and dynamic time warping. We will elaborate further on these modifications and the CycleGAN-VC implementation itself in the next section.

The second model is a MaskCycleGAN-VC implementation by [25]. MaskCycleGAN-VC is the latest iteration of CycleGAN-VC [18]. This model has the latest improvements and represents a state-of-the-art CycleGAN model for our experiments. This implementation includes pre-processing pipelines which are meant for voice conversion datasets such as VCC2018. We have modified the pipelines so that speech data from the UASpeech corpus can be processed, while leaving everything else unaltered. Further details are provided in Section 3.2.2.

3.2.1. CYCLEGAN-VC IMPLEMENTATION

The initial CycleGAN-VC implementation was first sanity-checked to validate that the it is working as intended. We have tested the model by letting it perform its intended purpose of voice conversion. Similar to [15], we tested the model using VCC2016 data [26]. We converted source speaker SF1 to target speaker TF2 and compared the results with the baseline samples from the CycleGAN-VC website [15].

We have also conducted some preliminary experiments by training the model with a leave-one-out training scheme as described in Section 3.1.2. We have picked one combination from Table 3.2 to test the training scheme that we have chosen for our main experiments, and to validate that speech data from the UASpeech corpus can be used with the implementation.

Following our validation of the model we haven chosen to change the default F0 conversion model. The F0 conversion model takes care of transforming the fundamental frequency during the mapping process. The main reason for changing the conversion model is that conceptually, the default F0 conversion model was intended for single speaker conversion, while our training schemes require mapping speech sourced from multiple speakers.

The default F0 conversion model uses a logarithm Gaussian normalized transformation to map the F0. This method of transformation assumes the F0 can be modeled with a single Gaussian model, which does not hold up when multiple speakers are involved. The use of the default F0 conversion model in our preliminary experiments resulted in either unnatural sounding speech or sonic artifacts.

The default F0 conversion model has been replaced with a framework proposed by [21], in which a separate CycleGAN model is used to map the fundamental frequency. Two

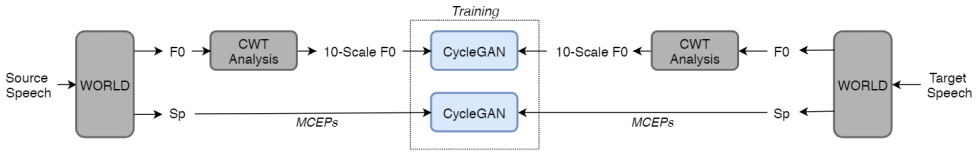

Figure 3.2: A diagram by Zhou et al. which visualizes the proposed framework [21]. Note that the F0 is first decomposed into 10 scales using CWT.

separate CycleGAN models, one for MCEPs mapping and one for F0 mapping, are trained sequentially. In the work of [21], both models are inferred for the purpose of emotional voice conversion.

We have validated this framework by running the preliminary experiment again using the new F0 conversion model. Following a subjective evaluation we confirm that the converted speech sounds more natural, and that sonic artifacts are absent.

TIME ALIGNMENT AND TWO-STEP ADVERSARIAL LOSS

For our experiments for **RQ2** we look to modify the CycleGAN model to potentially improve the baseline model performance. Firstly, we propose to add a time alignment method to the training process of the CycleGAN-VC implementation. A time alignment technique was also used by [9], but it remains unclear if this had a favorable impact on the results. We would thus like to find out if this is the case.

A time alignment step would require the model to train with parallel data. We can only time align a dysarthric recording and a control recording if they have the same linguistic content. Since CycleGAN-VC is by default a non-parallel voice conversion method, we have customized the implementation to allow parallel utterances as well. We use Dynamic Time Warping (DTW) to implement time alignment as an optional step during data loading. In our experiments we will compare the ASR performance between models with parallel data setups with DTW and models with non-parallel data setups without any time alignment.

Finally, we look into implementing a two-step adversarial loss for the CycleGAN-VC implementation. The two-step adversarial loss is one of the improvements introduced for CycleGAN-VC2 and aims to improve the model object and to mitigate the negative effect of oversmoothing [16]. Two-step adversarial loss is thus also included in MaskCycleGAN-VC, since it is an extension of CycleGAN-VC2. We have implemented two-step adversarial loss into the CycleGAN-VC implementation, and we conduct experiments to compare the performance of models with and without two-step adversarial loss.

3.2.2. MASKCYCLEGAN-VC IMPLEMENTATION

The MaskCycleGAN-VC implementation by [25] was taken as is. We did not perform any additional modifications to the model parameters or framework. Changes were only made to the pre-processing pipeline to adapt the UASpeech dataset for this model.

As with the CycleGAN-VC implementation, we performed sanity checks to ensure the implementation works as intended. We use the VCC2018 dataset and convert source speaker VCC2SF3 to target speaker VCC2TM1. We train the model and compare the converted samples with samples from the MaskCycleGAN-VC website [18]. We did not

find any inconsistencies after assessing the results. We thus proceeded with the given implementation without modification.

3.2.3. MODEL CONFIGURATION

We set up our CycleGAN-VC model with a fixed set of hyper-parameters. The model configuration is similar to the configuration proposed in [15]. The network is trained using an Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$, with an initial learning rate of 0.0002 for the generator and 0.0001 for the discriminator. After 2×10^5 iterations, the learning rates linearly decay over 2×10^5 iterations. $\lambda_{\text{cycle}} = 10$ and $\lambda_{\text{id}} = 5$ are set to regularize the cycle-consistency loss $\mathscr{L}_{\text{cycle}}$ and the identity loss $\mathscr{L}_{\text{id}}$. After 1×10^4 iterations, $\mathscr{L}_{\text{id}}$ is set to 0. A batch size of 1 is used during the training procedure, and a segment of 128 frames is randomly selected from a training sample. The CycleGAN-VC model is trained for 1000 epochs

The MaskCycleGAN-VC model is by and large similarly configured as the CycleGAN-VC model. However, the learning rates are set to decay after only 1×10^4 iterations. For MaskCycleGAN-VC, 80-dimensional mel-spectrogram features are extracted and precomputed using librosa [27]. Segments of only 64 frames are randomly selected from the training samples. The MaskCycleGAN-VC model is trained for 300 epochs.

3.3. DATA PRE-PROCESSING AND FEATURE EXTRACTION

To perform our experiments we require a way to extract the desired features from the speech data and modify the speech data available to us. The former is important for all our experiments, as the model only accepts certain features for training. The latter is important to conduct experiments for **RQ3**, where speech data must be denoised or time-stretched before features are extracted.

We have set up a pre-processing pipeline so that we can modify speech data and extract features. This pipeline will extract for every 5 milliseconds: Mel-cepstral coefficients (MCEPs) of predetermined dimensions from the raw WAV files using the WORLD vocoder [28], the logarithmic fundamental frequency ($\log F_0$), and the aperiodicities (APs). A recording will thus be decomposed into the following features: A vector of $\log F_0$ values, a vector of MCEP coefficients (also known as a frame; each frame represents a segment of the total recording), and lastly a vector of APs.

We have set up a separate audio pre-processing pipeline to enable us to modify the waveform so that we can for example apply denoising or time stretching before the feature extraction is performed. For denoising we use a Python package called noisereduce [29]. However, the UASpeech corpus has already been denoised using the same noise reduction algorithm, thus no additional denoising was performed [1]. The time stretching procedure is given more attention in Section 3.3.1.

The next audio pre-processing step is related to loud clicking noises present in the recordings, which occur at either the start or the end of a recording. Following this discovery, we have added a pre-processing step which removes the first and last 200 milliseconds of recording. This solution removes most traces of the undesired clicking noise.

Another point of attention is related to the training process of the model implementation. When a training sample is processed, a fixed-length segment of 128 frames is randomly selected from the extracted frames. The selected frames are then fed into the CycleGAN model. If 128 frames cannot be selected, the process will malfunction and training cannot continue. For this reason we require recordings used for training to have a minimum length. We have determined that a minimum length of one second suffices to extract at least 128 frames. To ensure that each recording is long enough, we trim the existing leading and trailing silences and zero-pad the remaining waveform to a 1 second minimum duration in a pre-processing step.

No additional work is required for the MaskCycleGAN-VC model. The model implementation already contains a complete pre-processing pipeline that takes care of the necessary feature extraction. However, instead of MCEPs, 80-dimensional mel-spectrogram features are extracted and precomputed using librosa [27].

MCEP DIMENSIONALITY

We were interested in discovering the consequences of increasing the feature dimensionality of MCEPs. The experiments performed by [9] used 40-dimensional MCEPs as features [9]. CycleGAN-VC however uses 24-dimensional MCEPs by default [15]. The increase in

[1] An earlier version of UASpeech corpus was considerably more noisy. This prompted us to implement a way to denoise the recordings using noisereduce, which became unnecessary upon obtaining the latest version of UASpeech.

dimensionality can be beneficial for improving the perceived quality of the converted speech. This is not without risk however, as increasing dimensionality increases the complexity of the network and can thus negatively influence the models ability to learn. This phenomenon is also known as the Curse of Dimensionality.

We have conducted preliminary experiments in which we train a model with 34-dimensional MCEPs and 40-dimensional MCEPs, both models were trained for 1000 epochs. We noticed instability during training for both 34-dimensional MCEPs and 40-dimensional MCEPs. In the end, the model was not able to converge and the resulting converted speech samples were noisy and did not represent speech. Following this result we have decided to leave the features as 24-dimensional MCEPs.

3.3.1. TIME STRETCHING

A common and often pronounced effect of severe dysarthria is an abnormal or reduced rate of speech. This is a widespread occurrence for many types of dysarthria. The possible intelligibility improvements of adjusting the speech rate of dysarthric speech is already well known [30]. A more recent study [31] found that adjusting abnormal speech rate on typical speech can improve performance on ASRs. CycleGAN-based voice conversion does not change the speech rate of the converted speech which thus requires an separate solution.

Since we will conduct experiments with both time-stretched speech and non-time-stretched speech, we have added an optional audio pre-processing step where we can adjust the speech rate of dysarthric speech utterance using audio time stretching. With time stretching we can increase or decrease the speed or duration of a recording without affecting the pitch.

The main idea is to adjust the speech rate of the dysarthric speaker based on the speech rate of the paired control speaker (e.g. M05 is adjusted according to CM05). We compute the speed-up ratio by dividing the duration of the dysarthric utterance by the duration of the parallel healthy utterance. The duration of the dysarthric sample will be decreased (or increased) by a factor of the speed-up ratio.

Time stretching is performed using a function available in librosa [27], a Python package for music and audio analysis. In our experiments for **RQ3** we will measure the ASR performance of models that use time-stretched dysarthric speech data as input. The ASR performance will then be compared with our other results.

3.4. ASR SYSTEM

For objective evaluation of our experiments, we emulate ASR systems present in home devices and smart phones, which are trained to recognize healthy English speech. With these ASR systems we can measure the phoneme error rate (PER) or word error rate (WER) of dysarthric speech, converted speech produced by CycleGAN models and healthy control speech. This allows us to compare the performance metrics between the different types of speech. The ASRs must be pre-trained with an English read speech corpus since the UASpeech corpus is an English read speech corpus as well. The purpose of measuring PER is for a more nuanced performance analysis, as improvements in intelligibility can sometimes only occur in the phoneme-level rather than the word-level.

We use a pre-trained Kaldi ASR with the same specifications as the ASR used by [9] for the purpose of phoneme recognition. The phoneme recognition system was trained with the TIMIT dataset and uses a HMM acoustic model. The TIMIT corpus is an English read speech corpus specifically designed for acoustic-phonetic studies [32]. The Kaldi scoring system uses a tool called SCLITE to measure the PER. In order to measure the PER, we require phonemic transcriptions of the UASpeech utterances. We have used a Python package called g2p-en [33] to transcribe the word transcriptions of the utterances into phonemic transcriptions.

3.4.1. PRELIMINARY WER EXPERIMENTS

To measure WER, we have used a pre-trained ESPNet ASR [34] trained with LibriSpeech. An ESPNet-based ASR provide state-of-the-art performance and is easily accessible and configurable using their Python API. LibriSpeech is an English-speech dataset comprised of a collection of 1000 hours of audiobook data [35]. Due to sheer quantity of data it is a popular choice for training English ASRs. Following preliminary experiments where we measured the WER of converted speech of earlier preliminary experiments, we have decided against using measuring WER for our main experiments.

We initially planned to perform and include experiments where we measure WER as well, but we have since then reconsidered this idea after taking into account the context of our experiments. The UASpeech corpus which we use in our experiments only provides word-level utterances. This means the ASR cannot rely on contextual cues that could be found in sentence-level utterances to decode the word utterance. It also means that the ASR relies on, and is more prone to, cues on the phoneme-level. If the ASR is not able to recognize a certain phoneme in a word utterance, it can cause the ASR to recognize it as a completely different word. While WER measurements are more representative of the real world problem, within the limitations of our experiments we believe that measuring WER will not give us further insight into the problem that we are trying to dissect.

4

Experiments and Results

To find answers for our research questions, we conduct various experiments. The basis of our experimental design is a reproduction of [9]. In their work they trained a DiscoGAN model using speech data from the UASpeech corpus with the purpose of converting dysarthric speech into typical speech. Afterwards, they measure the ASR performance in terms of Phoneme Error Rate (PER) using an ASR specified in Section 3.4.

As outlined in Section 3.3, four male speakers and four female speakers are selected for training and testing. Separate models are trained for male and female speakers, and a leave-one-out training scheme is used to train and evaluate a model. To evaluate a model, the speech data of the test speaker is converted using the trained model, and the ASR performance of the converted speech is measured and compared to the ASR performance of dysarthric speech as well as the (healthy) control speech.

In order to find an answer for **RQ1**, we conduct a similar experiment as [9]. We substitute the model with the models that we have outlined in Section 3.2, namely CycleGAN-VC and MaskCycleGAN-VC. Using the selected speakers and training scheme, we train and evaluate a total of eight models per gender.

To answer **RQ2**, we repeat the experiment while substituting the model with modified CycleGAN-VC models. We will train four models per gender for the following modified models:

- CycleGAN-VC with parallel data and DTW
- CycleGAN-VC with two-step adversarial loss
- CycleGAN-VC with parallel data, DTW and two-step adversarial loss

A total of 12 models per gender will be trained and evaluated. The exact modifications are detailed in Section 3.2.

For **RQ3**, we will revisit all the models that we have trained so far. We first pre-process the selected speech data to create time-stretched dysarthric speech for every speaker as explained in Section 3.3.1. Afterward we will re-evaluate all 20 models per gender using time-stretched dysarthric speech.

4.1. RESULTS

4.1.1. BASELINE EXPERIMENTS

Our first experiment sets a baseline ASR performance for the control speech, the dysarthric speech and the converted speech of the chosen models. The PER results for male speakers are shown in Table 4.1 and the PER results for female speakers are shown in Table 4.2. The PER results are shown for individual speakers separately and averaged over all speakers. We also include the results of [9] for comparison. Note that empty cells refer to data that was not given or specified by [9].

We have measured a lower average PER for the control speech than [9]. We similarly measured a lower average PER for male dysarthric speakers than [9]. This is not the case however for female dysarthric speakers, where we measured a small increase in PER compared to [9].

The converted speech produced by CycleGAN-VC and MaskCycleGAN-VC does not improve on the ASR performance of the dysarthric speech. An increase in average PER relative to the average PER of dysarthric speech is observed for both genders. However, MaskCycleGAN-VC does outperform CycleGAN-VC for both genders.

Table 4.1: The ASR performance in PER of male speakers. **Bold** highlights column-wise the best result.

	M05	**M08**	**M09**	**M10**	**Average**
Control	53.9%	48.2%	57.6%	55.6%	53.8%
Dysarthric	**94%**	**64.1%**	**70%**	68.8%	74.2%
CycleGAN-VC	106.8%	72.6%	76.6%	81.7%	84.4%
MaskCycleGAN-VC	102.5%	73.8%	77.1%	**67.3%**	80.2%
Control by [9]	-	-	-	-	64.7%
Dysarthric by [9]	-	-	-	-	77.9%
DNN by [9]	-	-	-	-	82.9%
DiscoGAN by [9]	-	-	-	-	**73.3%**

Table 4.2: The ASR performance in PER of female speakers. **Bold** highlights column-wise the best result.

	F02	**F03**	**F04**	**F05**	**Average**
Control	56.9%	61.6%	74.0%	53.1%	61.4%
Dysarthric	**109%**	**89.8%**	79.9%	**85.9%**	91.2%
CycleGAN-VC	122.6%	99.2%	91.4%	105.3%	104.7%
MaskCycleGAN-VC	116.9%	96.3%	**78.8%**	89.9%	95.6%
Control by [9]	-	-	-	-	65.4%
Dysarthric by [9]	-	-	-	-	87.1%
DNN by [9]	-	-	-	-	75.7%
DiscoGAN by [9]	-	-	-	-	**71.1%**

4.1.2. MODEL MODIFICATION EXPERIMENTS

We investigate the performance effects of the individual modifications. We label the models with parallel data and DTW with 'DTW' in our results table. A model with a two-step adversarial loss is labeled as '2-STEP'. We also highlight performance improvements or deterioration with respect to the baseline model with green and red respectively. The results of the experiments are shown in Table 4.3 and 4.4.

Overall, we observe that no modified model improves on the baseline dysarthric speech performance. However, some modified models improve on the performance of the baseline CycleGAN-VC model. For CycleGAN-VC with DTW, we measure a PER decrease for speakers M05, M09 and M10. The average PER for male speakers decreases relatively by 1.8%. For female speakers, we measure small improvements for F02 and F03 and the average PER decreases relatively by 0.5%.

When two-step adversarial loss is introduced, we find small performance gains for speakers M05, M09 and M10. Speaker M08 however degrades with a relative increase in PER of 5.7%. This causes the average PER of CycleGAN-VC with two-step adversarial loss to increase slightly with respect to the baseline model. We similarly measure small performance gains for female speakers F04 and F05, and on average we measure a relative PER decrease of 0.6%.

Introducing two-step adversarial loss together with DTW does not improve the ASR performance for male speakers. We measure an increase in PER for every male speaker, and the highest average PER overall. However, we measure the lowest average PER for female speakers, with a relative PER decrease of 1.0%. This is caused by speaker F03 and F04 improving on the performance of CycleGAN-VC with two-step adversarial loss, and speaker F05 improving on the performance of the baseline model.

Table 4.3: The ASR performance in PER on modified CycleGAN-VC models trained with male speakers. 'DTW' denotes a model with a parallel data setup and DTW. '2-STEP' denotes a model with two-step adversarial loss. Green highlights performance improvements with respect to the baseline model, while red highlights performance deterioration. **Bold** highlights column-wise the best result.

	M05	**M08**	**M09**	**M10**	**Average**
Dysarthric	**94%**	**64.1%**	**70%**	**68.8%**	**74.2%**
CycleGAN-VC	106.8%	72.6%	76.6%	81.7%	84.4%
CycleGAN-VC + DTW	103.6%	74.6%	75.2%	78.3%	82.9%
CycleGAN-VC + 2-STEP	105.7%	77.0%	76.5%	80.4%	84.9%
CycleGAN-VC + 2-STEP + DTW	107.1%	75.8%	76.8%	83.2%	85.7%

Table 4.4: The ASR performance in PER on modified CycleGAN-VC models trained with female speakers. See Table 4.3 for clarification.

	F02	**F03**	**F04**	**F05**	**Average**
Dysarthric	**109%**	**89.8%**	**79.9%**	**85.9%**	**91.2%**
CycleGAN-VC	126.1%	101.5%	89.6%	103.2%	105.2%
CycleGAN-VC + DTW	122.6%	99.2%	91.4%	105.3%	104.7%
CycleGAN-VC + 2-STEP	126.6%	103.3%	89.0%	99.0%	104.6%
CycleGAN-VC + 2-STEP + DTW	127.2%	102.1%	86.6%	100.0%	104.1%

4.1.3. TIME-STRETCHED SPEECH EXPERIMENTS

We measure the ASR performance of converted time-stretched speech for all our previously trained models. These results are labeled with 'TS'. The results of these experiments are shown in Table 4.5 for male speakers and Table 4.6 for female speakers.

For time-stretched dysarthric speech, we measure a decrease in the average PER for both genders with respect to the baseline dysarthric speech. Speakers M05, F02, F03 and F05 benefit the most from time stretching, and small performance gains were found for speakers M10 and F04. The average PER decreases relatively by 19.8% for female speakers and 5.5% for male speakers. This is similar to the 22.6% and 6.25% improvement found for the DiscoGAN-based method by [9].

None of the CycleGAN-VC-based models improve on the ASR performance of time-stretched dysarthric speech, but some modified models do improve on the performance of the baseline model. Introducing DTW improves the performance of male speakers M05, M09, M10 and female speakers F02 and F03. Introducing two-step adversarial loss results in slight performance gains for male speakers M05 and M10, and female speakers F03 and F05. When both modifications are introduced we measure a slight PER decrease for all female speakers with respect to the baseline model, and for male speakers we measure a slight PER decrease for speaker M05 and M10.

The MaskCycleGAN-VC model shows on average a very similar ASR performance to time-stretched dysarthric speech for both genders. It achieves the lowest PER for speakers M05, F02 and F04. For the remaining speakers we find that the PER is higher than the PER of time-stretched dysarthric speech, but lower than the PER of CycleGAN-VC models.

Table 4.5: The ASR performance in PER of all models trained with male speakers, evaluated with time-stretched dysarthric speech. 'DTW' denotes a model with a parallel data setup and DTW. '2-STEP' denotes a model with two-step adversarial loss. 'TS' denotes time-stretched dysarthric speech data is used as input. **Bold** highlights column-wise the best result. Green and red highlight positive and negative performance differences respectively between the modified CycleGAN-VC model and the baseline model.

	M05	**M08**	**M09**	**M10**	**Average**
Dysarthric	94%	**64.1%**	**70%**	68.8%	74.2%
Dysarthric + TS	76.6%	68.9%	70.4%	**65.4%**	**70.3%**
CycleGAN-VC + TS	78.8%	73.3%	76.3%	77.1%	76.4%
CycleGAN-VC + DTW + TS	78.1%	74.4%	75.5%	74.6%	75.7%
CycleGAN-VC + 2-STEP + TS	78.7%	79.3%	77.8%	76.4%	78.1%
CycleGAN-VC + 2-STEP + DTW + TS	76.7%	75.6%	78.0%	76.4%	76.7%
MaskCycleGAN-VC + TS	**69.2%**	69.3%	75.0%	69.4%	70.7%

Table 4.6: The ASR performance in PER of all models trained with female speakers, evaluated with time-stretched dysarthric speech. See Table 4.5 for further clarification.

	F02	**F03**	**F04**	**F05**	**Average**
Dysarthric	109%	89.8%	79.9%	85.9%	91.2%
Dysarthric + TS	81.8%	**79.3%**	75.5%	**67.9%**	**76.1%**
CycleGAN-VC + TS	83.5%	86.3%	80.9%	79.0%	82.4%
CycleGAN-VC + DTW + TS	81.2%	85.6%	84.2%	77.3%	82.0%
CycleGAN-VC + 2-STEP + TS	83.5%	85.7%	81.6%	77.8%	82.2%
CycleGAN-VC + 2-STEP + DTW + TS	83.4%	85.8%	80.2%	78.8%	82.1%
MaskCycleGAN-VC + TS	**75.4%**	82.8%	**73.9%**	72.5%	76.2%

5

Discussion

5.1. Baseline Experiments

We outline a complete overview of all our results in Table 5.1 for male speakers and Table 5.2 for female speakers. The PER results are shown for individual speakers separately and averaged over all speakers. The results of [9] are also included in both tables for comparison. Note that empty cells refer to data that was not given or specified by [9].

The first thing to note is that our average PER for the control speech is lower than the average by [9]. This can be caused by a few factors. One such factor is the application of a denoising algorithm to the UASpeech corpus, as described in section 3.1, which was performed after the work by [9]. We have also manually removed noisy artifacts such as clicking noises from the speech data.

This could also explain the improved dysarthric speech baseline for male speakers; however, this improvement was not found for female dysarthric speakers. We suspect that speakers with low speech intelligibility, such as speakers F02, F03 and F04, are less susceptible to improvements in the audio quality. Furthermore, our male dysarthric speakers consists mostly of low severity cases with high speech intelligibility, with speaker M05 being an exception. We conclude that the improved audio quality of the UASpeech corpus is beneficial for improving the ASR performance of speakers with high speech intelligibility.

It seems that on first glance using CycleGAN-based speech conversion does not directly improve the intelligibility of the converted speech in terms of PER. Both our CycleGAN-VC model and MaskCycleGAN-VC model show less than stellar results. The PER of converted speech often increases the PER instead of decreasing it which is especially noticeable for CycleGAN-VC. In contrast, the work by [9] show on average a decrease in PER for DiscoGAN-based converted speech.

There are multiple reasons why no improvements are observed. Because of our improved baseline results, there might already be less room for improvement. Further performance differences might be caused by differences in the experimental setup. A notable difference between our setup and [9] is the use of a different vocoder. We have used

Table 5.1: A complete overview of the ASR performance in PER for all models trained and evaluated with male speakers. The percentage between the parentheses indicates the speech intelligibility. 'DTW' denote a model with a parallel data setup and DTW. '2-STEP' denotes a model with second adversarial losses. 'TS' denotes time-stretched dysarthric speech data is used as input. The results are separated into several blocks. **Bold** highlights column-wise the best result within a block.

	M05 (58%)	**M08 (93%)**	**M09 (86%)**	**M10 (93%)**	**Average**
Control	53.9%	48.2%	57.6%	55.6%	53.8%
Dysarthric	94%	64.1%	70%	68.8%	74.2%
CycleGAN-VC	106.8%	**72.6%**	76.6%	81.7%	84.4%
CycleGAN-VC + DTW	103.6%	74.6%	**75.2%**	78.3%	82.9%
CycleGAN-VC + 2-STEP	105.7%	77.0%	76.5%	80.4%	84.9%
CycleGAN-VC + 2-STEP + DTW	107.1%	75.8%	76.8%	83.2%	85.7%
MaskCycleGAN-VC	**102.5%**	73.8%	77.1%	**67.3%**	**80.2%**
Dysarthric + TS	76.6%	**68.9%**	**70.4%**	**65.4%**	**70.3%**
CycleGAN-VC + TS	78.8%	73.3%	76.3%	77.1%	76.4%
CycleGAN-VC + DTW + TS	78.1%	74.4%	75.5%	74.6%	75.7%
CycleGAN-VC + 2-STEP + TS	78.7%	79.3%	77.8%	76.4%	78.1%
CycleGAN-VC + 2-STEP + DTW + TS	76.7%	75.6%	78.0%	76.4%	76.7%
MaskCycleGAN-VC + TS	**69.2%**	69.3%	75.0%	69.4%	70.7%
Control by [9]	-	-	-	-	64.7%
Dysarthric by [9]	-	-	-	-	77.9%
DNN by [9]	-	-	-	-	82.9%
DiscoGAN by [9]	-	-	-	-	**73.3%**

Table 5.2: A complete overview of the ASR performance in PER for all models trained and evaluated with female speakers. See Table 5.1 for additional clarification.

	F02 (29%)	**F03 (6%)**	**F04 (62%)**	**F05 (95%)**	**Average**
Control	56.9%	61.6%	74.0%	53.1%	61.4%
Dysarthric	109%	89.8%	79.9%	85.9%	91.2%
CycleGAN-VC	126.1%	101.5%	89.6%	103.2%	105.2%
CycleGAN-VC + DTW	122.6%	99.2%	91.4%	105.3%	104.7%
CycleGAN-VC + 2-STEP	126.6%	103.3%	89.0%	99.0%	104.6%
CycleGAN-VC + 2-STEP + DTW	127.2%	102.1%	86.6%	100.0%	104.1%
MaskCycleGAN-VC	**116.9%**	**96.3%**	**78.8%**	**89.9%**	**95.6%**
Dysarthric + TS	81.8%	**79.3%**	75.5%	**67.9%**	**76.1%**
CycleGAN-VC + TS	83.5%	86.3%	80.9%	79.0%	82.4%
CycleGAN-VC + DTW + TS	81.2%	85.6%	84.2%	77.3%	82.0%
CycleGAN-VC + 2-STEP + TS	83.5%	85.7%	81.6%	77.8%	82.2%
CycleGAN-VC + 2-STEP + DTW + TS	83.4%	85.8%	80.2%	78.8%	82.1%
MaskCycleGAN-VC + TS	**75.4%**	82.8%	**73.9%**	72.5%	76.2%
Control by [9]	-	-	-	-	65.4%
Dysarthric by [9]	-	-	-	-	87.1%
DNN by [9]	-	-	-	-	75.7%
DiscoGAN by [9]	-	-	-	-	**71.1%**

the WORLD vocoder as proposed by [15] for CycleGAN-VC, while [9] used AHOCODER in their experiments [36]. We did not investigate AHOCODER as it was not readily available as a Python package, but we suspect that the different vocoder is a major factor for the discrepancy in the results.

There is furthermore a minor difference between the models. While DiscoGAN used by [9] is architecturally similar to CycleGAN, CycleGAN uses a single cycle-consistency loss which sums the losses of both mapping directions, while DiscoGAN uses individual reconstruction losses for both mapping directions. These losses are conceptually very similar however, and we suspect that this cannot be the sole reason for the performance differences.

When comparing the CycleGAN-VC model with the MaskCycleGAN-VC model, we observe that the MaskCycleGAN-VC is a better performing model overall. It outperforms the CycleGAN-VC model for both genders, but it is not able to surpass the average baseline dysarthric speech performance. There are only two occurrences where the converted speech of the MaskCycleGAN-VC decreased the PER slightly with respect to the baseline dysarthric speech, with speakers M10 and F04. However, these performance gains are negligible and there is no commonality between the speakers. Speakers M10 and F04 have wildly different speech intelligibility (93% and 62% respectively) and also differ in the type of diagnosed dysarthria. We thus conclude that converting speech using CycleGAN-based models does not improve the ASR performance of dysarthric speech.

The differences between the models serves to highlight the reasons for MaskCycleGAN-VC outperforming CycleGAN-VC. Firstly, the MaskCycleGAN-VC model converts mel-spectrograms, instead of MCEPs; mel-spectrograms are high-dimensional representations of audio, while MCEPs are compressed representations of spectrograms. Mel-spectrograms thus retain more information, which might be beneficial for the purpose of speech conversion. Secondly, a MelGAN vocoder is used to synthesize speech using the converted mel-spectrograms. The vocoder is responsible for converting features back into speech, and can thus directly affect the audio quality of the converted speech. Depending on the implementation, (noisy) artifacts can be introduced in the reconstructed audio. This in turn can influence the ASR performance on converted speech.

Both [9] and [7] used higher dimensional features, in the form of 40-dimensional mel-cepstral coefficients and wideband spectrograms respectively. We were not able to push for higher dimensional features with CycleGAN-VC due to instability during training. We conclude that the use of higher dimensional features will likely lead to better results. Furthermore, the vocoder has likely a large impact on the quality of the converted speech.

5.2. MODEL MODIFICATION EXPERIMENTS

We found mixed results in our model modification experiments. The modified models were not able to improve on the ASR performance of dysarthric speech; however, some modifications did on average improve on the performance of the baseline CycleGAN-VC model. This was observed with CycleGAN-VC with DTW for both genders, which was in line with our expectations, but the overall performance differences between the models are almost negligible.

The effects of introducing two-step adversarial loss with and without DTW were mixed. Male speakers did on average not benefit from two-step adversarial loss, while

small performance gains were on average found for female speakers. We found no discernible pattern that would indicate why certain speakers benefit more from certain improvements over others. Speaker F04 and F05 for example seem to benefit slightly from the introduction of two-step adversarial loss, yet the speakers are very different in terms of speech intelligibility and the diagnosed type of dysarthria.

The small performance fluctuations resulting from the various models and speakers border on random, and seem to indicate that the chosen modifications did not have a major effect on the quality of the converted speech. It is more probable that other factors play a larger role in this. Two-step adversarial loss for example is also included in MaskCycleGAN-VC, but it still manages to perform better than the modified CycleGAN-VC models that we have tested. This points to another property of the model, such as the vocoder, being the most likely culprit for the performance differences.

Overall, we conclude that the small performance gains can be had by introducing parallel data and DTW to the model; however, the improvements are not enough to surpass the baseline dysarthric speech performance. Introducing two-step adversarial loss does generally not improve the performance of the CycleGAN-VC model. The seemingly random and fluctuating ASR performance caused by the modifications indicates that another factor, such as the vocoder, is most likely at play for the overall quality of the converted speech.

5.3. TIME-STRETCHED SPEECH EXPERIMENTS

A quick glance over the results seem to indicate that time stretching speech seems beneficial for improving ASR performance. First off, the average PER of time-stretched dysarthric speech is reduced for both genders. We notice that the male speaker with the lowest speech intelligibility, speaker M05, gained the biggest performance improvement. The remaining male speakers (with low severity) do not benefit or benefit only a little from time-stretched speech. This is likely because their rate of speech already match the speech rate of the control speakers.

The performance gains are most pronounced for female speakers. We observe that all female speakers benefit from time-stretched speech, and that it is notably effective for the high severity cases of speakers F02 and F03. The performance improvement found for speaker F05 is rather surprising, considering the high speech intelligibility of 95%. However, the speech rate of speaker F05 is slightly higher than the speech rate of its control speaker CF05, and it seems that slowing this speech down accordingly using time stretching made it more intelligible for the ASR system.

The performance of time-stretched dysarthric speech on a (modified) CycleGAN-VC model is consistent with our previous findings. The baseline model performance does not surpass the performance of time-stretched dysarthric speech. The modified models for female speakers seem to slightly improve on the performance of the baseline model, but the gains are again negligible. The performance fluctuations between the various speakers and model modifications are again seemingly random. Our conclusions on the modified CycleGAN-VC models from the previous section remains unchanged.

The MaskCycleGAN-VC model shows for both genders a very similar ASR performance to time-stretched dysarthric speech. It does this while also reducing the PER significantly for certain speakers. These speakers are M05, F02 and F04. What is notable about these

speakers is that they all suffer from severe cases of dysarthria, with 58%, 29% and 62% speech intelligibility for M05, F02 and F04 respectively. However, it does not seem to be effective for speakers with low severity dysarthria. Speaker F03 is also an exception to this; speaker F03 is a very severe case of dysarthria with a speech intelligibility of only 6%.

Overall, we conclude that purely time stretching dysarthric speech is effective for improving the ASR performance of dysarthric speech when the atypical speech rate of dysarthric speech is pronounced. We also conclude that for mid to high severity cases, where atypical speech rate heavily influences speech intelligibility, MaskCycleGAN-VC is effective for further reducing the PER for time-stretched dysarthric speech. However, it is not effective for further reducing the PER of dysarthric speakers with low or very high severity dysarthria.

6

CONCLUSION

To investigate if CycleGAN-based methods can improve intelligibility for ASR systems, we conducted three experiments. Firstly, we reproduced the experiments by [9] but instead of using DiscoGAN, we use CycleGAN-VC and MaskCycleGAN-VC. Secondly, we trained CycleGAN-VC with parallel data and DTW and introduced two-step adversarial loss and measured the performance. Thirdly, we changed the speech rate of the dysarthric speech data and converted it with our trained models. These experiments were set up to find answers for the following research questions:

- **RQ1**: Does the use of CycleGAN-based speech conversion improve ASR performance of converted dysarthric speech in terms of phoneme and word error rate?
- **RQ2**: Can the effectiveness of CycleGAN-based speech conversion further be improved with parallel training and additional modification such as time alignment using dynamic time warping?
- **RQ3**: Does additional audio pre-processing such as denoising, time stretching and loudness normalization improve the dysarthric speech signal such that it increases ASR performance?

We now conclude our findings. Despite our best efforts to reproduce [9], we were not able to achieve similar results. The (Mask)CycleGAN-VC models were unable to enhance the dysarthric speech to be more intelligible for ASR systems on its own. Through our experiments, we found several clues that points to the feature dimensionality and the vocoder being the culprit. We also found however that increasing the speech rate of dysarthric speech does make it more intelligible for ASR systems. Furthermore, using MaskCycleGAN-VC we were able to further improve the intelligibility for speakers with mid to high severity cases.

For **RQ1** we conclude that depending on the model specifications, CycleGAN-based speech conversion does not improve the performance of converted dysarthric speech in terms of PER. Our CycleGAN-VC model using 24-dimensional MCEPs and a WORLD

vocoder was not able to improve on the dysarthric speech baseline. The MaskCycleGAN-VC model using 80-dimensional mel-spectrograms and a MelGAN vocoder performed better than CycleGAN-VC overall. We found negligible PER improvements for a few speakers with respect to the baseline, but on average it did not improve on the dysarthric speech baseline.

We conclude for **RQ2** that small improvements can be gained by training CycleGAN-VC with parallel data and dynamic time warping. However, it does not improve with respect to the dysarthric speech baseline. Adding second adversarial losses to the model did not improve the ASR performance. Since two-step adversarial loss is included in MaskCycleGAN-VC, we conclude that other factors such as the vocoder are likely the reason for surpassing CycleGAN-VC in terms of ASR performance.

Lastly, we conclude for **RQ3** that adjusting the speech rate of dysarthric speech using time-stretching, improves the ASR performance. By increasing or decreasing the speech rate of dysarthric speech data, we measured a relative improvement of 19.8% and 5.5% for female and male speakers respectively. We also found that using MaskCycleGAN-VC to convert time-stretched dysarthric speech further improved the performance for mid to high severity dysarthric speakers; however, it did not improve the performance for speakers with low and very high severity dysarthria.

For future work, we suggest investigating the use of different types of vocoders. We find that the vocoder likely plays a major role in the quality of the converted speech. Further work is needed to investigate the quality of synthesized converted dysarthric speech from different vocoders. We also suggest looking into separate solutions for adjusting the speech rate of dysarthric speech, for example using attention-based sequence-to-sequence models [37]. The speech rate is also a major factor to make dysarthric speech more intelligible for ASR systems. Using such solutions in combination with CycleGAN-based speech conversion will likely lead to better results.

BIBLIOGRAPHY

[1] P. Enderby, "Chapter 22 - disorders of communication: Dysarthria," in *Neurological Rehabilitation*, ser. Handbook of Clinical Neurology, M. P. Barnes and D. C. Good, Eds., vol. 110, Elsevier, 2013, pp. 273–281. DOI: https://doi.org/10.1016/B978-0-444-52901-5.00022-8.

[2] K. F. McCoy, J. L. Arnott, L. Ferres, M. Fried-Oken, and B. Roark, "Speech and language processing as assistive technologies," *Computer Speech & Language*, vol. 27, no. 6, pp. 1143–1146, 2013, Special Issue on Speech and Language Processing for Assistive Technology, ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2013.04.005.

[3] L. Moro-Velazquez, J. Cho, S. Watanabe, M. A. Hasegawa-Johnson, O. Scharenborg, H. Kim, and N. Dehak, "Study of the performance of automatic speech recognition systems in speakers with parkinson's disease," vol. 2019-September, ISCA, Sep. 2019, pp. 3875–3879. DOI: 10.21437/Interspeech.2019-2993.

[4] V. Young and A. Mihailidis, "Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review," *Assistive technology : the official journal of RESNA*, vol. 22, 99–112, quiz 113, Jun. 2010. DOI: 10.1080/10400435.2010.483646.

[5] E. Yılmaz, M. Ganzeboom, C. Cucchiarini, and H. Strik, "Multi-stage dnn training for automatic recognition of dysarthric speech," 2017. DOI: 10.21437/Interspeech.2017-303.

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

[7] S. H. Yang and M. Chung, "Improving dysarthric speech intelligibility using cycle-consistent adversarial training," Jan. 2020.

[8] J.-Y. Zhu, T. Park, P. Isola, and A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," Oct. 2017, pp. 2242–2251. DOI: 10.1109/ICCV.2017.244.

[9] M. Purohit, M. Patel, H. Malaviya, A. Patil, M. Parmar, N. Shah, S. Doshi, and H. A. Patil, "Intelligibility improvement of dysarthric speech using mmse discogan," IEEE, Jul. 2020, pp. 1–5, ISBN: 978-1-7281-8895-9. DOI: 10.1109/SPCOM50965.2020.9179511.

[10] H. Ackermann, I. Hertrich, and W. Ziegler, "Dysarthria," in *The Handbook of Language and Speech Disorders*. John Wiley & Sons, Ltd, 2010, ch. 16, pp. 362–390, ISBN: 9781444318975. DOI: https://doi.org/10.1002/9781444318975.ch16.

[11] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," Jan. 2008, pp. 1741–1744.

[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14, Montreal, Canada: MIT Press, 2014, pp. 2672–2680.

[13] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2813–2821. DOI: 10.1109/ICCV.2017.304.

[14] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., ser. Lecture Notes in Computer Science, vol. 9907, Springer, 2016, pp. 702–716. DOI: 10.1007/978-3-319-46487-9_43.

[15] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," Nov. 2017.

[16] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion," May 2019, pp. 6820–6824. DOI: 10.1109/ICASSP.2019.8682897.

[17] ——, "Cyclegan-vc3: Examining and improving cyclegan-vcs for mel-spectrogram conversion," *ArXiv*, 2020.

[18] ——, "Maskcyclegan-vc: Learning non-parallel voice conversion with filling in frames," *ArXiv*, 2021.

[19] T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," *Proc. Int. Conf. Machine Learn. (ICML)*, Mar. 2017.

[20] M. Parmar, S. Doshi, N. J. Shah, M. Patel, and H. A. Patil, "Effectiveness of cross-domain architectures for whisper-to-normal speech conversion," in *2019 27th European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5. DOI: 10.23919/EUSIPCO.2019.8902961.

[21] K. Zhou, B. Sisman, and H. Li, "Transforming spectrum and prosody for emotional voice conversion with non-parallel training data," May 2020. DOI: 10.21437/Odyssey.2020-33.

[22] F. Rudzicz, A. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, pp. 1–19, Jan. 2010. DOI: 10.1007/s10579-011-9145-0.

[23] E. Yılmaz, M. Ganzeboom, L. Beijer, C. Cucchiarini, and H. Strik, "A dutch dysarthric speech database for individualized speech therapy research," 2016.

[24] C. Pajot, H. Hotta, and S. Zalouk, *Maskcyclegan-vc*, https://github.com/GANtastic3/MaskCycleGAN-VC, 2020.

[25] B. Halpern, *Cyclegan-vc pytorch*, https://github.com/karkirowle/cyclegan_pytorch, 2020.

[26] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The voice conversion challenge 2016," in *Interspeech 2016*, 2016, pp. 1632–1636. DOI: 10.21437/Interspeech.2016-1066.

[27] B. McFee, C. Raffel, D. Liang, D. Ellis, M. Mcvicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in python," Jan. 2015, pp. 18–24. DOI: 10.25080/Majora-7b98e3ed-003.

[28] M. Morise, F. Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, pp. 1877–1884, Jul. 2016. DOI: 10.1587/transinf.2015EDP7457.

[29] T. Sainburg, M. Thielk, and T. Q. Gentner, "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires," *PLoS Computational Biology*, vol. 16, no. 10, 2020.

[30] G. Van Nuffelen, M. De Bodt, J. Vanderwegen, P. Van de Heyning, and F. Wuyts, "Effect of rate control on speech production and intelligibility in dysarthria," *Folia Phoniatrica et Logopaedica*, vol. 62, pp. 110–119, 2010. DOI: 10.1159/000287209. (visited on 10/27/2020).

[31] X. Zeng, S. Yin, and D. Wang, "Learning speech rate in speech recognition," Jun. 2015.

[32] L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, Nov. 1992.

[33] K. Park and J. Kim, *G2pe*, https://github.com/Kyubyong/g2p, 2019.

[34] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proceedings of Interspeech*, 2018, pp. 2207–2211. DOI: 10.21437/Interspeech.2018-1456.

[35] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," Apr. 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

[36] D. Erro, I. Sainz, E. Navas, and I. Hernáez, "Hnm-based mfcc+f0 extractor applied to statistical speech synthesis," May 2011, pp. 4728–4731. DOI: 10.1109/ICASSP.2011.5947411.

[37] C.-c. Yeh, P.-c. Hsu, J.-c. Chou, H.-y. Lee, and L.-S. Lee, "Rhythm-flexible voice conversion without parallel data using cycle-gan over phoneme posteriorgram sequences," Dec. 2018, pp. 274–281. DOI: 10.1109/SLT.2018.8639647.